TEXT FLY WITHIN
THE BOOK ONLY

UNIVERSAL
LIBRARY

OU_156745

UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 519 | S86 B Accession No. G. 16037

Author Stockton, John R.

Title Business Statistics. 1958.

This book should be returned on or before the date last marked below.

Copyright ©, 1958

Philippine Copyright, 1958

by

SOUTH-WESTERN PUBLISHING COMPANY

Cincinnati 27, Ohio

All Rights Reserved

The text of this publication, or any part thereof, may not be reproduced in any manner whatsoever without the permission in writing from the publisher.

Library of Congress Catalog Card Number: 58-7932

K-259

Printed in the United States of America

Preface ♦ ♦ ♦

Business Statistics is designed to serve as the text for the basic course in statistics taught in departments and schools of business administration. This text has been written to meet the following objectives:

1. To make students of business aware of the value of statistical analysis in the solution of business problems.
2. To give students sufficient acquaintance with the most commonly used methods of analysis and to enable them to use statistical data effectively in making decisions.
3. To give a strong foundation for more advanced study to those students who want professional training in statistics.

These objectives were used as the criteria for selecting the topics that are included in *Business Statistics*. In making the selection, it was considered important to keep in mind the kinds of statistical analysis businessmen use frequently and to resist including a subject solely because of its interest to statisticians. Thus, this text places strong emphasis on descriptive statistics, including the analysis of time series. Since it is not possible to present a complete discussion of statistical inference in an introductory text, it has been necessary to be highly selective and limit this discussion to the basic concepts.

The wide acceptance of statistics as a basic part of the curriculum in business is recognition of the fact that statistical analysis is an important tool of business management in controling operations and in making decisions. Thus, in Parts I through V the discussion is aimed primarily at describing the various methods of statistical analysis and at pointing out how these methods are used in making business decisions. In order to explain the methods of analysis, the illustrative material throughout the book is based upon authentic business problems.

The development of this text has followed the basic principle that a discussion of the problems to which statistical analysis can be profitably applied is as important as the methods of analysis themselves. Therefore, in Part VI a broader approach to decision-making has been adopted and this part is devoted entirely to a discussion of business problems, the solutions to which are based upon the many different methods of analysis presented earlier in the text.

The end-of-chapter materials are divided into two groups: (1) study questions, and (2) problems. The study questions are designed for use in developing an understanding of statistical analysis through class discussion. The problems require the use of analytical skills on the part of the student as he applies the principles he has learned to problem solving and decision making. The tables needed for solving problems are conveniently grouped in the Appendix. *Laboratory Problems for Business Statistics*, which has been designed to accompany the text, provides students with twenty comprehensive, realistic problems that provide additional opportunity to apply important principles in the analysis of business data. It is hoped that students working these problems will be forced to think through to conclusions, using statistical data and statistical methods.

Presenting statistical data graphically is particularly important in business, and a large number of charts has been included in the text to illustrate the various uses of the graphic method. All of the charts, except a few reproduced from published sources, were drawn by Mrs. Roberta Steele in the style used by the Bureau of Business Research, The University of Texas.

The manuscript has been read by E. A. Gaumnitz, Dean, School of Commerce, The University of Wisconsin, and Wesley C. Ballaine, Professor of Business Administration, University of Oregon. To them I express my appreciation for their valuable suggestions.

J. R. S.

CONTENTS

Introduction

Chapter		Page
1	STATISTICS IN BUSINESS	1
	Use of Statistical Data in Management Decisions	2
	Basic Steps in Statistical Methods	4
	Internal and External Sources of Statistical Data	9
	Management of a Statistical Office	11
	Statistical Pitfalls to Avoid	12

Part I ♦ Assembling Statistical Data

2	STATISTICAL TABLES	19
	Types of Classification	19
	Presenting Statistical Data	25
	Construction of Tables	34
	Major Parts of a Table	37
	Mechanical Features of Tables	43
3	INTERNAL STATISTICS — RECORDS AND REPORTS	48
	Collecting Internal Data	49
	Basic Types of Internal Records	50
	Machine Methods of Recording and Summarizing Statistical Data	54
4	PUBLISHED DATA	69
	Primary and Secondary Sources of Data	70
	Guides to Published Data	70
	Important Sources of Statistical Data	71
5	COLLECTION AND TABULATION OF EXTERNAL DATA	85
	Collecting External Data	85
	Methods of Collecting External Data	88
	Tabulation of External Data	97

Part II ♦ Statistical Description

6	STATISTICAL ANALYSIS	105
	Purpose of Statistical Analysis	105
	Ratios, Percentages, and Relatives	106

Chapter Page

7 CHARTS 128
Simple Bar Charts 128
Two- and Three-Dimension Charts 132
Pictograms 133
Line Charts 136
Component Parts Charts 147
Statistical Maps 149
Mechanical Features of Chart Construction 153

8 AVERAGES 160
Use of Averages 160
Definition and Computation of the Principal Averages 163
Chief Characteristics of the Principal Averages 183

9 DISPERSION 192
Importance of Dispersion 192
Measures of Absolute Dispersion 194
Chief Characteristics of Measures of Dispersion 204
Relative Dispersion 206
Measures of Skewness 207
The Normal Distribution 209

Part III ♦ Statistical Induction

10 SAMPLING 215
Importance of Sampling 215
Sampling Error 216
Selection of a Sample 217
A Simple Random Sample 218
Confidence Intervals 221
Computation of $\sigma_{\bar{x}}$ 223
Bias in the Sample 224
Importance of the Size of the Universe 225
Standard Error of a Percentage 228
Determining Size of Sample Needed 231
Sample Design 235
Cluster Sampling 236
Testing Hypotheses 238

Part IV ♦ Analysis of Business Change

11 TIME SERIES 245
Types of Fluctuation 248
Importance of Time Series Analysis 255
Adjustment for Calendar Variation 256
Semilogarithmic Charts 258
Deflation of Time Series 267

12 SECULAR TREND — THE STRAIGHT LINE 276
Nature of Trend 276
Characteristics of a Straight Line 279
Measuring Secular Trend 281
Use of Linear Trend 292

Chapter Page

13 SECULAR TREND — NONLINEAR 295

Straight Line Fitted to Logarithms by Method of Least Squares 296
Second-Degree Parabola 299
Growth Curves 301
Choice of Trend Line 307
Making Use of a Measure of Secular Trend . . 308
Forecasting Secular Trend 314

14 SEASONAL VARIATION 318

Indexes of Seasonal Variation 318
Measuring Seasonal Variation — A Problem of Averages . . 321
Specific Seasonals 324
Averaging Specific Seasonals 333
Evaluation of the Different Methods of Measuring Seasonal Variation 339
Changing Seasonal Pattern 340
Making Use of a Measure of Seasonal Variation . . 344
Adjusting Data for Seasonal Variation . . . 352

15 CYCLICAL FLUCTUATIONS IN TIME SERIES 356

Importance of Business Cycles in Business Management . . 356
Isolating Cyclical Fluctuations 358
Forecasting Cyclical Fluctuations in Time Series 379

16 CONSTRUCTION OF INDEX NUMBERS . . 384

The Nature of an Index Number 384
Simple Relatives 386
Averages of Relatives 386
Factors Test 400
Method of Aggregates 401
General Comments on Index Number Construction 405

Part V ♦ Analysis of Relationship

17 CORRELATION 411

Line Charts . . . 411
Scatter Diagrams 411
Relationships Involving More than Two Variables . 415
Estimates Based on a Scatter Diagram . . 418
Nonlinear Relationships . . . 419
Cross-Classification Tables 420
Correlation and Cause-and-Effect Relationships . . 422

18 MEASUREMENT OF SIMPLE LINEAR CORRELATION 430

The Linear Regression Equation 430
Estimates 433
Reliability of Estimates 434
Shorter Method of Computing S_y 436
Abstract Measures of Relationship 437
An Alternative Method of Computation (Product-Moment) . . . 440
Grouped Data (Product-Moment Method) 442
Some Examples of the Use of Correlation Analysis . . 446

19 MEASUREMENT OF NONLINEAR AND MULTIPLE CORRELATION . . 451

Nonlinear Correlation 451
Multiple Correlation 460

Chapter Page

Part VI ♦ Applications in Business

20 BUSINESS BAROMETERS 471

The Need for Business Barometers 471
Measures of Total Activity 473
Components of Gross National Product 478
National Income and Personal Income 482
Measures of Individual Business Processes 484

21 BUSINESS FORECASTING 495

Secular Trend 496
Seasonal Variation 502
Cyclical Fluctuations 503
Forecasts for an Industry and Individual Business 517

22 MARKETING RESEARCH 521

Agencies and Functions of Marketing Research 521
Types of Marketing Research 526
Comments on Some Statistical Techniques Used in Marketing Research 533

23 STATISTICAL QUALITY CONTROL 538

Variability in the Manufactured Product 538
Control Charts for Variables 540
Control Chart for Fraction Defective 550
Acceptance Sampling 554

24 COMMUNITY ECONOMIC SURVEYS 559

Nature of Economic Surveys 559
Users of Economic Surveys 562
Basic Data for Economic Surveys 563
Economic Potential 571

Appendixes

A Glossary of Symbols 577
B Statistical Formulas 581
C Table of Squares, Square Roots, and Reciprocals 587
D Table of Values of $\sum X^2$ and $\sum X^4$ for Computing Least Squares Trend Lines 597
E Use of Logarithms 598
F Table of Four-Place Logarithms 602
G Table of Ordinates of the Normal Curve and Areas Under the Normal Curve 604
H Selected Reading References 606

INDEX 609

INTRODUCTION

chapter 1 ♦ Statistics in Business

1 STATISTICS IN BUSINESS

One of the outstanding features of the Industrial Revolution has been the increase in the scale of business operations. When production was in the handicraft stage, individual businesses were small and production was on a small scale. The management of such a business enterprise was very different from the management of a large-scale business at the present time. The information needed by the executives was much less extensive than at present; therefore it was much easier for them to keep themselves well informed. The market was close at hand; the customers were, for a large part, personal friends of the owner of the business; and there was obviously little need for an elaborate analysis of the market. The businessman had easy methods of finding out what people thought of his product and what they wanted to buy. All he had to do was to ask questions and listen to his customers' remarks.

Under our present economic system, most of the goods bought by consumers are manufactured long before they are offered for sale. Producers and consumers are strangers to each other, and goods are manufactured in anticipation of consumer demand. All of this means that the facts about consumer preferences and desires are not so easy to get or so simple as under the earlier systems; yet, since production is carried on so far ahead of consumption, such information becomes more necessary than ever.

When production was by hand in small shops, there was little need for elaborate production and personnel records. The manager of a business, who was usually also the owner, worked in the shop with the men. He knew them personally and had no need for personnel data. He likewise knew the progress being made on the work in the shop, and did not need extensive records on production. Any facts he needed he could learn from direct observation; most of the facts he needed he had in his head.

In the modern factory the executives need elaborate systems of reporting facts about operations to keep them informed on what is happening. It is impossible to walk around the shop and see how things are going, just as it is impossible to know all the employees and be familiar with their abilities and faults. Methods of gathering and filing information must be provided to keep the executives informed on the progress of the business.

USE OF STATISTICAL DATA IN MANAGEMENT DECISIONS

Many facts about business problems are capable of being expressed numerically, or *quantitatively.* Sales are expressed as a given number of dollars or a given physical amount of product. Production data are usually given in the form of amounts. Facts about markets are expressed in terms of numbers of potential customers; quotas are expressed as quantities; changes in business conditions are expressed in terms of quantitative changes in different lines of activity. The quantitative method of expressing facts is superior to the more general, descriptive form of expression, known as *qualitative description.* Qualitative description is used by the office manager who classifies employee performance according to categories such as satisfactory, improving, and unsatisfactory. There is an increasing tendency to express facts quantitatively whenever possible.

In order to make use of quantitative data it is necessary to have some systematic method of organizing and summarizing the individual facts. A large, unorganized mass of facts gives little information of value. On most occasions the executive is not interested in the individual facts recorded in the accounting system. He wants the information summarized in reports that will give the facts about a large number of transactions. This is typical of all kinds of quantitative data; to be of maximum usefulness the data must be classified, organized, analyzed, and summarized. The technical methods employed in doing this constitute the major part of statistical methods.

The word "statistics" is sometimes given different meanings; and in order to avoid misunderstandings, it is probably better to avoid using the word alone. If the terms "statistical data" and "statistical methods" are used, the meaning is generally clearer. *Statistical data* may be defined as facts expressed in quantitative form. *Statistical methods* may be defined as techniques used in collecting, summarizing, analyzing, and interpreting such facts.

A Tool of Management

Because of the importance of statistical facts, every business executive should have, as one of the important tools of management, a familiarity with the methods of using statistical data. Some businesses can make more use of statistical methods than others; certain functions in a business are able to use this technique more than others. But the methods of statistics have so many applications to the problems of business that

it is highly advantageous for all executives to be familiar with the fundamental principles of statistical methods and to know how they may be applied to the problems of business. This does not mean that every businessman should have the technical training of a statistician, but it does mean that the business executive should understand the ways in which statistical data can be used.

A proper use of statistical data requires some knowledge of the methods of analysis used on the data. This is not so much a knowledge of how to make the analysis as an understanding of what the analysis means, what it does *not* mean, and how the results of the analysis can be used. Having an understanding of statistical methods will better enable the businessman to judge the validity of information coming to his attention. Understanding the use of statistical methods in business can best be acquired by making some study of the technique used in analysis. This should equip one to interpret the results in the most effective manner and, at the same time, develop a critical attitude toward methods of analysis and thinking. It is important, further, that a student learn not only the uses that can be made of statistical methods, but also the limitations of these methods, which again simply means that the basic principles of statistical analysis should be understood if the results are to be used in the most intelligent manner.

In addition to understanding statistical methods, the businessman should know in a general way the important sources of statistical data relative to business problems that are available to him. It is easier to describe the methods of collecting and analyzing data than to describe all the actual data that may be of use to businessmen. Methods are very much the same regardless of the facts the data represent, while the information in statistical form that is of interest to businessmen is as varied as the facts about business.

Statistical Units

The inherent advantage of the methods used by the statistician in analyzing facts is that greater precision can be attained when facts are expressed as definite quantities than when they are expressed in purely descriptive terms. In any particular problem, however, it is an obvious prerequisite that the facts should be capable of expression in quantitative form; they must be characteristics that can be measured or individual cases that can be counted. Anything that exhibits differences of magnitude or of number may be called a *variable* and may serve as the raw material for statistical analysis. The values of some variables are arrived at by merely counting individual cases, such as the number of employees in the departments of a business. In this instance, the statistical unit is one employee.

When the number of units in a certain group has been obtained, this number becomes a value of the variable.

Other values of variables are obtained by measuring some characteristic. A company's retirement and pension program may require the businessman to know the age of each employee in the business. The values of the variable in this case are arrived at, not by counting, but by measuring in years the age of each employee. A sales manager may want to know the size of orders received from different types of customers. This characteristic is measured by computing the value of each order.

Still other values of variables are obtained by computing the ratio between different factors, such as the per capita consumption of different kinds of meat in different years. These derived variables, computed by one method or another, may assume many forms. A few examples are: death rates per thousand population, per capita income for each state, average yield of cotton per acre for the states in the cotton belt, and average daily consumption of cotton by textile mills.

BASIC STEPS IN STATISTICAL METHODS

STATISTICAL METHODS — THE SIX BASIC STEPS

1. Defining the problem
2. Assembling relevant information
3. Original investigation
4. Classification
5. Presentation
6. Analysis

Defining the Problem

The first step in rational thinking on business problems, as well as in thinking on any other subject, is the proper definition of the problem at hand. A great deal of time may be wasted in a business conference in discussing a problem if an exact definition of the problem has not been worked out. One individual may be thinking of the problem as something entirely different from another's conception of it. Naturally, since they are not talking about the same thing, they do not make much progress toward a solution. Even when only one person is thinking about a problem he will not reason most effectively unless he states definitely what his problem is.

Assembling Relevant Information

When a problem has been isolated and it is definitely known what must be solved, the next logical step is to assemble all relevant information that can be obtained. Although sometimes the facts will be predominantly statistical, on other occasions the most important information may be in other than numerical form. In a large percentage of problems, however, the statistical information will be at least an important part of the total information assembled.

In assembling the statistical data about a problem, the first thing to do is to discover what information has already been collected. This information, if it deals with internal operations, may be in the records of the business; or, it may have already been collected and published by some agency outside the business. Unfortunately, it is not uncommon for a business to make decisions without using information that could be obtained easily from that business's own records. One reason for this is the fact that records may be kept by one executive for one purpose without other executives knowing that such records exist. Many times the information is not available in the form that would serve the purpose at hand, but within a short period of time the data could be assembled into a meaningful table, graph, or report for the person or persons making the decision.

There are a large number of fact-gathering agencies which publish information of value to executives. Some of these agencies are listed and described in Chapter 4. It is always important to remember that such agencies are potential sources of pertinent and helpful information. For the most part, it is easier and less expensive for an individual business to secure information from these agencies than for that business to collect the same information itself. For example, governmental agencies distribute statistical data at a price that approximates the cost of printing, and trade associations normally offer information to members at nominal cost.

Original Investigation

After a complete study of the problem and the available information, it may be discovered that not all the facts needed are at hand. If the information is important enough to the decision to justify the cost of collecting, and if the decision can wait long enough, an original investigation is the next logical step.

It is unlikely that an executive concerned with the data used in a problem will make the survey himself, but it is necessary that he know a great deal about the problems involved in making such a survey. It will usually be his decision whether the investigation should be made or whether the problem should be settled on the basis of the facts already available. There

is always a limit to the amount of money that can profitably be spent in collecting information on a particular subject, and there is usually a limit to the length of time a decision can wait. Furthermore, the facts it is possible to collect about a problem will often not be entirely satisfactory, no matter how skillfully the collecting is done. Whether to decide the question at hand on the basis of the information already available or to make additional investigation is frequently a difficult matter to determine. The executive who is to assume the responsibility of solving the business problem is the one who should decide whether or not to secure additional facts. If he makes an intelligent choice, it is only because he knows a great deal about the problems of collecting data, the different methods of collecting data, and the advantages and disadvantages of each method.

Classification

The executive of a business or government agency will normally make use of statistical data in classified form. A mass of unclassified information cannot be used effectively until it has been organized.

This principle is so obvious that it is perhaps absurd to labor the point, but the following example may emphasize the importance of classification of data. It will also illustrate the fact that a great deal of classification is taken for granted. Assume that a concern manufacturing cosmetics and selling to wholesalers and retailers made 11,765 individual sales during a certain month. If the sales manager had in front of him the 11,765 sales invoices, he would have all the facts available about the sales of the company for that month. These invoices would represent a complete account of each individual sale.

It is obvious that the invoices in this form would be practically worthless to the sales manager as far as giving him significant information about the business for the month is concerned. On the other hand, properly classified and organized, this same mass of information would tell the sales manager a great deal he would want to know.

The following list suggests some of the classifications of the data the sales manager would ordinarily want to have:

1. Amount of sales by sales territories.
2. Amount of sales by salesmen.
3. Amount of sales by price groups.
4. Amount of sales on which excise tax must be paid, and, if the rate of tax is different for some items, the amount of sales in each tax group.
5. Comparison of all of these with other periods of time.

Most sales managers would want more classifications than are given here if they could be obtained without too much work and expense; but

it is safe to say that the sales manager would not want to see the individual invoices unless there was something unusual about certain ones. The significant facts about a mass of detailed information are seen more easily in a classification of these facts than in the original data.

Classification is essentially a process of separating a mass of individual facts or observations into groups on the basis of some similarity. In other words, the items put into one group have certain characteristics in common which differentiate them from the items put in other groups. If sales were classified on the basis of salesmen, all the invoices in one group would be alike in that they were all sales of one man. If the invoices were classified on the basis of territories, the characteristic common to all members of a class would be the geographical location of the dealers buying the goods.

Because classification is the first step in analysis of a mass of statistical data, the manner in which the classification is made determines largely what further analysis is possible. There are several recognized basic types of statistical classification, and the methods of analyzing these different types of classification represent a major part of statistical method. This fact will make it necessary to devote some time in the next chapter to consideration of various types of classification.

Presentation

The most common method of presenting data after they have been properly classified is the *tabular form*, which is an orderly arrangement of data in columns and rows. If the number of groups in the classification is small, the data can be presented in *paragraph form*, as illustrated on page 9. But if the number of classes is large, the tabular form is the only practicable method of presentation.

The advantages of the tabular form may be readily understood by comparing the data presented in Table 1-1 on the next page with the data presented in the paragraphs opposite the table, for both the table and the paragraphs contain the same information.

The written paragraphs are tiresome to read, since it is necessary to repeat the same words so many times. In the table the same result is accomplished with much less repetition, since the caption at the head of a column applies to all the figures in the column. The economy of the tabular form of presenting data is so well recognized that it is generally accepted as the most efficient method of presenting nearly all types of distributions. There is a minimum number of figures, however, below which it is probably a saving not to use the tabular form. This number may vary with circumstances, but in general, the smaller the number of figures, the easier it is to give them in the text of a discussion rather than in a table.

Table 1-1

TENURE OF OCCUPIED DWELLING UNITS IN THE UNITED STATES: 1900 TO 1950

Census year	Total	Reporting tenure	Owner-occupied		Renter-occupied	
			Number	Per cent	Number	Per cent
1900	15,963,965	15,428,987	7,205,212	46.7	8,223,775	53.3
1910	20,255,555	19,781,606	9,083,711	45.9	10,697,895	54.1
1920	24,351,676	23,810,558	10,866,960	45.6	12,943,598	54.4
1930	29,904,663	29,321,891	14,002,074	47.8	15,319,817	52.2
1940	34,854,532	34,854,532	15,195,763	43.6	19,658,769	56.4
1950	42,826,281	42,826,281	23,559,966	55.0	19,266,315	45.0

SOURCE: Department of Commerce, Bureau of the Census, *U.S. Census of Housing: 1950*, Vol. I, Part 1.

Tabular Form of Presenting Data

The ability to organize data into a compact, logical table must be acquired by anyone who deals with statistical data. The construction of ordinary tables is not a highly technical problem, although it does require some experience and a thorough understanding of the data and what the table is to show. For this reason it is not unreasonable to expect that every businessman should be familiar enough with tabular forms to be able to set up ordinary tables in the best form, and particularly to be able to read and understand easily any properly drawn and correctly designed table. The construction of very complex analytical tables is properly the function of the statistician, but most tables used by business executives are not of this type.

Analysis

In many situations the information supplied in the form of a statistical tabulation may give the executive all the information he needs, but in many other situations the data need further analysis to extract from them all the valuable information they contain. This leads to the last step in statistical methods, the analysis of statistical data. Most of the chapters in this text will be devoted to methods of analyzing statistical data after they have been collected and put into tabular form. The methods of analysis discussed in these chapters will be those most generally used, since this is only an introduction to the subject. These methods, however, are so generally applicable to statistical data that everyone who has occasion to use such data should understand their significance.

According to the Department of Commerce, Bureau of the Census, *U. S. Census of Housing: 1950*, Vol. I, Part 1, there were 42,826,281 occupied dwelling units in the United States in 1950. Of these, 23,559,966 or 55.0%, were owner-occupied, and 19,266,315, or 45.0%, were renter-occupied. Information on tenure was secured for all of the occupied dwellings.

In 1940 there were 34,854,532 occupied dwelling units, of which 15,195,763, or 43.6% were owner-occupied, and 19,658,769, or 56.4%, were renter-occupied. In this year, information on tenure was secured for all occupied dwellings.

In 1930 there were 29,904,663 occupied dwelling units, but information on tenure was secured for only 29,321,891 units. Of these dwelling units for which information was secured on tenure, 14,002,074, or 47.8%, were owner-occupied, and 15,319,817, or 52.2%, were renter-occupied.

In 1920 there were 24,351,676 occupied dwelling units, but information on tenure was secured for only 23,810,558 units. Of these dwelling units for which information on tenure was secured, 10,866,960, or 45.6%, were owner-occupied, and 12,943,598, or 54.4%, were renter-occupied.

In 1910 there were 20,255,555 occupied dwelling units, but information on tenure was secured for only 19,781,606 units. Of the dwelling units for which information on tenure was secured, 9,083,711, or 45.9%, were owner-occupied, and 10,697,895, or 54.1%, were renter-occupied.

In 1900 there were 15,963,965 occupied dwelling units, but information on tenure was secured for only 15,428,987 units. Of the dwelling units for which information on tenure was secured, 7,205,212, or 46.7%, were owner-occupied, and 8,223,775, or 53.3% were renter-occupied.

Paragraph Form of Presenting Data

The refinement of the methods of analyzing data has been going on constantly, with the result that more and more significant information is being extracted from collections of statistical data. Since the collecting of data is expensive and time consuming, any improvement in the methods of analyzing data to make them provide more information is a profitable procedure.

It should be emphasized that the collecting of data is related to the analysis of the data. The information to be collected frequently is determined by the analysis intended. Actually there is no hard and fast line between the collection and analysis of statistical data, and in a sense, collection and tabulation of data become a part of the analysis. As was noted on page 2, all these processes together constitute what is called statistical methods.

INTERNAL AND EXTERNAL SOURCES OF STATISTICAL DATA

It is convenient to distinguish the kinds of statistical data used by the businessman according to their source, the chief distinction being between data that originate within a particular business organization and

data that originate outside of it. Information that relates to the operations of an individual business is classified as *internal data* for that business. The source of the data is a record or report originating within the business. Information that relates to some activity outside the individual business is classified as *external data*. This classification is significant only with respect to an individual concern, since internal data for one concern are external data for another. For example, the sales of Sears, Roebuck & Company are internal data for that company but external data for any other organization.

The classification of statistical data as internal or external is important to the business statistician because of the difference in the methods that must be used in collecting the two types. Since internal data relate to the operations of an individual business, all the information needed is under the control of the concern. The problem of collecting the information is one of designing the proper records in which to enter the information so that it will be readily available when needed. Usually the record is kept in such a manner that it provides the classification of the data desired.

The collection of internal statistical data is usually closely related to the accounting system, and in many businesses the scope of the accounting system is enlarged to provide statistical data as well as strictly accounting data. The discussion of the collection of internal statistical data inevitably overlaps the field of accounting systems, for it is frequently impossible to distinguish between internal statistical data and accounting data.

External data, however, are information relating to phenomena outside the business and therefore not under the direct control of the concern collecting it. The problem of collecting external data is a distinctly different problem from that of recording internal data. Persons not under the control of the organization collecting the data must be persuaded to supply the information needed, and this fact gives rise to many problems not met in the operation of an accounting system or of a system of internal statistical records.

It has already been pointed out that the businessman wanting information on phenomena outside the organization should rely as far as possible on data already collected and available for use, rather than make the collection himself. Public and private agencies which make statistical data available in published form include federal and state governments, trade associations, banks, newspapers, business periodicals, universities, and even private business concerns. Although it is cheaper to use data already available, many problems arise which require information that has never been collected. If these facts are necessary, the only source is a statistical investigation to collect them.

The sources of statistical data on business problems, then, may be summarized as follows:

1. Internal records of the organization.
2. Governmental and private agencies that make statistical data available to businessmen.
3. Statistical surveys to collect the specific data needed.

Since these sources of data are all important to the businessman, each will be discussed in a separate chapter.

Some concerns may make use of information from all three sources given above. Others may use very little data other than that supplied by their own internal accounting and statistical records. Accurate internal records (primarily accounting records, but also statistical records) represent the basic data that are practically a necessity to a modern business. Probably most concerns make some use of published external data, but much more effective use could be made of them than is generally the case. With public agencies making a steadily increasing amount of such data available to business, it is a mistake for management not to have all the published data on a particular problem and not to take full advantage of it when making decisions.

The collection of external data by the individual business is so expensive that it is generally limited to large concerns. In most situations the cost of making a study involving the collection of original external data would be substantial. For a large concern the cost might be a relatively small percentage of total operating costs, but for a small business it could easily be prohibitive.

MANAGEMENT OF A STATISTICAL OFFICE

The amount of clerical work involved in the collection and tabulation of both internal and external statistical data on a large scale is considerable. The operation of a statistical office becomes a typical problem of office management with emphasis on speed because the results are always wanted as soon as possible. For internal data and for external surveys that recur at regular intervals, a systematic procedure must be developed to provide for a smooth flow of work through the different steps. This involves a knowledge of office procedures and the services of a good office manager, who may not be a statistician but rather a man with experience in organizing clerical work and maintaining morale in a large staff of clerks.

A survey to be made only once must be planned in detail and the procedures worked out in advance since it is difficult to introduce improvements after operations have started. The questionnaire, the size of the sample (if the study is based on a sample) and the way it is to be secured,

the work of mailing forms and receiving returns, the supervision of interviewers, and the tabulating procedures, whether by hand or machine, should be worked out step by step. In a large survey these procedures may cover a great many pages, for they should contain instructions for all situations that are likely to arise. If an unforeseen situation should arise, the method of handling it should be added to the procedure. Only by such a method can a large survey be carried out and give results that are accurate.

STATISTICAL PITFALLS TO AVOID

Because the use of quantitative methods of analysis has made possible the substitution of measurements for opinions and hunches, it is sometimes easy to place too much confidence in the information secured by statistical methods. Securing accurate information on a problem may not be merely a matter of making mathematical calculations and precise measurements. Because of the abstract nature of much of the methodology of statistics, it is possible to make serious errors in an analysis without their being apparent to the persons using the information. When a simple set of facts is observed by the person making a decision, he has a certain amount of protection against drawing incorrect conclusions. The more elaborate the analysis of the data on which he must base his conclusions, the less opportunity he has to know how reliable the information is that he is using.

Assume that the owner of a weekly newspaper in a small town is planning to hire and train a young man to take over the sale of advertising and subscriptions. Several high school graduates apply for the position. The owner has known all of them all of their lives and has watched their schoolwork, their performance on vacation jobs, and their participation in various activities in the town. The facts upon which he bases his decision to hire one of the applicants will be very different from the information that would be compiled by the personnel department of a large business.

As the scope of the operation of a business becomes greater, more reliance must be placed on compilations of statistical data; but it is well to keep clearly in mind that the improper collection or analysis of statistical data may lead the user of that data to completely wrong conclusions and can result in decisions that are seriously in error. This fact means that persons placing reliance on statistical analysis must make certain that the analysis was correctly performed, since it is frequently not possible to determine from the information that any incorrect methods were used.

A great many examples of the incorrect use of statistical data have been published,[1] but an understanding of what is wrong with the analysis usually requires some knowledge of the correct method of making the analysis. Throughout this text, the various kinds of errors that commonly arise in the use of different methods of analysis will be pointed out; but in the following paragraphs a few simple examples of fallacious reasoning will be presented as a general warning of the pitfalls that are always present in the use of statistical data. The subject of statistical pitfalls might include mistakes in recording data or in making calculations, since either would produce incorrect information. However, the subject will be limited to include improper analysis that is not immediately or easily recognizable but that leads to faulty conclusions.

Bias

One of the most prevalent pitfalls in the use of statistical analysis is an unconscious bias on the part of the analyst or the user of the data. It is almost impossible for one to be completely objective or without preconceived ideas on a subject, and there are many ways in which such a state of mind can influence the results of any collection or analysis of data. In its simplest form, *bias* means that one gives more weight to the facts that support his opinion than to conflicting data. The failure of the election polls to forecast the re-election of former President Truman in 1948 probably received more publicity than any other statistical error in history. Although the reasons for the error were complex and technical, to a certain degree the overwhelming belief on the part of most analysts that President Truman would not be re-elected made them fail to recognize the significance of the information they were analyzing. After the election, it was not difficult to see that the polls had given a strong warning that the election would be close and that any forecast would be risky.

An extreme case of bias would be a situation in which a course of action is decided upon, and then statistical analyses are used to find reasons for doing what had already been decided. Some business statisticians have been likened to the boy in school who looked up the answers to his problems in the back of the book, and who now finds reasons to support the decisions made by his boss. The cynical pronouncement that "figures don't lie but liars figure" reflects the same belief that decisions are based on preconceived ideas and statistics are used only as a pretense that the decision had a logical basis. When statistical data or the results of statistical analyses are used in selling, there is always the suspicion that the facts have been manipulated to tell the most favorable story.

[1]For a recent entertaining treatment of this subject, see *How to Lie with Statistics* by Darrell Huff (New York: W. W. Norton and Company, Inc., 1954).

Noncomparable Data

An ever-present pitfall in the use of statistical data is the failure to make certain that the data used in making comparisons are comparable in every way. Comparison of the cost of living at the present time with the cost of living fifty years ago raises questions of comparability, since many of the items in the present day budget either did not exist or were relatively unimportant fifty years ago. In comparing from one period to another the number of deaths attributed to a particular disease, the more recent year may show an increasing rate due to more accurate reporting of the causes of death; thus, the figures for two periods are not comparable.

One of the most important problems in comparing changes in economic activity over a period of time is determining whether the statistical data mean the same thing in each period. In the Census of Population prior to 1950, students and military personnel were counted in the population of their home towns; but in 1950, students and military personnel were counted in the population of the city in which they were attending school or in which they were stationed. This change in classification caused a considerable degree of noncomparability in cities with either large military establishments or educational institutions with a large number of students that were not residents of the community. Totally wrong conclusions might be drawn from the comparison of population in 1940 and 1950.

In earlier years the data on water consumption in many cities represented the amount of water pumped from the central pumping station, while in later years water consumption represented the amount of water sold, as measured by customers' meters. Since a considerable amount of water is normally lost in the distribution process, the data on total water consumption would not be comparable for the different years. It is imperative that data always be comparable or the conclusions drawn may be highly inaccurate.

Uncritical Projection of Trends

Making decisions with regard to the future is almost inevitably based on existing trends, since the only information with regard to the future must be derived from a study of past relationships. Many decisions are made by accepting a naive projection of an existing trend without making any analysis of the stability of this trend. The population of a city can be forecast by projecting the average rate of growth for the past ten years without making any study of the factors that might cause the rate of growth to change. A careful analysis of the factors involved might indicate that unusual forces had pushed the trend up at a very rapid rate in the past, but changes already taking place might foretell an entirely different rate

of growth in the future. An uncritical projection of past trends is a pitfall that has done much to discredit the use of statistical analysis. A considerable amount of space is devoted in future chapters to the techniques that can be used in making projections and the dangers of all such techniques.

Improper Assumptions Regarding Causation

Understanding the cause-and-effect relationships between events is an extremely important element in making business decisions, but securing an accurate explanation of the cause of a given event is not easy. Even if accurate statistical data are compiled to describe what happened, it is not a simple matter to determine why it happened. It is a very simple fallacy to conclude that because two events occur together one was the cause of the other. Frequently it is possible to make a case for either factor being the cause, and it may be equally likely that both events were the result of some third factor.

Many theories of the business cycle have selected a single factor and attempted to explain all fluctuations of business activity as results of the fluctuations in this one factor. For example, it has been argued that variations in the interest rate have caused the variations in business volume, with periods of decline following a rise in interest rates, and an upturn in business following a drop in interest rates. There is much evidence that this simple cause-and-effect relationship does not explain the whole phenomenon of the business cycle.

One of the simplest examples of this type of fallacious reasoning would be to measure two factors that are both related to growing population and then designate one factor as the cause of the change in the other. An obviously incorrect analysis would be to conclude that because church attendance in a growing city had increased by about the same percentage as the number of arrests, that either the increase in church attendance was the cause of the increase in arrests, or that the increase in arrests was the cause of the increase in church attendance. There probably would be little argument with the statement that both arrests and church attendance increased as the population increased and that neither was the cause of the increase in the other. However, all such relationships are not so simple. A study showed that areas of a city with the highest degree of crowding (largest ratio of persons to square feet of floor space in their homes) also showed the highest incidence of tuberculosis. The conclusion might be drawn that overcrowding was a cause of tuberculosis, although a more careful analysis might suggest that the economic factors that resulted in overcrowding were also important factors in the number of tubercular cases. At least, it would be unwise to conclude uncritically that a direct cause-and-effect relationship existed.

Comparison with an Abnormal Base

Comparisons of business and economic data with previous periods are made on the assumption that the period used for comparative purposes is approximately average or normal. If the period used as the basis for comparison is abnormal, the conclusions drawn may not be correct. It is always wise to check to be certain that the base used for comparison was not an extreme value that gave an incorrect impression.

If a salesman reported that farm income in a portion of his territory was 400 per cent higher this year than last, it would imply that business had been unusually good. However, if income last year was very low because of drought, the comparison would not be so favorable as the statement seemed to indicate. An extreme case of the use of an abnormal base is to use a very small number, such as one or two, as the base for a percentage. Very large percentage increases will occur in such cases, but they will generally not be so significant as they appear to be.

Improper Sampling

A major proportion of statistical analysis makes use of samples to draw conclusions concerning the larger group from which the sample is taken. If the sample is properly selected, it will have essentially the same characteristics as the group from which it is taken; but if the sample is not selected properly, the results may be meaningless. Since it is not easy to select a sample correctly, there is great danger that a sample study will not give accurate information. Because the methods to be used are rather technical, the discussion of sampling will be deferred to Chapter 10.

The preceding discussion has all been rather general, since specific examples of fallacious reasoning with statistical data are usually rather subtle and require a fairly complete discussion of the techniques of analysis involved. Further examples of the misuses of statistical methods will be given more detailed treatment in connection with the discussion of the different methods of analysis.

STUDY QUESTIONS

1. What does the following statement mean: "A man's decisions are no better than his information"?

2. Describe the differences in the amount and kind of information needed in the management of the two following types of eating establishments:

(a) A small neighborhood tearoom operated by the owner.
(b) A chain of restaurants operating in 35 cities in 9 states.

3. Give three examples (in addition to those in the text) of a variable for which successive values are obtained by:

(a) Counting.
(b) Measuring a given characteristic.
(c) Computing ratios.

4. If someone tells you that a certain book is a "statistics book," how would you determine whether it should properly be called a "book containing statistical data," or a "book describing statistical methods of analysis"?

5. In compiling data on the production of automobile tires, the unit most likely to be used would be the number of tires. However, technological progress in the manufacture of tires has improved so much in the past 30 years that a tire may now be expected to be usable for several times the number of miles that it could be used in the early days of the industry. This seems to mean that the unit (one tire) has not remained the same over the years. In other words, the production of 1,000 tires in 1957 has a different significance than the production of 1,000 tires in 1927, in terms of the amount of usage in the tires.

Does the fact that the unit of measurement has changed its significance over the years make it improper to show data on the number of tires produced? Can you suggest any way to take into account the change in the unit of measurement?

6. Assume that you plan to secure certain agricultural information by interviewing farm families. In determining which farm families to interview, it will be necessary in each case to decide whether the home should be classified as a farm or a nonfarm home.

Write a definition of "farm" that you would use in selecting the farm homes to be interviewed. Obviously there will be no question about many homes; for example, a family living in an apartment house is probably not a farm family. But you may also have to classify a family living at the edge of town on 10 acres, some of which are planted in crops and some in pasture on which a number of animals graze. The family members may also have jobs in town.

7. The compilation of internal data in many business organizations is delegated to the accounting department, even when the businesses have a statistical office that compiles external data for use by the management. How can you justify dividing the collection of statistical data in this way rather than concentrating all the work in the statistical office?

PROBLEMS

1. Let us say that you are working for a wholesale grocery business, and the question has come up as to whether or not it would be profitable to close out a large number of very small accounts. One group argues that the cost of keeping the records, delivering the small orders, and collecting the accounts absorbs all the profit on the sales. Another group counters with the argument that the small customers may become large customers and when they do, their business will be profitable. Neither group has any information on which to base its arguments.

You are instructed to explore the possibilities of assembling enough facts to settle the argument. As the first step in your assignment, make a list of the

various items of information that you think would be needed to make the right decision. In this problem, it is not necessary to develop methods of securing the information; the step to be taken is to decide what information would be valuable in making a decision. Later the decision can be made as to whether or not the information can be secured at a reasonable cost.

2. The question of air conditioning a business office is being considered. One of the advocates of installing air conditioning insists that the increased efficiency of the office staff would pay for the cost of the air conditioning. The annual cost of air conditioning has been compiled, but the question of measuring the increased efficiency of the office employees raises many questions.

Outline a method for measuring the increase in efficiency that could be expected as a result of air conditioning the office. The measurement should make possible the computation of the total monetary savings of the air conditioning installation.

3. A research organization that compiles rather extensive data on the business and economic conditions in a section of the United States publishes a monthly magazine in which business conditions in the region are analyzed. The data compiled by this organization are published monthly in the magazine in tabular form. For a number of years the text of the magazine has repeated in discussion form most of the data shown in the tables, in the belief that some persons would not read the tabular information but would read it if it were written out in the text.

After several years of presenting statistical data in both tabular and paragraph form, a proposal was made to eliminate from the text all repetition of the data in the tables and confine the text to analysis of business conditions. The argument presented in favor of this change was that anyone interested in the statistical data could secure it from the tables, and repeating the information in the text was needless duplication.

Write your recommendations for the policy you think should be adopted to present the statistical data collected by the organization. Support your recommendations with reasons.

Part

I ASSEMBLING STATISTICAL DATA

chapter 2 ♦ Statistical Tables

chapter 3 ♦ Internal Statistics—Records and Reports

chapter 4 ♦ Published Data

chapter 5 ♦ Collection and Tabulation of External Data

2 STATISTICAL TABLES

Facts expressed in quantitative form inevitably result in a great many numbers, and unless the facts are presented in some organized manner their significance is easily lost. Tabular form is the universal method of presenting statistical data except when a very small number of items is being presented. Table 2-1 on page 20 is a typical statistical table.

It was pointed out in Chapter 1 that statistical data are usually presented in classified form. This is so generally true that the principles of the design of statistical tables are based on the presentation of classified rather than unclassified data. This chapter will consist, therefore, of a discussion of the kinds of classifications, followed by the generally accepted rules for presenting these different types of data in tables under varying circumstances.

TYPES OF CLASSIFICATION

The number of ways in which statistical data may be classified is almost limitless, but the possible classifications can be grouped into the four following types:

1. Classifications based on differences of kind.
2. Classifications based on differences of degree of a given characteristic.
3. Geographical divisions.
4. Time series.

Differences of Kind

An analysis of the market for home equipment, such as refrigerators, would make use of a considerable amount of statistical data classified in a number of ways. The 1950 Census, for example, reported that there were 42,826,281 occupied dwelling units in the United States. The total number of occupied dwelling units is classified by the Bureau of the Census according to tenure, whether owner-occupied or renter-occupied. Each of these groups is further classified as urban, rural nonfarm, or rural farm. In Table 2-1 occupied dwelling units are classified according to the kind of refrigeration reported. The classes of refrigeration are "mechanical," "ice," and "other or none." These classes accounted for all but 766,431 dwellings, for which no information was available as to the kind of refrigeration.

Table 2-1

REFRIGERATION IN UNITED STATES DWELLING UNITS, 1950

CLASSIFIED BY TYPE

Type of refrigeration	Number of dwelling units
Mechanical	33,720,250
Ice	4,445,335
Other or none	3,894,265
Total reporting	42,059,850
Not reported	766,431
Total occupied dwelling units	42,826,281

SOURCE: Department of Commerce, Bureau of the Census, *U.S. Census of Housing*: *1950*, Vol. I, Part 1.

The above classification is based on *differences of kind.* There are different kinds of refrigeration, and these different kinds become the classes into which the data are grouped. The classes are set up on the basis of *qualitative* differences.

The following is a list of some typical examples of classifications of business data based on differences of kind:

Internal Data:

1. Sales classified by products.
2. Sales classified according to terms of sale: cash, charge, C.O.D.
3. Sales classified by salesmen.
4. Employees classified by type of work performed.
5. Customers classified according to credit rating.

External Data:

1. Retail sales classified according to type of store: chain, independent, and other.
2. Retail sales classified according to kind of business.
3. Population classified by race.
4. Manufactures classified by industries.
5. Exports classified by commodities.

Differences of Degree of a Given Characteristic

This type of classification is illustrated by Table 2-2, which gives the owner-occupied urban dwelling units in the United States classified

according to value of homes. The 11,160,046 one-family owner-occupied urban dwelling units are classified on the basis of different degrees of a particular characteristic. The classes set up are varying degrees of value, and all the dwelling units valued between certain amounts are put into the same class.

Table 2-2

ONE-FAMILY OWNER-OCCUPIED URBAN DWELLING UNITS IN THE UNITED STATES, 1950

CLASSIFIED BY VALUE

Value of dwelling units	Number
Less than $2,000	445,674
$2,000 to $2,999	397,101
$3,000 to $3,999	597,464
$4,000 to $4,999	696,145
$5,000 to $5,999	873,181
$6,000 to $7,499	1,612,806
$7,500 to $9,999	2,178,400
$10,000 to $14,999	2,576,673
$15,000 to $19,999	872,724
$20,000 or more	614,851
Not reported	295,027
Total dwelling units	11,160,046

SOURCE: Department of Commerce, Bureau of the Census, *U. S. Census of Housing: 1950*, Vol. I, Part 1.

Classifications based on differences of degree are sometimes referred to as *quantitative* classifications, since the most common method of expressing differences of degree is in quantitative terms. In some cases it may be found that differences of degree are expressed in qualitative language rather than in quantitative terms, but this is not done frequently and it is usually not satisfactory. For example, employees might be classified according to ages — the classes being young, middle-aged, and old. The superiority of using quantitative classes is obvious in this case.

It is frequently difficult to distinguish between classifications based on differences of kind and those based on differences of degree if the classes are expressed qualitatively rather than quantitatively. For example, accounts receivable may be classified as good, doubtful, and bad. This might be called a classification based on degrees — degrees of certainty as to whether the accounts will be paid. Or it might be called a classification

based on differences of kind, the different classes being different kinds of accounts. Actually it makes little practical difference which term is used, since all such classifications must be treated in much the same way as classifications based on differences of kind. This is merely another way of saying that, for all practical purposes, differences of degree may be considered as being expressed as quantitative differences.

The following are some typical classifications of business data based on differences of degree of a particular characteristic:

Internal Data:

1. Employees classified according to earnings.
2. Employees classified according to age.
3. Customers classified according to amount of annual purchases.
4. Accounts classified according to age.
5. Sales invoices classified according to amount of sale.

External Data:

1. Retail stores classified according to annual sales volume.
2. Factories classified according to number of employees.
3. Business concerns classified according to amount of capitalization.
4. Employees of an industry classified according to earnings.

When the data classified on the basis of differences of degree consist of the number of units in each class, the classification is called a *frequency distribution*. It may be said, therefore, that a frequency distribution shows the frequency of occurrence of different values of a single variable. Table 2-2, for example, shows the frequency of occurrence of dwelling units valued at less than $2,000 to be 445,674 for the whole United States. The frequency of occurrence of this and other values is shown in the column headed "Number." Two other examples of frequency distributions are the two columns in Table 2-7 (page 28) showing the number of places in urban territory in the United States. It will also be noted that the two columns showing the population of places in urban territory are not frequency distributions.

Some statisticians insist that a frequency distribution must have equal class intervals. It is obvious that comparisons of classes can be made more accurately when the intervals are all the same size, but many types of business data cannot be put into equal groups. The setting up of classes for a frequency distribution is discussed later in this chapter.

Table 2-3 shows the classification of occupied dwelling units in the United States in 1950 according to the number of persons living in them. This is another example of a quantitative classification. A discussion of a significant difference between the data in Tables 2-3 and 2-8 is given on pages 29 to 31.

Table 2-3

OCCUPIED DWELLING UNITS IN THE UNITED STATES, 1950

CLASSIFIED BY NUMBER OF OCCUPANTS

Number of occupants	Number of occupied dwelling units
1 person	3,993,399
2 persons	12,023,065
3 persons	9,762,531
4 persons	7,878,063
5 persons	4,466,292
6 persons	2,257,898
7 persons	1,159,544
8 persons	594,365
9 persons	324,224
10 persons or more	366,900
All occupied dwelling units	42,826,281

SOURCE: Department of Commerce, Bureau of the Census, *U. S. Census of Housing: 1950*, Vol. I, Part 1.

Geographical Divisions

The third type of classification is the separation of the data into groups according to different *geographical divisions*: continents, countries, cities, or other subdivisions. The data in the preceding tables are for the whole United States, but in Table 2-4 the 42,826,281 occupied dwelling units in the United States are classified according to their geographical location. In addition, the percentage of dwelling units with mechanical refrigerators is given for each of the geographical divisions, as well as for the United States.

In Table 2-4 the nine census geographical divisions of the United States are used, but each of these divisions can be classified further into states, and the states into counties. Since much of the data of use to businessmen needs to be classified according to geographic divisions, this type of classification is very important.

Time Series

The last type of classification is the separation of the data into groups on the basis of *time intervals*. According to Table 2-1 there were 42,826,281 occupied dwelling units in the United States in 1950. When the number of

Table 2-4

NUMBER OF OCCUPIED DWELLINGS AND PERCENTAGE WITH MECHANICAL REFRIGERATORS, 1950

CLASSIFIED BY GEOGRAPHICAL DIVISIONS

Geographical divisions	Occupied dwelling units	Percentage of dwelling units with mechanical refrigerators[1]
New England	2,613,421	86.2
Middle Atlantic	8,614,655	90.1
East North Central	8,823,056	86.2
West North Central	4,148,910	78.9
South Atlantic	5,539,058	69.5
East South Central	2,991,832	60.4
West South Central	4,101,717	69.5
Mountain	1,445,433	77.1
Pacific	4,548,199	84.0
United States	42,826,281	80.2

[1] Based on a 20 per cent sample.

SOURCE: Department of Commerce, Bureau of the Census, *U. S. Census of Housing: 1950*, Vol. I, Part 1.

dwellings for a sequence of dates is compiled, the resulting series of data is called a *time series*. Table 2-5 shows the number of dwelling units in the United States at each census date from 1890 to 1950. These amounts

Table 2-5

OCCUPIED DWELLING UNITS IN THE UNITED STATES: 1890 TO 1950

Year	Number of units
1890	12,690,152
1900	15,963,965
1910	20,255,555
1920	24,351,676
1930	29,904,663
1940	34,854,532
1950	42,826,281

SOURCE: Department of Commerce, Bureau of the Census *U. S. Census of Housing: 1950*, Vol. I, Part 1.

represent the result of a count made every ten years. The data in Table 2-6 are also classified according to time intervals, but in a somewhat different way. The total number of dwelling units built in nonfarm areas during the ten years 1947 to 1956 are classified according to the year in which they were started. These two kinds of time series will be discussed further in Chapter 7; for the purposes of this discussion it is sufficient to note that they both represent classifications based on time.

Table 2-6

NEW PERMANENT DWELLING UNITS STARTED IN NONFARM AREAS: 1947 TO 1956

Year	Number of units
1947	849,000
1948	931,600
1949	1,025,100
1950	1,396,000
1951	1,091,300
1952	1,127,000
1953	1,103,800
1954	1,220,400
1955	1,328,900
1956	1,118,100

SOURCE: Department of Labor, Bureau of Labor Statistics.

PRESENTING STATISTICAL DATA

In this section the problems of setting up the classes for the different types of classifications will be discussed. This is the basic problem of presenting statistical data; once the classes have been defined accurately, the problem of separating the items into the classes is usually a routine clerical operation.

Defining the Classes

For qualitative classifications, the problem of setting up classes varies greatly from one situation to another. Sometimes the classes are generally recognized categories, but in other cases they must be carefully worked out. Classification of products by brand name, for example, is relatively simple. If consumers are asked the make of automobile owned, each make is a separate class. On the other hand, in classifying business concerns by

industry or by commodities produced, the classes may not be self-evident. A great deal of work has been done by government statisticians in developing standard commodity and industry classifications so that classified data published by different agencies can be compared or combined. If each agency uses its own grouping, the classified data of one cannot easily be compared with the classified data of others.

The most common classes for geographical distributions are political divisions. World data are ordinarily classified by countries, and data for the United States by states. If further classification of data for individual states is desired, counties are commonly used. Many statistical data are classified by cities; and political subdivisions, such as city wards, are sometimes used.

In recent years certain special geographical classifications have been used by statisticians. Prior to the 1950 Census of Population the governmental agencies used a number of these special areas, each defined in somewhat different ways for different purposes. Examples of these geographical areas were the metropolitan districts of the Census of Population, the industrial areas of the Census of Manufactures, and the labor market areas of the Bureau of Employment Security. The usefulness of data published for any of these geographical classifications was limited by the lack of comparability with other similar areas. This led to agreement among the various federal agencies on the establishment of the *standard metropolitan area* so that a wide variety of statistical data might be presented on a uniform basis. The general concept applied was that an area should be an integrated economic and social entity, with an attendant large volume of daily travel and communication between the central city and the outlying parts of the area. There were 168 standard metropolitan areas designated in the United States in 1950.

Another special geographical classification is the system of state economic areas set up by the Bureau of the Census. The state economic areas are relatively homogeneous subdivisions of states. An economic area consists of single counties or groups of counties with similar economic and social characteristics. The boundaries of these areas have been drawn in such a way that each state is subdivided into relatively few parts, with each part having certain significant characteristics which distinguish it from adjoining areas.

Time series are classified on the basis of the recognized time units. Years, quarters, months, weeks, days, and hours may be used as classes. Since the classes are already set up, the only problem, normally, is choosing which interval to use. There are, however, several undesirable features of the calendar from the viewpoint of the statistician. This matter will be discussed in Chapter 11.

Principles for Setting Up the Classes for a Frequency Distribution

Setting up classes for a quantitative distribution offers some problems not encountered with other types of classifications. There are usually no natural divisions into which the figures tend to fall, although sometimes there is a tendency for the values to be concentrated at specific amounts such as the numbers ending in five and zero. The fact that the data do not themselves suggest the classes into which they should be divided makes the problem of setting up the groups of a quantitative classification an important one. The following are some of the principles that should be considered.

1. *The number of classes should not be so large as to destroy the advantages of summarization, yet condensation must not be carried too far.* The classes must provide enough groups to show the chief characteristics of the data; at the same time, they must not be so numerous that it is difficult to grasp the picture of the distribution as a whole.

The correct number of classes depends somewhat upon the number of figures to be classified. A small number of items to be classified justifies a small number of classes. If a hundred items are classified into 25 classes, there are likely to be many groups with no frequencies or with a very small number. On the other hand, 25 classes might not be too many for several thousand items. A rule that is often cited is that an interval that casts the variable into 15 to 25 classes is most satisfactory.

2. *Whenever possible, all classes should be the same size.* In analyzing a quantitative classification it is a decided advantage to have all the classes the same size, although when the items being classified contain a few extremely large or extremely small items it is usually impossible to set up equal class intervals. Table 2-7 is an example of a classification of this kind. Here the population of places ranges from 2,500 up to one million or more. If the data were divided into 20 equal classes, the limits would have to be 0 and under 50,000; 50,000 and under 100,000; etc. — or something very close to this. For 1950, such equal classes would bring 3,788 of the items in the first class, 128 in the second, and leave only 107 to be distributed among the remaining 18 classes. Most of the information wanted about the distribution would be hidden in the single group of 3,788; in other words, a much more detailed breakdown of this group is needed, which means that the class intervals below 100,000 should be much smaller than is needed for that part of the distribution above 100,000. The rule that class intervals in a frequency distribution should be equal if possible does not mean that they should be equal if the characteristics of the data can be shown better by using unequal classes.

Table 2-7

NUMBER OF PLACES AND POPULATION IN URBAN TERRITORY IN THE UNITED STATES: 1940 AND 1950

Size (Population)	1940		1950	
	Number of places	Population	Number of places	Population
Places of 2,500 and under 5,000	1,422	5,025,911	1,570	5,565,269
Places of 5,000 and under 10,000	965	6,681,894	1,133	7,878,675
Places of 10,000 and under 25,000	665	9,966,898	814	12,467,229
Places of 25,000 and under 50,000	213	7,417,093	271	9,495,862
Places of 50,000 and under 100,000	107	7,343,917	128	9,073,363
Places of 100,000 and under 250,000	55	7,792,650	66	9,614,111
Places of 250,000 and under 500,000	23	7,827,514	23	8,241,560
Places of 500,000 and under 1,000,000	9	6,456,959	13	9,186,945
Places of 1,000,000 or more.....	5	15,910,866	5	17,404,450
Total urban territory.....	3,464	74,423,702	4,023	88,927,464

SOURCE: Department of Commerce, Bureau of the Census, *U. S. Census of Population: 1950*, Vol. I.

When the extremes of the distribution are a long distance from the center, it may be necessary to set up an *open-end class*, as in Tables 2-2 and 2-7. Table 2-7 shows that there were five places with a population of more than one million. Table 2-2 shows that there are 614,851 homes valued at $20,000 and over. The obvious advantage of using such a class is that it eliminates the need for a large number of classes of which some show only a few frequencies and many show none. Just as obviously, the disadvantage is that there is no way of telling how large the items in this open-end class are unless a special note is made of the facts at the time of tabulation. In making a tabulation it helps a great deal to give in the table the total amount of the items going into an open-end class. This procedure removes the most striking disadvantage of this class, but many classifications are published without this information.

3. *The position or starting point of the intervals is, except under special conditions, more or less a matter of indifference.* Sometimes the classes are so chosen that the mid-values are integers. When the values tend to be stated in round numbers, as for example, in hundreds, tens, or fives, the data are concentrated at regular intervals, and the classes should be set up so as to make the round numbers fall, as nearly as possible, *in the middle of the classes.* It is also advisable, if possible, that the same number of concentration points should fall in each of the classes of the same size. There may be many times when this cannot be accomplished, but the frequency distribution will represent the data more accurately if this can be done.

4. *Class limits should be determined accurately and the classes should be given appropriate designations.* The determination of class limits needs to take into consideration the nature of continuous and discontinuous variables. Quantities that can take on any value within a given range are called *continuous variables,* while quantities that occur only at certain values are called *discontinuous,* or *discrete, variables.* Examples of discontinuous variables are size of families, number of rooms in a house, and gasoline tax rates in cents per gallon. Various types of measurements and ratios may have any value, although they are rounded to some desired number of places, depending on the degree of accuracy desired in a particular situation. When the values occur in steps that are small in comparison with the total range of the variable, the data are for practical purposes continuous, even though in reality they are discontinuous. Monetary values are usually expressed only in whole cents, but when the amounts are large the steps between the values are so small that the data are essentially continuous.

For discontinuous data little choice is possible in the determination of the class limits, once the width of the class is decided upon. If each value is in a different class, the value itself designates the class, as in Table 2-3 for the number of persons. When several values of the variable are included in one class, the class limits are designated by stating the largest and the smallest units in the class. For example, if business concerns are classified according to number of employees, the classes would be stated as follows:

1 to 5
6 to 10
11 to 15, and so on.

When the variable is continuous, however, the designation of the class limits cannot always be accomplished so simply. The gross profit as a percentage of total net sales, as shown in Table 2-8, is a variable that does not fall on certain values to the exclusion of the fractional values between. Some percentages may come out even, but most of them can be carried out to any desired number of decimal places. In computing the ratios of

Table 2-8

GROSS MARGINS OF 238 LUMBER YARDS, 1956

RATIO OF GROSS PROFIT TO NET SALES
(Per cent)

Gross margin (Per cent of net sales)	Number of yards
10.0–12.4	2
12.5–14.9	9
15.0–17.4	16
17.5–19.9	37
20.0–22.4	43
22.5–24.9	52
25.0–27.4	35
27.5–29.9	20
30.0–32.4	18
32.5–34.9	3
35.0–37.4	2
37.5–39.9	1
Total	238

SOURCE: Reports on file in the Bureau of Business Research, The University of Texas.

gross profit to total net sales, the results were rounded off to the nearest tenth of 1 per cent. In other words, .3 per cent might have originally been .33 per cent or it might have been .29 per cent. If a ratio should be exactly .25, it could be rounded to .3 or to .2 with equal logic. The best practice is to round alternately to the larger and to the smaller amounts, and one way to do this is to round to the even number. However, many people always round one half to the larger value. Which rounding practice is used will affect to a slight degree the real limits of the classes.

Since the data in Table 2-8 are expressed to one decimal place, the class limits can be stated clearly as follows:

10.0–12.4
12.5–14.9
15.0–17.4, and so on.

In making the tabulation, the real limits of the class are not a problem, since a ratio of 12.4 is put in the first class and the next largest ratio, which is 12.5, belongs in the next class.

It is clear, however, that the real upper limit of the first class is not 12.4, since ratios larger than 12.4 would be rounded to 12.4. Where the real limit falls depends upon how the data are rounded. If 12.45 were always rounded to 12.5, the limits of the class would be 10.0 and under 12.45. Obviously 12.45 is the beginning of the next class, and no ratios of this size belong in the first class. The class limits under these conditions are:

10.0 and under 12.45
12.45 and under 14.95
14.95 and under 17.45, and so on.

If 12.45 were rounded to 12.5 and to 12.4 alternately, the real limits of the class are 10.0 to 12.45, including 12.45 in the first class. But since 12.45 is also rounded to 12.5, the percentage 12.45 is also in the second class. In other words, the class limits in this situation are:

10.0–12.45
12.45–14.95
14.95–17.45, and so on.

The real class limits overlap and this method of locating the class limits is ambiguous, since 12.45 may sometimes be rounded to 12.5 and sometimes to 12.4.

It should be pointed out again that the classes for continuous data are usually stated by giving the smallest and largest rounded amounts that are put in each class, just as the classes for discontinuous data are stated by giving the smallest and largest amounts in each class. In other words, tabulations of continuous data usually do not state the real limits of the classes. However, occasions will arise in the analysis of the data when it is important to use the real limits in further calculations. For example, a considerable saving in time is possible in computing averages and certain other statistical measures by using the assumption that the average value of the items in a class is equal to the mid-value of the class. This mid-value is computed by averaging the *real upper and lower limits* of the class.

Cumulative Classifications

The classifications discussed in the preceding pages show the amounts in each class, but for some classifications it is useful to cumulate the amounts. Table 2-9 gives the sales of Sears, Roebuck & Company for each month of 1956, and also the total sales from January 1 to the end of each month. Each figure in the Cumulative column represents the total sales from the beginning of the year to end of the month represented by the particular line in the table. Sales during April, 1956, were $280,424,000 and sales

Table 2-9

SEARS, ROEBUCK & COMPANY SALES: 1956 AND 1957

MONTHLY AND CUMULATIVE THROUGHOUT EACH YEAR
(Thousands of dollars)

Month	1956		1957	
	Monthly	Cumulative	Monthly	Cumulative
January	228,084	228,084	246,908	246,908
February	217,628	445,712	223,750	470,658
March	265,612	711,324	269,815	740,473
April	280,424	991,748	307,394	1,047,867
May	317,556	1,309,304	338,262	1,386,129
June	328,976	1,638,280	335,812	1,721,941
July	276,030	1,914,310	308,538	2,030,479
August	326,855	2,241,165	344,491	2,374,970
September	310,817	2,551,982	314,876	2,689,846
October	327,558	2,879,540	329,811	3,019,657
November	362,433	3,241,973	344,687	3,364,344
December	444,955	3,686,928	...	...

SOURCE: *Survey of Current Business*, February, 1957, and December, 1957.

from January 1 to April 30 appear in the next column on the April line as $991,748,000. As each month's sales are entered in the table, the cumulative figure can be obtained by adding that month's sales to the previous cumulative figure.

A cumulative frequency distribution can be derived in much the same manner as a cumulative time series. Tables 2-10 and 2-11 show cumulative frequency distributions obtained from the data in Table 2-2. Table 2-10 shows the number of homes less than each of a series of values. There were 455,674 one-family owner-occupied urban dwelling units in the United States valued at less than $2,000; 842,775 at less than $3,000; and so on through the distribution. It will be noted that the number of dwellings in each case includes all the dwellings in the preceding classes. If one wants to know how many are valued at less than $7,500, for example, he can see immediately that the number is 4,622,371.

If the chief interest is in the number of homes of more than a given value, the cumulative distribution can be made on a "more than" or "or more" basis instead of a "less than" basis. Table 2-11 is based upon the same frequency distribution, but it shows the number of homes at a given value or more.

Table 2-10

ONE-FAMILY OWNER-OCCUPIED URBAN DWELLING UNITS IN THE UNITED STATES, 1950

NUMBER OF HOMES LESS THAN A GIVEN VALUE

Value of homes	Number
Less than $2,000	445,674
Less than 3,000	842,775
Less than 4,000	1,440,239
Less than 5,000	2,136,384
Less than 6,000	3,009,565
Less than 7,500	4,622,371
Less than 10,000	6,800,771
Less than 15,000	9,377,444
Less than 20,000	10,250,168
Total reporting value	10,865,019
Total dwelling units	11,160,046

SOURCE: Department of Commerce, Bureau of the Census, *U. S. Census of Housing: 1950*, Vol. I, Part 1.

Table 2-11

ONE-FAMILY OWNER-OCCUPIED URBAN DWELLING UNITS IN THE UNITED STATES, 1950

NUMBER OF HOMES OF A GIVEN VALUE OR MORE

Value of homes	Number
$20,000 or more	614,851
15,000 or more	1,487,575
10,000 or more	4,064,248
7,500 or more	6,242,648
6,000 or more	7,855,454
5,000 or more	8,728,635
4,000 or more	9,424,780
3,000 or more	10,022,244
2,000 or more	10,419,345
Total reporting value	10,865,019
Total dwelling units	11,160,046

SOURCE: Department of Commerce, Bureau of the Census, *U. S. Census of Housing: 1950*, Vol. I, Part 1.

CONSTRUCTION OF TABLES

A number of tables have already been presented in the text, and the advantages of the tabular form of presenting statistical data have been pointed out. In the following pages the general principles of good table construction will be discussed, and a number of examples of various kinds of tables will be presented.

General-Purpose and Special-Purpose Tables

If the purpose of a table is merely to make figures available, it is called a *general-purpose* or *repository* table. The figures in most of the tables published by departments of the federal government are of this kind. The government collects data of general interest without trying to make the information applicable to any particular problem. The purpose in publishing the tables is to make as much information available to as many people as possible.

If a table is to show relationships between different groups of figures, it is called a *special-purpose* or *analytical* table. When a special analysis has been made of data, the information should be presented in such a way that the important parts of the analysis are easy to understand. Table 2-12 is a special-purpose table, giving an estimate of potential traffic on the Welland Canal and the St. Lawrence Seaway by direction of movement. This is a typical special-purpose or analytical table. Table 1-1 (page 8) on the other hand, is a general-purpose table from the Census; the data were not collected as part of an analysis. The sole purpose of the data in Table 1-1 is to make information available for anyone who may be able to use it.

Cross Classification

Data classified into groups may be further classified by dividing each class into subclasses. The classification into groups and subgroups is a particularly valuable device for showing significant relationships, and is in reality a rather advanced step in the analysis of the data.

The number of cross classifications possible with a given amount of data is very large, but care must be exercised to see that the division of each class into subclasses will give valuable information. The purpose of classification is to reduce the amount of detail by putting similar items together; subdividing the data into too many groups and subgroups defeats the purpose of classification and may result in more classes than can be readily used.

An example of cross classification appears in Table 2-12. The traffic was classified as St. Lawrence and Welland; and, within these two groups, the amount of traffic, in short tons, was presented for 1954 and the future.

The comparisons most easily made in this table are between 1954 and the future, for each of the two waterways. It is also possible to compare the amount of St. Lawrence traffic with that of Welland for 1954, and for the future; however, one column of figures comes between the columns to be compared. Obviously it is easier to make comparisons between adjacent columns than between columns not adjacent, and if this latter comparison were the most important, a rearrangement of the columns would be desirable. If possible, the columns for which the primary comparison is to be made should be placed adjacent in an analytical table. A further use of cross-classification tables in connection with correlation analysis is discussed in Chapter 17.

Table 2-12

ESTIMATE OF POTENTIAL TRAFFIC ON THE WELLAND CANAL AND THE ST. LAWRENCE SEAWAY BY DIRECTION OF MOVEMENT COMPARED TO 1954 TRAFFIC

(IN THOUSANDS OF SHORT TONS)

	St. Lawrence		Welland	
	1954	Future	1954	Future
Downbound:				
Coal and coke	1,382	3,200	4,993	4,171
Grain	4,104	8,200	5,304	10,200
Iron ore	...	...	2,119	...
Petroleum and products	207	50	1,582	1,687
Wood pulp and pulpwood	...	...	191	114
Other cargo	970	6,651	929	5,115
Total downbound	6,663	18,101	15,118	21,287
Upbound:				
Coal and coke	100	1,000	5	86
Iron ore	298	20,000	180	19,000
Petroleum and products	917	1,041	714	475
Wood pulp and pulpwood	537	1,165	380	990
Other cargo	1,122	3,225	1,117	2,580
Total upbound	2,974	26,431	2,396	23,131
Total	9.637	44,532	17,514	44,418

SOURCE: "The Effects of the St. Lawrence Seaway on Grain Movements" by Joseph R. Hartley. Indiana Business Report, No. 24. Bureau of Business Research, Indiana University School of Business, 1957.

In recent years the word *breakdown* and the plural form *breakdowns* have come into rather wide use among persons working with statistical data. The word is not found in the standard books on statistical methods, but it is used widely in certain fields of applied statistics, such as marketing research. *Breakdown* is used to refer to classification and often cross classification of data, and its meaning is derived from the fact that classification is a process of breaking a large mass of observations into groups. While the terms classification and cross classification seem to be adequate to describe this process, the continual use of the term breakdown as a synonym makes it desirable to call attention to its use, if not to recommend it.

Dropping Unnecessary Figures in Special-Purpose Tables

If the figures being used in a special-purpose table are large, consideration should be given to the possibility of reducing the size of the table and improving its readability by dropping a number of places from the figures. Table 2-13 is an example of the use of this device. Dropping six digits from each figure did not impair the usefulness of the data for the purpose at hand, and it did simplify the table considerably.

Table 2-13

TRANSFERS OF GOLD AND DOLLARS TO FOREIGNERS[1], 1956

(NET TRANSFERS FROM, OR TO (—), THE UNITED STATES, IN MILLIONS OF DOLLARS)

Item	Jan. – Mar.	Apr. – June	July – Sept.	Oct. – Dec.[P]
Net transfers, total	497	268	530	–381
Net gold sales	–5	–95	–155	–25
Net increase in dollar holdings	*502*	*363*	*685*	*–356*
Deposits	–24	290	308	70
U. S. Govt. securities:				
Short-term	639	–4	137	–410
Long-term	–251	60	108	–51
Other[2]	138	17	132	35

[P]Preliminary.

[1]Includes international institutions.

[2]Represents principally bankers' acceptances and short-term commercial paper.

SOURCE: *Federal Reserve Bulletin*, Vol. 43, No. 3 (March, 1957).

MAJOR PARTS OF A TABLE

The major parts of a table are the *title*, *main body*, and *source*. Within the main body of the table the data are arranged in columns and rows. The heading of a column is called a *caption*, and the explanation of the material in a row is called a *stub*. The caption applies to all the material in a column, and the stub applies to all the figures in a row. If some figure in a column or row is exceptional and does not belong under the caption or stub, a footnote may be used to describe the exact situation. This device is used in Table 2-15 (page 42).

In Table 2-12 the captions "St. Lawrence" and "Welland" are main captions, each with two subcaptions under it. "Downbound" is a main division in the stubs, and "Coal and coke" and "Grain" are two of the subdivisions under it. The stubs are arranged to bring related items close together, and the main divisions and subdivisions help accomplish this purpose.

A more detailed discussion of the different parts of a table is given in the following pages.

Title

For the sake of convenience in finding the information wanted, a table should have a complete title. When more than one table is presented, it is desirable that each table should be numbered or lettered.

The characteristics of a good title are similar to those of a good newspaper headline. The title should tell what the table contains, yet it should be as brief as possible. It should not be necessary to make any reference to the table itself to find out what is in the table. Frequently it is possible to give the most important facts in the main title and information of secondary importance in a subtitle. Table 2-13 shows a table with a title and subtitle. When information is put in the title, it must apply to the whole table. Only statements that do apply to the table as a whole belong in the title, except as the application of the title may be modified by the use of footnotes. This is discussed further on in the chapter.

The rule that a title should give a complete description of what is included in the table applies especially to repository tables. When one is looking for information in a volume of the Census, for example, it is essential that the titles listed in the table of contents tell exactly what facts are contained in each table. Or, if one is thumbing through a volume of statistical data, it should be necessary to read only the titles of the tables, rather than to look through each table, to see whether or not the information wanted is there.

When a table contains a special analysis and is part of a report, the above rule is not always observed. When such a table would require an abnormally long title, it frequently seems desirable to explain the table in the discussion of the report and use only a brief (and incomplete) title for the table itself. There are valid reasons for condemning this practice, but there are also practical advantages that give strong justification for it; and, as a matter of fact, the practice is widely followed.

Table 2-15 (page 42) shows an abbreviated title used in the *Federal Reserve Bulletin.* The article gave the results of a survey of consumer finances, and the discussion in the article explained in detail the meaning of the information in the table. Consumers were asked their opinion of the wisest thing to do with money not needed for expenses, and the answers were tabulated in Table 2-15. When the table is used in its context, the title is adequate; but if this title were listed in a table of contents at the beginning of a book, it probably would not give all of the information needed to know what was included in the table. If good judgment is used in constructing tables with abbreviated titles, the practice probably improves the readability of a statistical analysis.

The units in which the data in the table are expressed should be indicated in the title unless different rows or columns contain data in different units. In the latter case, the unit in which each column or row is expressed should be indicated in the caption or stub. Table 2-14 shows data in different units in different rows. In this table every row shows the unit in which the figures are expressed.

Table 2-14

AVERAGE PRICES RECEIVED BY FARMERS FOR SELECTED VEGETABLES: 1952 AND 1953

Commodity	Unit	Dollars per unit	
		1952	1953
Sweet corn	5 doz. ears (approx. 50 lb.)	2.85	2.75
Lettuce	crate (approx. 70 lb.)	3.15	3.05
Onions	sack (approx. 50 lb.)	2.31	.79
Potatoes	bu. (60 lb.)	2.44	1.01
Broccoli	crate (approx. 42 lb.)	3.59	3.25
Asparagus	crate (approx. 30 lb.)	4.03	3.86

SOURCE: Department of Agriculture, *Agricultural Statistics: 1954.*

Main Body: Columns and Rows

Since the columns and rows are the main body of the table, the arrangement of data into columns and rows is the central point in table construction. Simple tables ordinarily consist of one or two columns. In a simple classification of sales over a period of years, for example, the data would probably be put in the following form:

Year	Sales
1950	$4,567,000
1951	4,106,000
1952	5,207,000
1953	5,165,000
1954	4,947,000
1955	5,580,000
1956	6,098,000
1957	6,734,000

If the total sales of each year were to be broken down into domestic and foreign sales, two more columns would be added, one for each subdivision. The years in the stubs would serve for all three columns, and the word "Sales" would be used as a main caption over all three columns showing sales, with a subcaption for each individual column.

Year	Sales		
	Total	Domestic	Foreign
1950	4,567,000	2,460,000	2,107,000
1951	4,106,000	2,206,000	1,900,000
1952	5,207,000	3,061,000	2,146,000
1953	5,165,000	3,444,000	1,721,000
1954	4,947,000	3,054,000	1,893,000
1955	5,580,000	3,380,000	2,200,000
1956	6,098,000	3,598,000	2,500,000
1957	6,734,000	3,887,000	2,847,000

In most cases more classes can be put in the rows than in the columns, since a column requires more space than a row. In the above arrangement, it is obviously more economical to put the years in the stubs instead of running them across the page in the captions.

Another equally good reason for arranging the table in this form is the fact that it is easiest to make comparisons between figures when they are

arranged in a column. It seems natural to the eye to read down a column of figures and not quite so easy to read across a row. This means that, if possible, the items to be compared should be placed in the same column rather than in the same row. In the preceding table, it is somewhat easier to compare sales for 1957 with those for 1954 than to compare domestic sales and foreign sales for any year.

After the most important positions in the table have been assigned, comparisons will still have to be made between figures in columns and rows separated from each other. If one knows exactly what comparisons will be most important, it is often possible to make further adjustments in the table that will facilitate these comparisons. At least, it is possible in so far as the data will go into the space available in the form wanted. On numerous occasions the data must be put into a table in whatever form there is room for, without too much attention being paid to the most logical form.

After the distribution of the classes into rows and columns has been made, there is still the problem of arrangement of the classes both vertically and horizontally. This problem is different for the four different types of classifications and is best discussed for each one separately.

Classifications based on differences of kind. Sometimes a classification based on differences of kind will have in itself some logical basis for arranging the classes. If such a method of arrangement cannot be found, the classes may be arranged either alphabetically or in order of size. If the number of classes is small, there is very little to be said for an alphabetical arrangement, since the chief reason for this form is the ease with which the information wanted can be found. When there are only a few classes, it is easy to find the facts whatever the form of arrangement.

The most general form of arrangement is probably by order of size, with the largest class first. When there is a miscellaneous class, it is customarily placed last. For some classifications there has developed a customary method of arrangement that is always followed. One example of this is that a classification of sexes is set up as male, female, rather than female, male.

The principles pertaining to the location of the total are discussed in the next paragraph and apply here without any changes.

Classifications based on differences of degree of a given characteristic. Frequency distributions are always arranged in sequence, and it seems customary to put the smaller figures of the distribution at the top of a column or at the left of a row. The total is nearly always given with a frequency distribution, but there is some question as to its proper placement in relation to the individual items. The popular conception of the proper place for the total is at the bottom of a column of figures or at the

right of a row. However, if the total is the most important item of information in the distribution, the logical place for it is at the top of a column or at the left of a row. In all the frequency distributions in this book, the total is placed at the bottom of the column or at the right of a row of frequencies.

Geographical divisions. For geographical classifications there is no such natural order of arrangement as there is for frequency distributions and time series. One of the most satisfactory methods of arrangement is alphabetical. This may be arbitrary and without logic, but it nevertheless makes easy the location of any division wanted. In a reference table an alphabetical order of arrangement has many advantages.

The Bureau of the Census has a special classification of the states, grouping them by divisions and proceeding from north to south and from east to west. This probably comes nearer to being standard than any other grouping, although many individual organizations have their own order of arrangement. Sometimes the most logical arrangement is by order of size, although this is probably not used so much for geographical distributions as for classifications of kind.

The total should be shown if it is significant in a geographical distribution, and it usually is. The principles to follow in the location of the total are the same as for frequency distributions.

Time series. The arrangement of a time series in a table will normally be in chronological order. The only question that calls for decision is whether the current year should come first or last. Since we read from left to right and from top to bottom, the first figure in a column and the left-hand figure in a row will usually be seen first. Therefore a great many general-purpose tables published by the various departments of the federal government place the current year at the top of a column or at the left of a row, because the government statisticians believe the current figures are most interesting and used most often and should therefore be in the most accessible positions. A reason for arranging the years with the current year at the bottom or the right, however, is that people are accustomed to read a time sequence from first to last, and a reversal of this order is not natural. Both methods of arranging time series are used, and it does not seem worth while trying to settle the question here. It will be observed in Table 2-6 that the most recent year is at the bottom of the column.

Footnotes. Since the descriptive labels in a table must cover a group of figures, there may be exceptions to which the general title, caption, or stub does not apply. If this is the case, the use of a footnote explaining the exception would allow the general heading to stand; the footnote would call attention to the fact that the heading does not apply completely as it appears to do.

Table 2-15

INVESTMENT PREFERENCES

PERCENTAGE DISTRIBUTION OF SPENDING UNITS[1]

Preferred investment	1953	1952	1951	1949
Assets of fixed money value	65	65	69	79
Savings accounts	20	16	13	18
Savings bonds	38	43	49	54
Accounts or bonds	7	6	7	7
Assets of fluctuating money value	24	26	23	11
Common stock	9	8	6	2
Real estate	14	17	16	9
Common stock or real estate	1	1	1	(2)
Other[3]	9	6	5	8
Not ascertained	2	3	3	2
All cases	100	100	100	100
Number of cases	1,954	1,729[e]	1,000	1,751

[e]Estimated.

[1]Includes all spending units with incomes of $3,000 or more in previous year.

[2]No cases reported or less than one-half of 1 per cent.

[3]Includes combinations of assets of fixed and fluctuating value and assets other than those listed above.

SOURCE: *Federal Reserve Bulletin*, June, 1953.

Table 2-15 illustrates the use of footnotes to qualify the title, a stub, and the figures in a table. In the same manner a footnote might qualify one or more of the captions, although in this table there were no exceptions to be noted for the captions. The first footnote indicates that "spending units" includes all spending units with incomes of $3,000 or more in the previous year. The second footnote is used to show that the number reporting was less than one half of 1 per cent, although in this case the footnote might mean that the number was zero. It is rather common practice in this situation to use zero when no cases were reported, and to use a footnote when the number of cases was so small that it would round to zero. The third footnote gives detail for the stub, rather than taking the space in the table to make the detailed statement of all the cases covered by the stub "Other." The footnote designated "*e*" indicates that the number was estimated. This type of footnote designation is widely used instead of a number. It

has the advantage of suggesting the meaning, and it can be standardized in a report and have the same meaning wherever it appears. Other common letters used in this way are "*p*" for preliminary and "*r*" for revised figures.

Source

When statistical data are used by someone besides the organization or individual that collected them, attention must be given to the source of the data. When a table is made up of data taken from the publications of some organization other than the one that made the collection, the data are referred to as *secondary*. Data taken from the publications of the organization that collected them are *primary* data.

When using data collected by someone else, it is ordinarily preferable to get them from a primary source, although in many cases secondary sources are more convenient. The primary source is more likely to give a complete and accurate description of the data, which may not have been included in the secondary source. Every time the figures are copied there is a chance for error; when the original source is used, there is the least chance for this.

When data collected by someone else are used, the source should be shown. If the data are taken from a secondary source, it is often advisable to give the primary source from which the data were secured as well as the secondary source.

When information is released by the organization collecting it, there should also be released a description of the methods used to secure the data and a warning of any limitations or special features of the data that should be observed in using them. A statement of the methods used in collecting the data makes possible the intelligent use of the data by others.

MECHANICAL FEATURES OF TABLES

The arrangement of the material in a table is the most important element in the construction of a table, but the mechanical features of the table can exert a considerable amount of influence on the reader. It is possible to set up a table in the most logical manner, and yet lose some of the effectiveness that might be secured.

Position on the Page

Probably the first factor to consider in the construction of a table is the position on the page. As a matter of convenience to the reader, it should be possible to read the entire table without turning the report or book sideways. Frequently there is a temptation to write the captions vertically instead of horizontally, but such a form should be avoided.

Another factor to consider is space. Enough white space should be left around the table to give it an attractive appearance and to make the parts of the table stand out in the proper relation to each other. The whole problem of balance and appearance in a table is a problem of layout, which is merely another way of saying that the most effective use should be made of the space available for the table.

Ruling and Spacing

The purpose of ruling statistical tables is to emphasize the relationships between the different parts of the table. The amount and type of ruling varies with tables and with the individuals constructing the tables. Many tables have no ruling at all. It is easy to devote too much attention to the ruling; any form that makes the relationships in the table clear and easy to see is probably all right. Space taken by ruling may be used instead to set off totals and to emphasize separate groups.

Table 2-15 (page 42) shows the ruling generally used in this text. The main body of the table is set off from the title by a double line. At the bottom of the main body of the table is a single line. The columns are set off by single lines, except that each end of the table is left unruled. The only lines ruled between the rows of a table are used to set off the major groups in the table, or to emphasize the total. Columns may be separated by double lines to indicate totals or to divide the table into major groups.

When there are a number of columns in a table, reference to the table is facilitated if the columns are numbered. Table 6-8 (page 120) has the columns numbered to aid in referring to them in the discussion. Sometimes the rows are also numbered for the same reason.

When there are numerous items in the stubs, it helps to show relationships between them if the subordinate groups are indented, using a scheme that shows the coordinate and subordinate relationships between the various stubs. In Table 2-12 the different groups are set off by spaces and the relationships are shown by indenting. Frequently in general-purpose tables with a large number of stubs a double space is left after every fifth line. This aids materially in reading the table and in finding the figures wanted.

There are many circumstances that may arise in the construction of analytical tables that will call for special forms of arrangement and special ruling. The statistician's opinion of the best form to use in such special cases will normally determine the manner in which the table is to be constructed. No standard form can be made to fit all cases, and any arrangement that shows the facts effectively can be defended and will usually be used. The details of mechanical arrangement must be governed by the single objective of making the statistical table as easy to understand as the nature of the material will permit.

Size of Type

When tables are printed, it is possible to use different sizes of type to show the relationships between different parts of the table. The size and heaviness of the type can be varied to show different degrees of importance. The heading is ordinarily given considerable prominence. The subheading is made less prominent by using smaller type. The main captions should be given more prominence than the subcaptions. Total figures may be emphasized by using a heavier type face for them than for the other figures in the table.

Typewritten Tables

Most of the statistical data that are published are in the form of printed tables, but the tables that the businessman makes up, and most of the tables made up for the use of businessmen, are typewritten.

Table 2-16

INVESTMENT PREFERENCES

Percentage distribution of spending units[1]

Preferred Investment	1953	1952	1951	1949
Assets of fixed money value	65	65	69	79
Savings accounts	20	16	13	18
Savings bonds	38	43	49	54
Accounts or bonds	7	6	7	7
Assets of fluctuating money value	24	26	23	11
Common stock	9	8	6	2
Real estate	14	17	16	9
Common stock or real estate	1	1	1	(2)
Other[3]	9	6	5	8
Not ascertained	2	3	3	2
All cases	100	100	100	100
Number of cases	1,954	1,729[e]	1,000	1,751

[e]Estimated.
[1]Includes all spending units with incomes of $3,000 or more in previous year
[2]No cases reported or less than one-half of 1 per cent.
[3]Includes combinations of assets of fixed and fluctuating value and assets other than those listed above.

Source: Federal Reserve Bulletin, June, 1953.

Some of the cheapest means of reproducing material make use of the typewriter; practically all offices have a stencil duplicator, and offset printing is very common. When offset printing is used, the typewriter is

commonly used to prepare copy for reproduction. In planning a table to be typewritten, the designer cannot rely on italic, bold face, or differences in type size to bring out relationships. Spacing is the most effective method of showing relationships on the typewriter. The typewriter will make horizontal lines, but if vertical lines are desired it is necessary to draw them. This has resulted in the practice of producing tables on the typewriter with fewer lines than is common for printed tables.

If several typewriters with different kinds of type, or if a typewriter with interchangeable type, are available, it is possible to use techniques similar to those used by the typesetter. However, the majority of typewritten tables used in business are prepared on one machine. Table 2-16 shows that a typewritten table can be made to serve the same purpose as the more attractive printed tables. Comparison of the typewritten table with the printed form (Table 2-15) will give an indication of how well the typewriter can serve as a substitute for printing.

STUDY QUESTIONS

1. Describe each of the four ways in which statistical data may be classified and give an example of internal and external data for each type of classification. (Use examples different from those given in the text.)

2. Give an example (in addition to those in the text) of a classification based on differences of degree, but in which the classes are expressed qualitatively rather than quantitatively. Which is the better way to state the classes? Explain why you believe this to be true in your example.

3. Instructions were given to a group of students to report their age as of their last birthday on a questionnaire.

(a) What are the real limits of the following classes?

21 to 24 years
25 to 28 years

(b) What are the mid-values of the two classes?

4. Employees are instructed on a questionnaire form to report their age at their nearest birthday.

(a) What are the real limits of the following classes?

21 to 24 years
25 to 28 years

(b) What are the mid-values of the two classes?

5. The argument is frequently given that it is unnecessary to show in the title of a table all that the table includes, since the person using the table can easily determine for himself just what information the table contains. Do you consider this a sound argument in favor of always using abbreviated titles for tables?

6. Give two examples (in addition to those in the text) of continuous variables, and two examples of discontinuous, or discrete, variables that might be used by a business statistician.

7. What are the disadvantages of a large number of footnotes in a table?

8. What is the purpose of the title of a table? Why do statisticians usually observe the rule that a title should answer the questions: What? When? Where?

9. Sometimes the total of a column is placed at the top of the column and at other times at the bottom. What basic principles govern where the total of a column should appear?

10. Sometimes the total of a row is found in the last column on the right-hand side of the table, and at other times the total is found in the first column at the left-hand side of the table, immediately after the stubs. What basic principles govern where the total of the amounts in a row is placed?

11. Probably more users of statistical data get their information from secondary sources than from primary sources. Why is this the case? Do you think it is good practice to use secondary sources?

PROBLEM

Many of the consumer markets for petroleum products are located at considerable distances from the areas of production. The following data from *Petroleum Industry*, 1956, published by the Chase Manhattan Bank, show the production and consumption of petroleum by major geographical areas for the year 1956. Construct a statistical table to present the following data.

> "In the Western Hemisphere the United States during 1956 was the largest user of petroleum with consumption of 8,757,000 barrels per day. Production was 7,939,000 barrels per day, leaving a deficit of 818,000 barrels per day. Canadian consumption in 1956 was 720,000 barrels per day but production was only 484,000 barrels per day, leaving a deficit of 236,000 barrels per day. Production in Venezuela was 2,465,000 barrels per day but consumption was only 150,000 barrels per day, leaving an excess of production over consumption of 2,315,000 barrels per day. All other countries in the Western Hemisphere consumed 1,215,000 barrels per day but produced only 651,000 barrels per day, leaving a deficit of 564,000 barrels per day. The total production of petroleum in the Western Hemisphere was 11,539,000 barrels per day and consumption was 10,842,000 barrels per day. The excess of production over consumption was 697,000 barrels per day.
>
> "In the Eastern Hemisphere the largest consuming area in 1956 was Western Europe. Consumption was 2,835,000 barrels per day and production only 232,000 barrels per day, leaving a deficit of 2,603,000 barrels per day. Production in the Middle East was 3,451,000 barrels per day and consumption only 380,000 barrels per day. This excess of production over consumption gave an exportable surplus of 3,071,000 barrels per day. Production in Africa was 40,000 barrels per day and consumption 450,000 barrels per day, leaving a deficit of 410,000 barrels per day. Consumption in the Far East was 1,043,000 barrels per day and production was 411,000 barrels per day, with a deficit of 632,000 barrels per day.
>
> "Only data for countries outside the Soviet Union and its satellites have been included in this analysis. For 1956 total production in the free world was 15,673,000 barrels per day. Consumption in 1956 was 15,550,000 barrels, leaving an excess production of 123,000 barrels per day. Production of petroleum includes natural-gas liquids."

3 INTERNAL STATISTICS—RECORDS AND REPORTS

In surveying the information needed for the solution of a problem, it is perhaps natural for the businessman to think first about the data available in the records of the business. To management, the most essential facts about a business are probably the internal data arising within the business unit from such sources as records of shipments, personnel files, production records, and the accounting system. Reports and summaries of facts of this type have long been recognized as vital instruments in the control of business. Practically every phase of the management of a business enterprise is more or less dependent upon this information, and without it executives would find it virtually impossible to control the operations of the large units in which present business activity is carried on. Although there are important items of information not recorded in its records, the accounting system supplies great quantities of data about the most important phases of a business. The methods of collecting, summarizing, and analyzing this information have become highly developed under the name of *accounting*. Statistical methods are not thought of as including the methods of collecting and analyzing accounting data, although occasionally statistical methods are used to make further analysis than the accountant's methods provide. In other words, the accounting records may be the source from which the business statistician secures data to which he applies his specialized methods of analysis. Accounting and statistics are both engaged in supplying executives with quantitative information about problems of business.

Many factors have impressed upon business executives the need for keeping records, and in all except the smallest concerns at least a set of accounting records is maintained. The regulations imposed upon industry by the war agencies increased greatly the number of records that had to be kept in order to give detailed statistics on operations, prices, material requirements, and every other phase of a concern's operations. However, it should not be thought that records of internal operations are more important than information about the external environment in which the business operates; both types of information are essential to the efficient management of a business. The present chapter will discuss some of the

principles and methods involved in the collecting of internal statistical data; the collecting of external data will be discussed in Chapter 5.

COLLECTING INTERNAL DATA

Since internal data originate within the business, collecting the information desired does not usually offer insuperable difficulties. The particular procedure depends largely upon the nature of the facts being collected and the form in which they exist. The problem of collecting is chiefly the problem of having the proper record made at the time of securing the information. The information wanted is frequently to be found in more than one department of the business, which increases the difficulty of getting just the information one wants recorded. In general, business organizations are not likely to keep records of facts unless there is an assurance that they will be used.

It often happens that facts are wanted after it is too late to make the necessary records. If the facts wanted are not in permanent form, it may be impossible to secure them when the temporary record has been destroyed and the facts are forgotten. Sometimes it is still possible to get the facts, but the cost of doing so may be prohibitive. For example, a sales manager needed to know the amount of goods which his company had sold in Iowa each year for the preceding ten years. The records had been kept so as to show the amount of sales made by each man in his territory. Iowa was included with Minnesota, North Dakota, South Dakota, and Nebraska in one territory, and a record of the sales for this territory had been kept; but no record had been made of the sales in any one state. It was found that copies of most of the sales invoices were available for the last five years of the period, which made it possible to classify them one by one according to the state in which the buyer's business was located. Since sales were numerous and in small amounts, this was a long job and the results were not entirely accurate. For the first five years of the period, however, there were no invoices. As the files had filled up, the oldest invoices had been destroyed. Thus, for half the period it was impossible to secure any of the information wanted.

The above case is fairly typical of the problem of collecting internal data. By far the most satisfactory method of collecting internal data is to know what information is wanted and make the necessary records as the facts become known. The methods of collecting internal data under these conditions are largely determined by the information wanted and are ordinarily fairly simple. When an attempt is made to collect internal data for a number of years back, difficulty is almost always encountered.

BASIC TYPES OF INTERNAL RECORDS

Since it is impossible to draw a distinct dividing line between accounting records and internal statistical records, this chapter may contain some material also found in textbooks in accounting. In business practice, many concerns do not try to distinguish between the two, with the result that in many businesses there is a department of accounting and statistics. When such a department exists, it is usually concerned with internal rather than external statistics. The statistician who specializes in internal data is partly an accountant. But the statistician who specializes in external data is more closely related to the economist, and in a business organization he may be given the title of economist rather than statistician.

Although internal statistical records are closely related to accounting records, there has not been developed a statistical procedure so uniform as double-entry bookkeeping. The statistical records of different businesses are usually much more varied than the accounting records of the same concerns. The basic reason for this is the fact that the accounting records of all businesses deal with monetary values, such as sales, expenses, liabilities, and asset values, and can be recorded in much the same manner in different types of businesses. On the other hand, a record of shoe sales by styles and sizes will be very different from the production records by products of an oil company.

There are, however, certain basic types of records and methods of analysis used for statistical data, even though they must be modified to fit the specific features of the individual business. The statistician has borrowed two basic devices of the accountant, the columnar journal and the ledger account. The problem of statistical records is not only to record the information wanted, but also to classify it as it is recorded so that reports can be taken from the records, when wanted, with the minimum amount of work. This is the same general objective as that of the accountant, and so it is not illogical that the general character of the records should be similar.

As defined earlier, classifying statistical data is a process of separating a mass of individual amounts into groups on the basis of some similarity. The records, therefore, must provide for this classification as the information is entered. The data to be classified in a statistical analysis will consist of a count of items or the sum of certain measurable characteristics.

Multiple-Column Record

Assume that a classification is to be made of sales by product groups by months. At the end of each month the sales manager wants the dollar value of sales in each of the groups. The *columnar record*, Fig. 3-A, is designed

SALES DISTRIBUTION

Month June, 19--

Date	Invoice number	Customer	Total sales	Department 1	2	3	4	5	6
19--									
June 3	3104	Perry-Scott	40.90	10.90		16.40	13.60		
3	3105	Mason Manufacturing	130.68		76.13	54.55			
3	3106	Trimble Co	10.15						10.15
4	3107	Exeter Hardware	184.77	14.50			69.04	14.98	86.25
4	3108	Merritt Co	39.82		14.40			25.42	
28	3374	Archer Co	289.03		99.76	19.86			169.41
28	3375	Merritt Co	76.46	30.00	46.46				
28	3376	Kelly Appliances	132.00		47.50		84.50		
28	3377	Johns Hardware	55.58			29.49		26.09	
28	3378	Trimble Co	29.19						29.19
28		Totals	6702.15	690.16	1980.70	548.88	902.50	806.43	1373.48

Fig. 3-A **Multiple-Column Record**

to record each sale and to divide the total amount of each sale among the product groups represented by the department numbers. Each sales invoice is analyzed, and the value of each group of products is entered in the proper column. After the entry is made, the amounts should be added and the total checked against the invoice total. When the last invoice for the month has been entered, the columns will be totaled. The result will be a classification of sales for the month by departments.

This columnar record, or work sheet, is typical of the journal with special columns, widely used by accountants as a book of original entry for the classification and recording of accounting data. It is a basic type of internal statistical record, and may be adapted to a great many situations. A single analysis sheet can be used for a special analysis, or a bound book may be used for recording data from day to day as it becomes available. The number of columns may be varied to suit the problem, and the number is limited only by the size of the sheet with which one can work.

Tally Sheet

When a classification is to be made by counting the number of items in each group, it is usually done with a *tally sheet*, instead of a columnar work sheet. A record is made on the tally sheet by entering a tally mark instead of writing an amount as in the columnar work sheet. The classes are set up on the tally sheet by allowing a line rather than a column for each class. A simple form of the tally sheet for classifying employees of a business according to length of service is given in Fig. 3-B on the next page.

Shipping Department Employees
Classified according to length of service
on July 15, 1958

Years of service	Tally	No. of employees
0 and under 5	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 111	33
5 " " 10	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 11	47
10 " " 15	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 11	52
15 " " 20	𝍸 𝍸 𝍸 𝍸 1	21
20 " " 25	𝍸 𝍸 1	11
25 and over	𝍸 1111	9*
Total		173

*Total 270 years of service

Fig. 3-B **Tally Sheet**

The source of the data was the personnel file for each employee. When the number of years of service was determined for an employee, a tally mark was made against the correct group. Crossing the fifth tally is a commonly used device to simplify counting the tallies. The tally sheet has the same purpose as the columnar work sheet, in that it provides a space for recording items for each class in order that they may then be summarized by classes.

Ledger Sheet

Instead of a separate column for each class, as used in the columnar work sheet, it may be desirable to use a different sheet for each class. This is the principle of the accountant's ledger. After the class in which an item belongs has been determined, it is entered on the proper page in the ledger. In early bookkeeping records, the ledger was a bound book; but in modern accounting systems and internal statistical records it may be a card file, a set of large analysis sheets, a loose-leaf ledger, or any other form that suits the particular problem at hand. The distinguishing characteristic of this record is the separate page for each class of items, regardless of the form of the page.

In many cases the ledger sheet is combined with the columnar record. In Fig. 3-A on page 51 it might be desired to classify sales by departments for each salesman. In such a case, a separate columnar analysis record could be set up for each salesman instead of for total sales. The one columnar sheet would be replaced by a book with as many sheets as salesmen, or a card file in which each card is a columnar record for one salesman.

Unit Records

In the previous discussion of three types of records, it has been assumed that each item was entered on the record individually, but it is obvious that this need not be done. With the ledger record, the invoices for a period such as a day or week might be accumulated. At the end of the period, the invoices might be sorted by salesmen, the sales added, and the totals entered on the ledger cards for each salesman. This would perform some of the summarization before making the entry, with a consequent saving in number of entries made but a loss in the amount of detailed information available in the record. However, the absence of detailed information in the record would not be serious, for the individual invoices would contain the detailed information; and the purpose of the record is to classify the data.

In the case just described, each item to be classified was entered on a different invoice; that is, the sale of only one salesman was entered on an invoice. This means that sales can be classified by salesmen by simply sorting the invoices on the basis of salesmen. This is probably the simplest method available for classifying data, and it permits the record to be kept by entering only the summary information wanted.

It does not always happen, however, that a separate document such as an invoice is made for each class into which the data are to be separated. It would be a reasonable assumption in most concerns that more than one product is entered on nearly every invoice. If sales are to be sorted by products, it is necessary to enter each product on a separate document. It would be possible to copy the information relative to each sale on a separate card, and then sort the cards by products in the same manner as the invoices were sorted by salesmen. This requires making a duplicate record from the original document unless that document can be designed in a form that will provide each item wanted as a separate unit.

The advantage of unit records lies in the fact that any summary desired can be prepared by sorting the items into groups and then accumulating the information wanted from the units in each group. The only records prepared, other than the unit record of each item, are the final summary reports. Since the unit record can be sorted as many times as desired, this method offers an efficient means of making several tabulations of the data. Unless a number of classifications are wanted, this method is usually not efficient for data that must be copied onto the unit records. But when the original business documents can be sorted, as, for example, when invoices can be sorted by salesmen, this method is very effective.

If a detailed analysis of sales were made by this method, the first step would be to copy the data from the sales invoices onto cards, putting each

sale of each individual product on a separate card. Assume that the following information would be entered on each card: name of the product, number of units sold, amount of sales in dollars, the customer's name and address, the terms of the sale, and the name of the salesman to be credited with the sale. If a sale consisted of twenty different items, twenty cards would be prepared. The cards could then be sorted into groups by products, and the amount of sales in each group could be found by merely adding the amounts on the cards in each group. The cards could then be sorted by terms of sale, by customers, by geographic location of customers, or by any other information entered on the cards, and summaries could be made by again adding the individual amounts by groups.

In actual practice, very little use is made of this method except in connection with one of several types of cards prepared specially for this type of analysis and designed for use with a mechanical method of sorting. In this connection, however, the unit record, which is the basis of machine methods of tabulation, has wide and important applications. This subject is so important that it is discussed in a special section of this chapter (pages 57 to 66).

Exhaust Method

The four types of records used in recording and classifying data just described represent the most important basic methods, but a fifth is sometimes useful. It is sometimes called the *exhaust method*, and although of limited usefulness, it will be described briefly.

Assume that a sales manager wants an analysis of sales by a few major groups. It would be possible to go through the sales invoices and record on an adding machine the sales belonging in the first class. When all the items in this group had been "exhausted," the total would be entered in the tabulation. To total the sales in the second class, the invoices would be gone through again and the items in the second class added.

The number of classes that can be exhausted efficiently in this manner is limited; if very many classes are wanted in the tabulation, it becomes more efficient to use one of the other methods rather than to go through the documents so many times.

MACHINE METHODS OF RECORDING AND SUMMARIZING STATISTICAL DATA

The methods described above do not require any mechanical devices, except that an adding machine can generally be used to advantage and that in the exhaust method the adding machine is an essential part of the process. The methods described in the following section, however, are

based on specialized machines or mechanical devices for performing one or more of the steps in the process of recording and classifying data.

The increased size of business concerns and governmental agencies has increased the volume of work involved in keeping adequate records, both accounting and statistical, to the point that simple hand methods are inadequate. To be useful to the management, statistical data must be available promptly, but the size of many operations is now so great that it is frequently impossible to prepare the reports desired in the time allowed. This has resulted in the development of more and more efficient machines for performing various steps in the recording of business data and the preparation of reports, until it is ordinarily assumed that any large-scale statistical tabulation will make use of machine methods.

Pegboard and Summary Strips

A variation of the columnar record known by various names such as *pegboard*, *shingle strip*, and *summary board*, requires certain special forms and devices as well as an adding machine. It is not correct to call this method machine tabulation, but it seems equally incorrect to consider it purely a hand method since special equipment is used. It is discussed here as the first of the machine methods, although it might have been discussed with the columnar record, since it resembles that type of record.

The basic device used in this method is an overlapping or shingling of forms on a special board or ring binder. This board or binder is provided with posts or rings for holding the shingled forms in correct alignment. By this method it is possible to create a flexible columnar work sheet. Fig. 3-C shows how the summary-strip method is used to analyze sales of a company by products by salesmen by months. The salesmen are identified by the numbers along the top of the board, and the products by the numbers in the stubs. A separate strip can be prepared monthly or as often as desired.

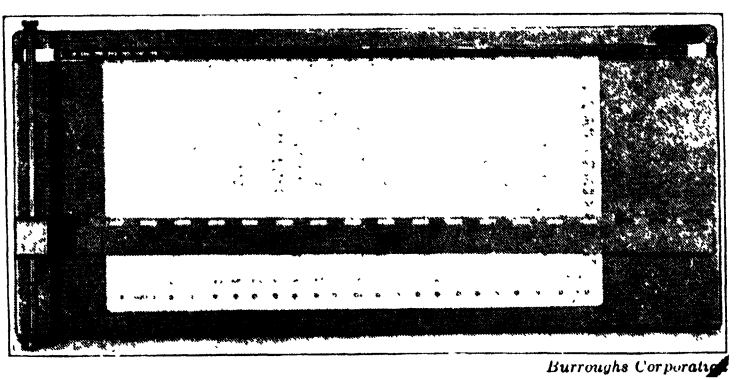


Fig. 3-C **Pegboard and Summary Strips**

The only mechanical devices, other than an adding machine, used in this method are those used to hold the forms in proper alignment. The strips could be held together with pins or clips, but rapid assembly is facilitated by the use of special pegs or rings designed to fit into holes punched into the top edges of the forms. These devices create a flexible columnar work sheet with all the advantages of this basic device, but with a saving of time in making possible the summary of the data in different ways with a minimum of recopying of the figures. In Fig. 3-D a clerk is shown verifying the totals on a pegboard.

Burroughs Corporation

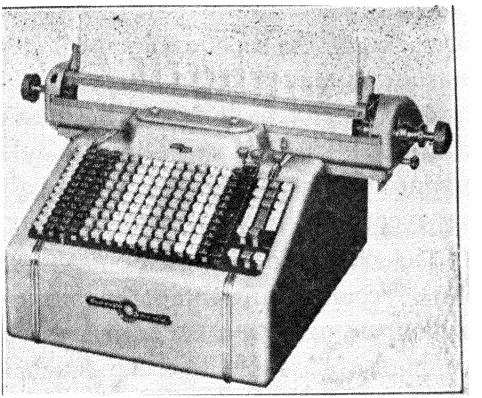
Fig. 3-D **Verifying Totals on a Pegboard**

Multiple-Register Machines

A simple method of classification is to use as many machines as there are groups to be added, one machine assigned to each class. The person making the tabulation goes through the original documents and enters each amount in the machine assigned to the group to which the item belongs. The total for each class is accumulated in the machine assigned to that class, and the work of entering the items on an analysis sheet is eliminated.

This method is not used, however, except as a temporary expedient; but the basic principle is so good that special machines have been developed to perform the same operations. These machines are built with a number of *registers*, which are merely individual adding machines built into one large machine. The multiple-register accounting machine shown in Fig. 3-E

contains nineteen registers or individual adding machines. The operator can determine into which register an amount will be added, and in this manner the machine operates as though nineteen adding machines were being used to accumulate amounts.



Burroughs Corporation

Fig. 3-E Multiple-Register Accounting Machine

In the analysis of sales invoices, nineteen classes of products could be set up and each class assigned to a specific register. The operator reads the first sale on the invoice, determines the product class to which it belongs, selects the register for that class, and enters the amount in the register. Each product is entered in the proper register, and the total amount of sales in each group is accumulated. When all the invoices have been entered, the totals are printed out of the registers and the tabulation has been made. The machines are ordinarily equipped with certain features that check the accuracy of the figures entered, and give warning that an error has been made in the distribution of the total. This sales invoice analysis is usually made at the time that the invoice is posted to the customer's account. If this is done, the statistical analysis is really a by-product of the routine accounting.

Posting Machines

The machines developed to record data on ledger accounts are primarily accounting machines, but they may occasionally be used for statistical data. The statistical uses of these machines, however, are so small that they will not be described here. It will be assumed that when the ledger account is used for statistical data, the entries are posted by hand.

Machine Tabulation of Unit Records

The highest degree of mechanization has been secured in tabulating data entered on unit records, and for the largest statistical tabulations — of either external or internal data — mechanical methods are normally used. The basic steps are the following:

1. Preparation of the record of the data in unit form.
2. Sorting the unit records into groups on the basis of the characteristics used for classification.
3. Adding the amounts being summarized (or counting the number of items).

Getting the information recorded in a form that will permit the performance of the second and third steps by mechanical means is a time-consuming procedure no matter how it is done. The biggest saving that can be effected is to design the business record in a form that will make each separate item of information available on a unit record. However, this cannot be done often, and it is usually necessary to make a completely new record in the special form needed for tabulation.

The most profitable use of mechanical devices has been in sorting items and in accumulating totals for different groups. Methods of performing these operations mechanically have been developed to such a point that the time saved is much greater than the time required to prepare the unit record in the required form.

Needle-sorting devices. Probably the simplest of all mechanical devices used for sorting cards makes use of a hole punched in a card in a certain position. When holes have been punched in this manner, mechanical devices can be used to bring together all the cards with holes punched in the same position. This device serves as the basis for several widely used methods of mechanical sorting by means of a long steel needle. In these systems the information to be used in sorting the cards into groups is entered on the cards by cutting a notch in the edge of the card in such a way that a hole located a small distance from the edge of the card becomes a slot. In other words, the card has holes punched close to the edge; when a notch is cut in the card opposite one of the holes, it removes the part of the card between the hole and the edge, thus forming a slot.

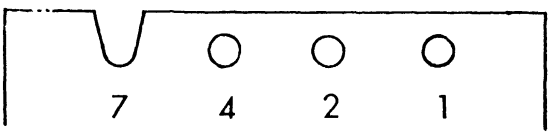
If some of the cards in a group have a notch cut out and other cards do not, the cards without the notch can be sorted out by inserting a needle through the holes in the pack of cards and lifting them. The cards with the notches cut out will drop free, while the others will hang on the needle. Assume that each card represents a sales invoice with the amount of the sale written on it. At the same time the entry was made, a notch was cut opposite a hole in the upper left-hand corner if the sale was for cash; but no notch was cut for a credit sale. At the end of the day, week, or any other period desired, the cards could be sorted by type of sale by simply lifting the cards without a notch out of the pack by means of the sorting needle. After sorting, the sales entered on each group of cards could be added.

The first step in this method is assigning a code number to each item to be recorded. The numerical code must be developed by determining the classes into which the data are to be distributed, and then always representing the classes by number. This is such a simple way of referring to a class that its use is not confined to mechanical methods of tabulation. Code numbers may be assigned to accounts in a ledger or to statistical classes in

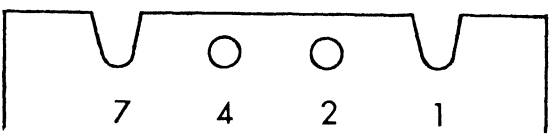
a hand tabulation. The only difference is that with mechanical tabulation of unit records codes *must* be used.

It will be noted that only qualitative classifications (based on differences of kind) and geographical divisions need to be coded. When quantitative data are to be classified according to differences of degree of a particular characteristic, the individual amounts can be punched directly into the card. In like manner, dates can be punched as quantities, August 30, 1958, being punched as 08/30/58, and time series can be tabulated using the numerical designation of the time units.

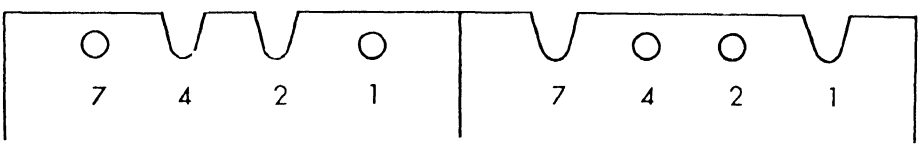
Holes are punched all around the card, and each one can be used to record information. In the example cited above there were only two groups, charge and cash sales, but usually there will be many more than two groups. For example, a common classification would be "Product." If there were less than ten different products, there would be less than ten numbers in the product code — each product could be identified by a different combination of notches cut opposite four holes. These four positions on the card would be labeled as follows: 7 4 2 1. The number seven would be notched as



and the number eight would be



If there were more than nine and less than 100 product classes in the code, two sets of four numbers would be needed and the number 68 would be notched as follows:



Each digit in a number must be assigned a group of four holes, and in this manner as large a number as is desired can be recorded, limited only by the size of the card.

With experience in operating a needle-sorting device it is possible to sort these cards with great speed, and thus one of the steps in tabulation can be performed very easily. However, this method does not provide a

mechanical means of summarizing the information on the cards after they have been sorted into groups. If the number in each group is wanted, the cards must be counted; if the sum of amounts written on the cards is wanted, they must be added. This addition can be performed no faster than is possible for figures entered in a columnar record. The saving effected by this method is a saving in the time required to sort the cards. Fig. 3-F is an example of a needle-sorting ledger card made by Royal McBee Corporation.

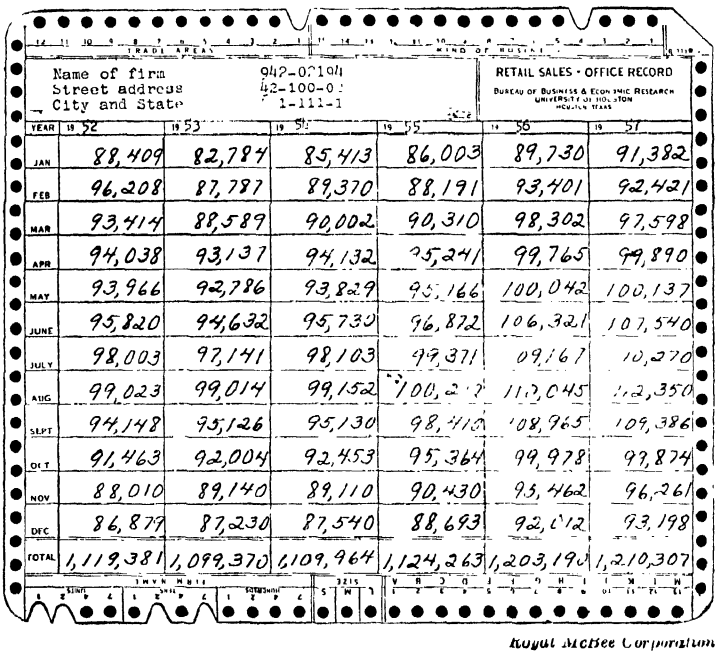
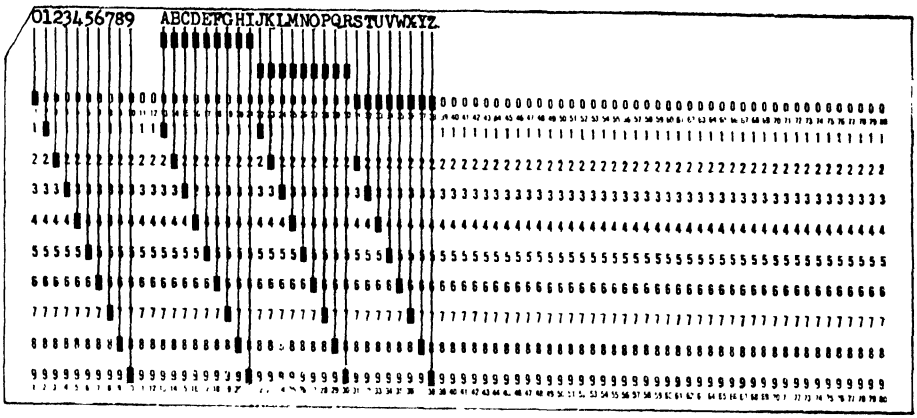


Fig. 3-F **Needle-Sorting Ledger Card**

Punched-card tabulation. The use of punched cards for tabulating statistical data has reached its most highly developed form with the machines manufactured by International Business Machines Corporation and Remington Rand Division of Sperry Rand Corporation. These systems are based on a unit record, and provide machine sorting of the items into classes, machine counting of the items in each class, and machine adding of amounts entered on the cards. The only one of the three basic steps listed on page 57 that must be done manually is the first, that is, the recording of the data on the card in the form of punched holes. The International Business Machines Corporation has developed a method of performing this step automatically, but this device is not in general use and for most tabulations the unit record must still be made by hand. But once a given set of facts has been recorded by punched holes, this set of facts can be copied as wanted on other cards.

The unit record is the tabulating card as shown in Fig. 3-G below. The International Business Machines card is marked off in eighty columns. In each of these eighty columns there are twelve positions, ten numbered from 0 to 9, and two positions above the zero position, referred to as "X" and "R" or "11" and "12." The two positions above the zero position are not printed on the card.



International Business Machines Corporation

Fig. 3-G **Punched Card**

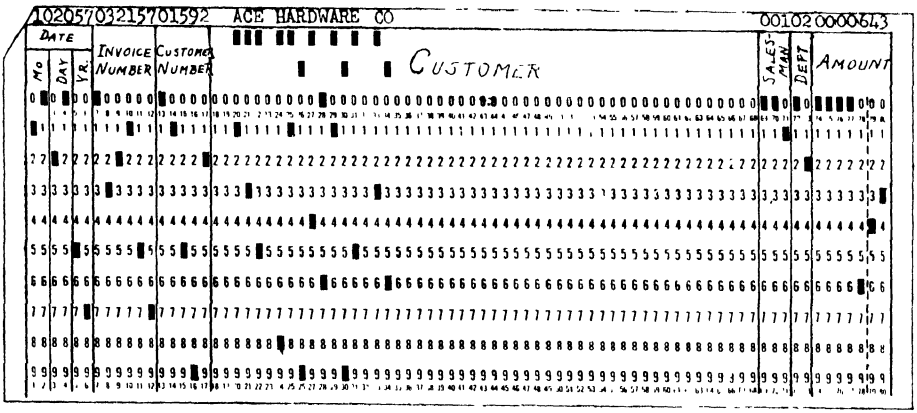
The card is essentially a line on which eighty letters or figures may be written, one character in each of the eighty columns. Figures are written by punching a hole in one of the positions from 0 to 9. This is done by a machine that resembles a typewriter, known as the alphabetic key punch. An illustration of a key punch appears in Fig. 3-J, page 64.

The card is fed into the machine and if the number 8 is to be entered on the card, the 8 key is depressed. A hole is punched in the 8 position; at the same time, with some punches, "8" is written at the top of the column. The machine then spaces and is ready to punch in the second column. The recording of letters requires the punching of more than one hole in a column, since there are twenty-six letters and only twelve positions in each column. Letters are recorded by punching two holes; this is done by the machine when the proper key is depressed on the punching machine. The numbers and the letters are recorded by the holes punched in the positions shown in Fig. 3-G.

The alphabetic punch makes possible the recording of names of salesmen, customers, products, and other qualitative data. Since there are only eighty columns on a card, however, there is a definite limit to the amount of data that can be recorded in this manner. Since numerical data are so much more compact, it is usually better to use a numerical code for as much qualitative information as possible. For example, each product may be assigned a number, and any number from 1 to 99,999 can be entered in

five columns. It would be impossible to write a description of the product in only five letters; thus this device records the information in a much smaller space than if a description were used. Only when the information is very important is it punched alphabetically on the card, and even then the name is usually also given a numerical code number as a check on the alphabetic description.

The card in Fig. 3-H is designed to record the sales data from the invoice shown in Fig. 3-I. The date is punched as the number of the month, the day of the month, and the last two digits of the year. The invoice number is punched directly from the invoice; the customer's name is recorded both by means of a numerical code, and by punching the name alphabetically. The salesman is recorded by code number only, the department by code number, and the amount of the sale by simply punching the amount numerically. Since this is to be a unit record, it is necessary to punch a separate card for each item on the invoice that is recognized as a separate unit.



International Business Machines Corporation

Fig. 3-H **Punched Card with Sales Data**

The invoice in Fig. 3-I shows sales for two departments, No. 2 and No. 11. The card in Fig. 3-H records the sale of the items from Department No. 2. The date is punched 10-20-57 in the first six columns to record October 20, 1957. The invoice number is punched as 032157, the customer's number is 01592, and the customer's name is punched alphabetically. The salesman's number, 001, is punched in the next "field," or column; the department number in the next is 02; and the amount of the sale, $6.43, is punched in the last field as 0000643.

The remaining items on this invoice were sales of the products of Department No. 11, and must be entered on another card. All the information on the second card is the same as that on the first card, except for the de-

HUNTER & MASON COMPANY
Merchandise Terminal
Chicago

Sold to ACE HARDWARE CO 01592
627 MAIN STREET
CHICAGO ILLINOIS

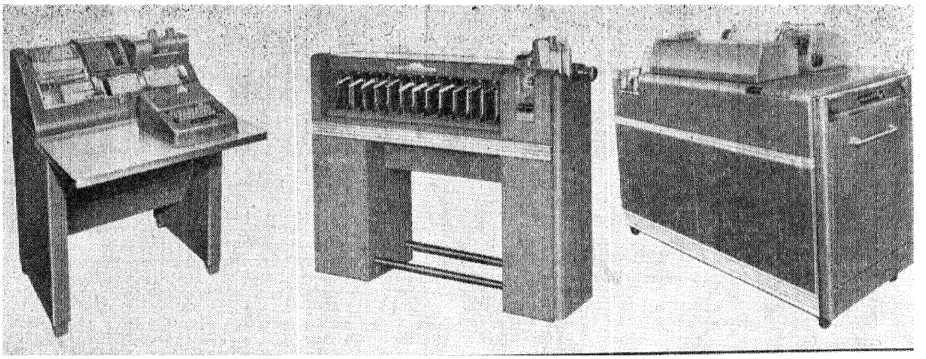
Invoice No. 32157
Date 10 20 57
Terms NET 30
Salesman MC DOWELL 001

Quantity	Dept No.	Product No.	Description	Price	Extension	Total
100	02	1056	USSSF HEX NUTS LESS 55			39
25	02	1149	CGE BOLTS	930	233	
25	02	1526	CGE BOLTS	1010	250	
25	02	4622	CGE BOLTS	1520	380	
25	02	5196	CGE BOLTS	1430	358	1221
			LESS 60 10			440
100	02	6173	CGE BOLTS	155	155	
50	02	6241	CGE BOLTS	320	155	
50	02	7152	CGE BOLTS	380	160	505
			LESS 60 10 10			164
						643
12	11	1932	UNIT CELL		4	48
12	11	1953	UNIT CELL		6	72 120
						763

Fig. 3-1 **Invoice Prepared from Punched Cards**

partment number itself and the amount of the sale. After the first card has been punched, it can be used to duplicate the information in the first 71 columns on a second card. This is done automatically by the duplicating key punch, and the operator then punches the department number and the amount of the sale in the last two fields. In this manner a unit record is made for the sales of each department of the business.

When all the sales of all salesmen for a week have been recorded on cards, a complete duplicate record of the business for the week has been created. The time required to make this record is well spent, since the unit records can be arranged into any grouping desired by the sorting machine illustrated in Fig. 3-J. The cards can be sorted at a rate up to one thousand per minute, dropping into a pocket in the machine according to the number punched in the column on which the sort is being made. If sales for each department were wanted, the cards would be sorted first according to the number punched in column 73, which would put all the the nines, all the eights, etc., together. Then sorting on column 72 would bring all the zeros and all the ones together. It will be remembered that no department was numbered over 18, so only zero or one would be punched in this column. When all the cards with zero in column 72 are brought together, they will already be in order according to the number punched in column 73; and when the cards with 1 punched in column 72 are brought



International Business Machines Corporation

Fig. 3-J Card Punch — Sorting Machine — Accounting Machine

together, they also will be in order according to the number in column 73. Thus the whole group will be in numerical order according to the number punched in the two columns, 72 and 73.

The next step in the tabulation of these sales data is that of adding the sales in each of the departmental groups. The electric accounting machine called a "tabulator" by most statisticians) shown in Fig. 3-J is used to compile these totals. This machine is an adding machine which is actuated by the holes punched in the cards. It is set to accumulate the total of the amounts punched in columns 74–80, and as the cards go through the machine, electric contacts are made that add whatever amount is punched in this group of columns on each card.

Since the total amount of sales is wanted for each department, the tabulator can be set to print a total for each group of cards indicated by the different numbers punched in columns 72–73. It will be remembered that the cards were previously arranged in order according to the numbers punched in columns 72–73. When all the cards with the number 01 have gone through, the machine stops, prints a total for that group, clears the machine of these figures, and then proceeds to add the group of cards with 02 punched in columns 72–73. At the end of each group of cards, the machine prints a total for that group and then goes on to the next. At the same time the total is being printed, the number punched in columns 72–73 to represent the group will also be printed, to serve in identifying what the total represents. After printing the total for the last group, the machine can be made to print a grand total for all the groups.

The resulting tabulation is a classification of sales for the week by departments. If it were desired to have a classification of sales by salesmen, and each salesman's sales classified by departments, the cards would first be sorted as described above by departments, and then further sorted ac-

cording to the salesman's number punched in columns 69–71. The tabulator would be set to print totals for each department, as described above, but since all the cards being run are for just one salesman, these totals would be the sales of the different departments for just that salesman. When all the cards for one salesman had run through the machine, it would then print the total sales made by this salesman before proceeding to add the sales of the next salesman. The resulting tabulation would be a classification of the sales of each salesman by departments, and also a classification of the total sales of the company for the week by salesmen.

The speed with which the unit records can be sorted into different groups is one important factor in this method of tabulation, but this advantage is also shared with the needle-sorting device previously described. With the punched-card methods of tabulation, however, the totals needed can also be accumulated automatically from the unit records. When a number of different classifications of the same data are wanted, the speed with which these groupings can be made and totaled more than offsets the time required in making the unit record in the special form needed. Just how many groupings of the data are needed before this method of tabulation becomes more economical than other methods usually depends upon the particular problem at hand.

Electronic data processing. In the years just prior to the outbreak of World War II, punched-card systems of tabulating data were generally accepted as a necessity in large-scale statistical and accounting operations. Punched-card systems were used by large business concerns in keeping accounting records and in compiling internal statistical data needed for making management decisions. Large-scale compilations of external data by statistical organizations, such as the Bureau of the Census, for all practical purposes would have been impossible without punched-card equipment. In addition to their use in compiling accounting and statistical data, punched-card systems were used to perform many clerical operations in a business, such as writing checks and preparing statements. The fact that the unit record could be used to perform other operations in addition to compiling accounting and statistical data increased the efficiency of punched-card installations.

The development of electronic computers during World War II was stimulated by the need for increased speed in solving mathematical problems. Punched-card equipment had been used in mathematical computations, but the top speed that could be obtained by computers using relays and counters operated by holes punched in cards or tape was a serious limitation. Electronic computers use the presence or absence of current in a circuit instead of the presence or absence of a hole punched in a card, with

the result that tremendous increases in speed have been achieved. For example, arithmetical operations on punched-cards reached a top speed of approximately 325 additions a minute, while a large computer can perform 2,000 additions or subtractions a second, 465 multiplications a second, and 257 divisions a second.

The increased speed alone would result in savings of time, but an additional feature of the computers is the fact that their operations can be directed automatically by a control unit. A storage device or memory unit receives instructions and data from an input device in the form of punched cards, punched tape, or magnetic tape. After the data and the instructions have been fed into the machine, it performs the entire processing in one continuous operation. The arithmetic unit performs the calculations on the data according to the instructions previously fed into the memory unit. Intermediate results are held in storage until needed, and the final results are written out by the output equipment, either on tape or on electrically controlled typewriters.

The preparation of the instructions, or *programming*, must be carried out in great detail and therefore requires a great deal of time; but the same program can be used to perform a given calculation on whatever data are put into the computer. Once the necessary programs have been prepared, a very large volume of routine operations can be performed at high speed. The use of the computers in business covers a wide range of clerical, accounting, and statistical operations. These operations can be performed very economically when the volume of work is large enough to take advantage of the high speed of the computer. The statistical computations performed by the computers frequently represent only a small proportion of the business uses, since so many other kinds of clerical work can be performed efficiently by the computers.

STUDY QUESTIONS

1. If you received a report compiled from sales invoices, in which an analysis of sales by types of retail outlets was given, how would you decide whether the report contained accounting data or internal statistical data?

2. In some organizations a person with the title "economist" appears to be doing work that in another organization is done by a person with the title "statistician." What kind of statistical work are these persons probably doing?

3. Sometimes the personnel representatives of companies hiring college graduates desire to interview accounting and statistics majors. If the representatives are looking for employees in these two fields, what kind of statistical work do you think they intend to have the employees do?

4. Under what conditions would it be worth while for a manufacturing concern to maintain its records on punched-card equipment? When would it be wiser for them to consider an electronic computer?

5. The statement has been made that, in making a decision, it is extremely wasteful to overlook the internal data. Is it any more wasteful to overlook internal data than to fail to use external data related to the solution of the problem? Give reasons for your answer.

6. The unit record system of tabulating statistical data offers the greatest possibility for expansion as the volume of work increases. Point out why this is true. Are there any inherent disadvantages of the unit record that tend to offset its advantage of expansibility?

7. Do you believe that the techniques of data processing have kept pace with the needs of business for more statistical data?

8. Under what conditions would the exhaust method of tabulating data be the best method to use? Under what conditions would this method have serious disadvantages?

9. The multiple-column record, the pegboard, and the multiple-register machine have certain basic features in common. In what way are these three methods of tabulation related?

PROBLEMS

1. A business starts a campaign to reduce the number of reports prepared and circulated, since there is some evidence that many of the reports no longer serve any useful purpose. In the course of the discussion there is some argument about the difference between "reports" and "records." Write a discussion of the differences, pointing out how you distinguish between reports and records.

2. The tabulating card shown on page 62 provides for tabulating sales by department but not by individual products. Design a card and describe the tabulating procedures needed to tabulate sales by individual product.

3. In this problem you are asked to set up a tabulating procedure for compiling data on the advertising sales of a newspaper. Information is desired on the amount of display advertising (excluding classified) run in the paper, classified by months, by kind of business that advertised, and whether the advertising was local or national. (It is suggested that the Standard Industrial Classification be consulted for use in preparing the classification "by kind of business." The amount of newspaper space devoted to advertising may be expressed in inches.)

(a) Prepare detailed procedures for tabulating the above information on IBM (International Business Machines) punched-card equipment.
(b) Design the card on which the information will be punched.
(c) Outline the tabulations that will be made, showing the tables to be prepared.

4. The following table gives the ages of 225 employees of a governmental agency employing a large number of field workers.

(a) Set up a tally sheet and classify the employees according to age.

(b) From the tally sheet prepare a table in proper form.

(c) Summarize briefly what this table shows about the distribution of ages of the employees in the agency.

(d) *Retain this solution for use in Problem 4 of Chapter 7, in Problem 3 of Chapter 8, and in Problem 1 of Chapter 9.*

AGES OF 225 EMPLOYEES IN A GOVERNMENTAL AGENCY

Empl. no.	Age (Yrs.)	Empl. no.	Age (Yrs.)	Empl. no.	Age (Yrs.)	Empl. no.	Age (Yrs.)	Empl. no.	Age (Yrs.)	Empl. no.	Age (Yrs.)
1	67	73	35	133	59	206	30	290	29	380	41
3	34	74	38	135	53	207	44	293	28	382	30
6	36	75	36	140	36	211	40	294	41	383	39
7	48	76	46	144	61	212	32	299	58	384	26
12	49	79	43	147	46	213	51	301	38	386	26
13	31	80	34	148	41	215	35	303	29	387	33
14	61	81	62	150	29	220	34	305	34	389	28
15	34	83	69	151	27	221	33	308	42	390	30
16	43	84	50	152	58	222	39	313	35	392	31
21	45	85	28	154	41	223	37	317	42	393	43
22	38	86	44	157	50	224	29	322	42	395	43
23	32	87	43	158	55	225	37	324	41	396	26
24	27	90	60	159	60	230	68	327	64	397	28
28	61	91	39	160	34	232	26	340	26	401	44
34	29	92	47	161	30	235	51	342	31	402	26
35	47	95	42	162	42	236	45	343	50	403	28
38	36	99	49	163	51	237	37	344	29	405	35
39	50	101	44	165	43	239	31	345	56	406	30
40	46	102	37	167	43	241	42	350	25	407	40
41	30	103	30	168	27	243	39	351	41	408	49
42	46	107	42	169	33	250	44	354	31	411	38
44	32	108	42	170	56	251	37	355	27	412	25
45	30	109	31	172	43	254	55	356	29	414	34
47	33	111	34	173	43	257	44	358	30	415	26
48	45	113	24	174	37	260	47	359	37	416	33
50	49	114	38	175	42	261	31	362	30	417	37
52	48	116	59	176	33	264	32	363	53	418	33
53	41	119	41	181	50	265	37	365	33	419	37
56	53	121	43	183	44	266	30	366	31	420	26
57	36	122	50	184	29	267	32	367	40	421	39
61	37	123	40	189	39	268	30	368	47	422	39
62	47	124	41	191	34	272	42	369	35	423	24
63	47	125	46	193	39	273	56	370	42	424	32
64	30	126	43	195	33	275	32	372	37	426	24
66	46	127	48	196	35	278	56	373	35	432	34
67	50	128	34	200	33	280	35	375	41		
70	28	131	37	201	43	284	30	376	30		
71	35	132	26	204	34	287	27	378	40		

SOURCE: Records of the agency.

5. From the data tabulated in Problem 4:

(a) Derive a cumulative frequency distribution on a "less than" basis.

(b) Derive a cumulative frequency distribution on a "more than" basis.

4 PUBLISHED DATA

After a businessman has assembled all the data on a problem that his own records make available, the next step is to find out what external data give significant information. The amount of external data now published has become so great that it is not a small undertaking to assemble all the information available on a particular problem. With experience a businessman learns, at least, the sources of the information he needs, even though he may make many decisions without taking into account significant data that are readily available.

The most important principle that can be learned at this point is that *no matter what problem a businessman has under consideration it is likely that information of importance in its solution is available from some source.* Every business executive should be aware of this fact and should make every effort to secure all the information that is already available before he makes a decision on a subject of importance.

An outstanding example of failure to make use of important data that could have been secured is the company selling a hay fever remedy. This company planned its advertising campaign without consulting a map showing sections of the country where people suffer from the allergy. The result was a distribution of too much of its advertising effort in localities where people had no use for this remedy. Such failure to use easily available data is inexcusable. Often it is difficult, if not impossible, to locate the best market accurately; but all the information that can be found should be used to make this location as accurate as possible.

A company selling antifreeze for use in automobiles can locate its market fairly accurately by the use of data already collected and published. The registrations of automobiles for the United States, by counties, can be secured. The United States Department of Agriculture publishes data showing the dates of killing frost (which means freezing temperatures) in different parts of the country. An automobile in a climate where it freezes for any appreciable portion of the year constitutes a market for antifreeze. This information would be basic, but it might be supplemented by data on income, type of roads, and driving habits in different sections of the country.

The amount of statistical data important to a specific business is so extensive that a separate volume on the subject could be written for each

industry. So many agencies are doing such an excellent job of collecting and releasing data of value to business that probably the most pressing problem is to teach businessmen to make effective use of what has already been collected. This does not mean that statisticians need to abandon their efforts to collect data on additional subjects, but it does mean that particular attention should be paid to using what is already available.

PRIMARY AND SECONDARY SOURCES OF DATA

In using published data it is important to recognize two types of sources, primary and secondary. The *primary source* is the agency that makes the compilation and first publishes it; a *secondary source* is any republication by another agency. For example, the Bureau of the Census collects and publishes the census of population; but many other agencies, both governmental and private, republish certain material in the original volumes. The census publications are primary sources and all subsequent publications are secondary sources.

It is generally preferable to use the primary source, since it usually contains a detailed description of the exact meaning and limitations of the information. For example, the 1948 Census of Manufactures presents twenty-two pages of detailed explanation of the data. This includes a detailed statement of the establishments covered by the census and the type of establishments excluded, the definition of the industries used in the classification, the degree of comparability of this census with other enumerations, and many other facts that are important in the use of the data. Secondary sources almost invariably give less explanation of the meaning of the statistics, and frequently present no explanation other than the captions, stubs, and footnotes in the tables.

The advantage of secondary sources is their convenience. Many of them, such as the *Statistical Abstract of the United States*, collect data from a large number of primary sources; accordingly, the user finds many of the facts he wants in one volume instead of having to consult a number of primary sources. In searching for information on an unfamiliar subject, it is good practice to look in a comprehensive secondary source such as the *Statistical Abstract*. The data wanted may not be there, but the secondary source may suggest primary sources that do contain the facts.

GUIDES TO PUBLISHED DATA

The volume of published statistical data on business problems is so great that it is wise to make use of published guides in locating information. Many large business organizations have special libraries with a staff trained in library science, and if such a department is available in a business

organization, it should be used whenever statistical data on a problem are needed. Concerns that do not have a librarian can make effective use of bibliographies and guides to published data, although experience in assembling published data is more valuable than any published guides.

The well-known *Readers' Guide to Periodical Literature* and *Industrial Arts Guide* are available in public libraries and may always be profitably consulted. However, their listings are usually not designed to provide the sources of specialized statistical data. After a little experience with the best-known general sources of statistical data such as the *Statistical Abstract*, the *Economic Almanac*, the *Survey of Current Business*, and others that are described later in this chapter, it becomes almost automatic to look in them for any desired information. The bibliography published in the *Statistical Abstract* is an excellent guide to sources of statistical data on nearly every subject of interest to the businessman.

Statistical Services of the United States Government, published in 1952 by the Office of Statistical Standards, gives an account of the federal agencies that compile statistical data, and the general principles and practices that are observed throughout the government in the compilation of statistics. The principal types of economic and social statistics are described and cross referenced to a bibliography of 125 official publications.

A useful guide to information on marketing is *Market Research: A Guide to Information on Domestic Marketing*, published by the Department of Commerce. This guide is an inventory of available information on markets in the United States and, since practically all external business data are valuable in making marketing decisions, it is a surprisingly complete inventory of business information.

The sources mentioned above are valuable timesavers in locating information, but since statistical data on business problems are found in so many kinds of publications, it is not safe to rely on any one guide or bibliography. The discussion of important sources in the following section is given to point out the most useful sources. If a search is started in these sources, additional leads are usually uncovered, but there is no assurance that all the available information will be found.

IMPORTANT SOURCES OF STATISTICAL DATA

The guides described in the preceding section are useful for even the most experienced statistician as well as the businessman who has only a slight acquaintance with the sources. One of the first things a student must do, however, is to become familiar with certain important sources. Since much of the needed information can be found in the sources, reference to the guides may be unnecessary.

To indicate something of the type and variety, several important sources are briefly described below and then an extensive list of rather specialized sources is given. Many are left out of this list, which does not attempt to duplicate the guides to published data.

The list every businessman should be acquainted with is divided into two groups:

1. Publications that contain comprehensive classifications of data available only once a year or less frequently.
2. Sources of current data published quarterly, monthly, weekly, or even daily.

Since the role of the federal government is so important in this connection, each group is subdivided into governmental and nongovernmental, or private, sources.

The collecting and publishing of external data is so expensive that the source of funds for the purpose is an important consideration. The collection of data of general interest to the public can in many circumstances be better done by government than by private organizations. Thus, there is a steady trend toward the collection of such data by government agencies. State governments and universities gather certain local data, but the agencies of the federal government are able to collect for the country as a whole as well as for the smaller geographical subdivisions.

Trade associations perform a type of cooperative collection, the cost being financed by the members. Some private businesses such as banks collect certain data as a service to their customers, but the amount of collection that any one institution can do is limited. Newspapers and magazines collect data to publish as part of their content and consider the cost a part of the total editorial cost. But probably the most important of the private organizations are those that collect and sell their data. Since a number of business concerns can use the same information, it can be sold a number of times, thus making the individual price much less than the total cost of collecting. This statement refers to organizations that make data generally available to anyone who wants to buy; the private research organization that makes special studies for individual concerns is a different type. It is discussed in Chapter 5.

Comprehensive Summaries of Statistical Data (Annually or Less Frequently)

Publications in this group usually consist of series covering long periods of time, or extensive compilations such as the Census of Population, which is compiled every ten years. For data extending any length of time into the past, and for comprehensive information on all phases of business,

comprehensive summaries should be consulted first. On many occasions no further search will be needed. When current data, such as monthly statistics on business conditions, are needed, the following sources should be consulted in the same manner.

A. Governmental sources. Under this heading the following sources, with which every businessman should be familiar, will be described:

1. *Bureau of the Census:* periodic censuses.
2. *Statistical Abstract of the United States.*
3. *Agricultural Statistics.*
4. *Minerals Yearbook.*
5. *County Business Patterns.*
6. *Banking and Monetary Statistics.*
7. *Handbook of Labor Statistics.*
8. *Economic Report of the President.*

1. *Bureau of the Census: periodic censuses*

The Bureau of the Census, Department of Commerce, is the largest data-collecting organization in the United States. Acquiring a knowledge of the information it collects and publishes is one of the first steps in becoming acquainted with the published statistical data available to the businessman.

The Constitution provides that a count of the population shall be made every ten years to serve as the basis for apportioning state representation in the House of Representatives. This count was first made in 1790, and since that date a Census of Population has been taken every ten years. Facts about the population of the United States and each of its political subdivisions are contained in the four volumes of the *U. S. Census of Population: 1950.*

Although its purpose was originally to provide information needed to apportion state representation in the House of Representatives, the Bureau of the Census long ago expanded its scope to provide a wide variety of facts about the people of the United States and their economic activities. Recent economic censuses are:

(a) 1950 Census of Housing (taken in connection with the 1950 Census of Population).
(b) 1954 Census of Agriculture.
(c) 1954 Census of Business. (Retail, wholesale, and service trades).
(d) 1954 Census of Manufactures.
(e) 1954 Census of Mineral Industries.

The basic data compiled and published by the Bureau of the Census is such a fundamental part of our information about the people and economy of the United States that there is some tendency to take the information for granted, as if it simply existed without any effort having been expended to secure the data. The story is told of an individual who proposed abolishing

the Census of Population and saving its cost, since all of the data published in the census volumes could be gotten from one of the large life insurance companies. No one had explained to him that the insurance company merely republished the information compiled by the Bureau of the Census.

The fact that most of the census information is based upon complete enumerations makes possible the tabulation of results in great detail, such as a tabulation of facts secured from the smallest geographical divisions of the country. However, there is a trend toward taking samples of some of the data wanted, since the sampling process gives results that are approximately correct, permits collecting more information at much less cost, and provides for more up-to-date facts. For small divisions of the country, however, the census tabulations represent practically the only information available about the population and the economy. Without these data, businessmen would be forced to make many decisions with respect to their operations without accurate information. It is unfortunately true that many businessmen still do not make the most effective use of this wealth of information. A good understanding of the types of data available in the periodic censuses is important information for every businessman, and it would be a useful habit to check the census publications whenever external data are needed as the basis for making a decision.

The use of the data published by the Bureau of the Census is facilitated by the *Catalog of United States Census Publications*, which is issued each quarter and is accumulated in an annual volume. A file of the annual issues will give a complete listing of the publications available, with each publication indexed by subject and by geographic areas. A monthly supplement to this catalog enables users to keep themselves currently informed about publications as they appear.

2. *Statistical Abstract of the United States*

This volume, published annually by the Bureau of the Census, contains over 1,000 tables. The facts are collected by government and nongovernment agencies, and represent what the compilers consider the most generally useful series of data on the industrial, social, and economic organization of the country. Most of the data are published elsewhere, and usually in more detail than in the *Statistical Abstract*; so if more detailed information is desired on a given subject, the original sources can be consulted. To make this easy, the source of every table is noted; and, in addition, a detailed bibliography is included. The *Statistical Abstract* is a typical secondary source which is extremely useful.

Supplements to the *Statistical Abstract* are published for the purpose of giving more detailed information. For example, the *County and City Data Book of 1950* brings together 128 items of information for each

county, standard metropolitan area, and state, and 133 items for each of the 484 cities with populations of 25,000 or more. The items included in this one-volume collection are the data most often used, and thus should be adequate for many users. Of course, if more details are needed, it is necessary to refer to the primary sources.

Historical Statistics of the United States, 1789–1945, a historical supplement, was published in 1949; and *Continuation to 1952 of Historical Statistics of the United States, 1789–1945*, was issued in 1954 to bring the series forward seven more years. These volumes contain approximately 3,000 statistical time series (usually on an annual basis) which cover various periods since 1789. The series included are the ones considered to be of the widest general interest. In addition to giving the statistical data, *Historical Statistics* provides brief descriptions of the series and information on the sources of the data.

3. *Agricultural Statistics*

This volume is published annually by the United States Department of Agriculture and brings together the most important series relating to agriculture collected by the Department or compiled from other sources for use in the Department. Unlike the *Statistical Abstract*, which is chiefly a secondary source, *Agricultural Statistics* is predominantly a primary source for data collected by the Department of Agriculture. The data for which it is a secondary source are facts of interest to agriculture, such as prices and production of rayon. *Agricultural Statistics* has been published as a separate volume since 1936; before then, these data were included in the *Yearbook of Agriculture*.

4. *Minerals Yearbook*

The United States Bureau of Mines publishes annually a comprehensive record of the production and marketing of minerals in the nation. The statistical data, which represent only a small portion of the book, give information on production, stocks, consumption, prices, foreign trade, and the condition of the industry in other countries.

5. *County Business Patterns*

Another important basic statistical summary issued by the federal government is *County Business Patterns*, prepared jointly by the Bureau of the Census and the Bureau of Old-Age and Survivors Insurance. The 1953 issue gives information on the amount of taxable payroll for the first quarter of the year, the number of employees for the payroll period ended nearest March 15, and the number of reporting employers classified by number of employees. The information is tabulated from the reports of employers who are subject to the provisions of the Social Security Act that

relate to old-age and survivors insurance. Only employment covered by this Act is included in the tabulations; but as the coverage has been extended to additional groups, the statistical data derived from the records have become more complete.

Since the information about employees covered by old-age and survivors insurance is reported each year, this source of data makes it possible to obtain information much more frequently than it could be collected by a census. As the coverage under the insurance program has been broadened, and as the data have been coordinated with the employment and payroll data collected in the economic censuses, this source of information has become increasingly significant. It is interesting, also, as an example of economic data that have been derived from the information reported as part of a tax-collecting program. A relatively small expenditure of funds can make a very large amount of data available for business use, since the major cost, collecting the basic information from the individual employers, is a necessary part of the cost of collecting the taxes.

6. *Banking and Monetary Statistics*

This volume was published in 1943 by the Board of Governors of the Federal Reserve System to make available on a uniform basis statistics of banking, monetary, and other financial developments since 1914, when the System was established. A few significant series are given for a longer period. For most of the series, current data are available in the *Federal Reserve Bulletin*, which is described on page 80, with the publications of current statistical data.

7. *Handbook of Labor Statistics*

This handbook is published by the Bureau of Labor Statistics, United States Department of Labor, at intervals of several years. The material consists largely of digests of reports or articles published by the Bureau, although material from other bureaus and divisions of the Department and a few outside sources is also used. The statistical material covers, rather completely, employment, payrolls, wages, working conditions, commodity prices, housing, business conditions, and other subjects of interest to labor. The current publications of the Department, particularly the *Monthly Labor Review*, which is discussed on page 80, provide current data on the same subjects covered more completely in the *Handbook*.

8. *Economic Report of the President*

In accordance with the Employment Act of 1946, the President submits regularly to Congress economic reports reviewing the economic situation and recommending needed legislation. The Act created the Council of Economic Advisers in the Executive Office of the President, and a com-

mittee of both houses of Congress called the Joint Committee on the Economic Report. The President's annual *Economic Report* contains extensive summaries of statistical series related to economic conditions.

It should be pointed out that the above list of comprehensive sources is a highly select list; and even if the information desired for a particular problem is found, it would be wise to check the guides mentioned on page 71 to see whether or not additional information about the problem is published in another source.

B. Private sources. Under this heading will be given a brief description of the following private sources of data:

1. *Standard and Poor's Trade and Securities Statistics.*
2. *Moody's Manuals of Investment.*
3. *The Economic Almanac.*
4. Newspaper almanacs.
5. *Commodity Yearbook.*
6. *Survey of Buying Power.*
7. *Editor and Publisher Market Guide.*
8. *Consumer Markets.*
9. Statistical publications of trade associations.

1. *Standard and Poor's Trade and Securities Statistics*

This is a loose-leaf service published by Standard and Poor's Corporation; it contains detailed information on individual companies, industries, and general business conditions. The complete service constitutes one of the most comprehensive collections of data on business available, and since it is in loose-leaf form it can easily be kept up to date. For this reason it is also listed under sources of current data.

2. *Moody's Manuals of Investment*

Five volumes are published each year by Moody's Investors Service covering the following groups of companies: (1) industrials, (2) railroads, (3) public utilities, (4) banks and financial institutions, and (5) governments and municipalities. These volumes contain information on individual corporations, including a history of the business, financial statements, and descriptions of each company's securities outstanding. In addition, a considerable amount of statistical data on business conditions is included.

3. *The Economic Almanac*

This volume is published by the National Industrial Conference Board as a selection of information most useful to the businessman. The major part of the data is statistical, and covers industrial and labor relations, business conditions, population, resources, industries, national income, wealth and savings, consumption and standards of living, private debt, public finance, and international trade.

4. *Newspaper almanacs*

The *New York World-Telegram* publishes annually *The World Almanac*, containing a summary of events of the year in all fields and many statistical series. In addition to information of national interest, it carries detailed information for New York City. The *Dallas Morning News* publishes the *Texas Almanac*, devoted to information and statistical data on Texas.

All the newspaper almanacs are supported by advertising and can therefore be sold at a low price. The fact that they appear promptly after the close of the year adds to their usefulness. For these reasons they are deservedly popular with businessmen as a source of information. While they probably do not have the ponderous and dignified appearance of the government or the nongovernment publications described above, they are apparently carefully compiled and give much the same information that is included in the other secondary sources already described.

5. *Commodity Yearbook*

The Commodity Research Bureau, Inc., publishes annually the *Commodity Yearbook*, a volume of statistical data on individual commodities important to the businessman.

6. *Survey of Buying Power*

The *Survey of Buying Power*, an annual publication, is a special issue of *Sales Management*. This source gives yearly estimates of population, income, retail sales, and wholesale sales for counties and cities in the United States and Canada, and additional data for large cities.

7. *Editor and Publisher Market Guide*

This annual publication contains statistical data on population, housing, income, banking, retail sales, newspapers, and other economic and marketing data for over 1,500 cities in the United States and Canada.

8. *Consumer Markets*

Consumer Markets is published annually by Standard Rate and Data Service, Inc., and gives information on population, households, income, retail sales, and farm income for counties and cities in the United States.

9. *Statistical publications of trade associations*

A number of annual statistical publications of trade associations are so widely used that they are listed here by name. Reference to a more complete list of trade associations is given in this chapter on page 83.

(a) *Annual Statistical Report of the American Iron and Steel Institute.*
(b) *Petroleum Facts and Figures*, American Petroleum Institute.
(c) *Gas Facts*, American Gas Association.
(d) *Automobile Facts and Figures*, American Automobile Association.
(e) *Life Insurance Fact Book*, Institute of Life Insurance.

Comprehensive Summaries of Current Statistical Data

This group of sources comprises those publications that present statistical data on business quarterly, monthly, or more frequently, and that contain series of rather general interest to businessmen. These sources are similar to the comprehensive sources already described in that the student should be familiar with all of them and with the data they contain. Since most of the sources described in the preceding section are published annually or less frequently, they cannot contain recent data. Generally the time series are not classified on a monthly basis, but if they are, current data must be secured from one of the sources described below. In some cases the current source is designed to supplement an annual comprehensive publication and the two should be used together.

*A. **Governmental sources.*** Under this heading the following sources will be described:

1. *Survey of Current Business.*
2. *Federal Reserve Bulletin.*
3. *Monthly Labor Review.*
4. *Economic Indicators.*
5. *Construction Review.*
6. Current publications of the Bureau of the Census.

1. *Survey of Current Business*

This magazine, published by the Office of Business Economics, U. S. Department of Commerce, is one of the most comprehensive collections of current statistical data and contains most of the monthly series of economic data available in the United States. Each issue contains the figures for the most recent month and the twelve preceding months. A weekly supplement is issued, giving certain weekly series and also early publication for a limited number of monthly figures that appear later in the monthly issue. In addition, a biennial supplement is published to give the data for past years, and to assemble for all of the series published monthly detailed information which is necessary for an understanding of the series. The supplement also includes detailed descriptions of the sources of the data published.

The supplement published in 1957, under the title *Business Statistics: 1957 Biennial Edition, A Supplement to the Survey of Current Business*, contains monthly data for approximately 2,600 series for the four years 1953–1956, and annual data from 1929. For monthly data prior to 1953, it is necessary to consult the earlier editions of the supplement.[1]

[1]The *1932 Annual Supplement to the Survey of Current Business* contained monthly figures for the years 1923 through 1931, and monthly averages for earlier years from 1913. The first annual supplement was issued in 1931. Prior to that time, semiannual supplements were issued in February and August. The data in the 1932 supplement superseded all previous data and thus became the basic volume of the comprehensive record. (*Continued on next page.*)

In compiling monthly data for a period of years, it is necessary to start with the most recent issue and work backwards to the last supplement. Any revisions since the most recent supplement will be noted in footnotes; it is therefore highly important that these footnotes be read carefully. If a revision has occurred, the footnote will indicate which monthly issue carries the revised data, which should be used instead of that published in earlier monthly issues or supplements.

2. *Federal Reserve Bulletin*

This monthly publication of the Board of Governors of the Federal Reserve System contains detailed current statistics on the Federal Reserve Banks and the member banks, domestic and international financial statistics, and statistics covering all phases of general business activity. It supplements the comprehensive volume, *Banking and Monetary Statistics*, also published by the Board.

3. *Monthly Labor Review*

This is the official publication of the Bureau of Labor Statistics, and is the method of release for a considerable amount of the current data compiled by the Bureau. It contains data on wages and hours of labor, employment and payrolls in manufacturing and nonmanufacturing industries, wholesale and retail prices, industrial disputes, building operations, and various miscellaneous statistics of interest to labor. It also contains current data for most of the series published in the *Handbook of Labor Statistics*.

Reports presenting more detailed data are issued monthly on the following subjects: employment and payrolls, labor turnover, current wage developments, hours and earnings, consumer price index and retail food prices, wholesale price index, labor-management disputes, and construction.

4. *Economic Indicators*

This leaflet is published monthly by the Congressional Joint Committee on the Economic Report. It contains data prepared by the Council of

[1] (*Continued from preceding page.*) The 1936 supplement carried monthly data for the years 1932 through 1935, and monthly averages back to 1913 where the data were available. Supplements were issued in 1938, 1940, and 1942, each presenting monthly data for the four preceding years and monthly averages back to 1913.

No supplements were issued during the war, but in 1947 the biennial publication was resumed. Subsequent issues have been dated 1947, 1949, 1951, 1953, and 1955. Each issue gives data for four years, which means that revised data can be given for two years published in the previous supplement. For example, the data for 1951 and 1952 in the *1955 Supplement* contain revisions of the data for these two years published in the *1953 Supplement*. When series are revised by the compiling source, the complete revised series is usually given in one of the monthly issues of *The Survey of Current Business*, and all subsequent issues carry a footnote on this series indicating where the complete revised data may be found. Sometimes minor revisions are not published, in which case a footnote states that the revised data may be secured from the Office of Business Economics.

Economic Advisers and presents in tabular and graphic form the most important indicators of current business activity. This publication serves to keep current the statistical data published in the *Economic Report of the President*. An airmail subscription is available for businesses that want to receive it as early as possible. Descriptions of the procedures used in compiling the series published in *Economic Indicators*, and an explanation of their uses and limitations are presented in *Historical and Descriptive Supplement to Economic Indicators*.

5. *Construction Review*

This monthly publication was first issued in January, 1955, replacing the monthly publications *Construction* (Bureau of Labor Statistics) and *Construction and Building Materials* (Business and Defense Services Administration). As the primary source of data on building activity, it presents detailed information on types of building by states and cities.

6. *Current Publications of the Bureau of the Census*

The Bureau of the Census collects important current series as well as the comprehensive enumerations described earlier in the chapter. These current series have been integrated with the periodic censuses to give an up-to-date measure of many aspects of economic life. The following list gives the divisions of the Bureau of the Census, and describes briefly the most important current information published by each division.

(*a*) *Agriculture Division.* In addition to taking the Census of Agriculture, this division collects statistics and issues periodic reports on cotton ginnings and cotton production.

(*b*) *Business Division.* This division takes the Census of Business and in addition issues monthly reports showing national and regional estimates of retail sales by kind of business. Monthly reports also are issued providing national and geographic division trends in wholesale sales and inventories by kind of business.

(*c*) *Foreign Trade Division.* Extensive statistical data on United States exports and imports of merchandise and of gold and silver are collected from shippers' export declarations and import entries. The reports issued by this division are given in the *Catalog of United States Foreign Trade Statistical Publications*.

(*d*) *Governments Division.* This division collects data and prepares reports showing current statistical data relating to the characteristics and functions of state and local governments and their administration in the United States.

(*e*) *Industry Division.* In addition to taking the Census of Manufactures and the Census of Mineral Industries, this division issues the *Facts for Industry* series and makes the *Annual Survey of Manufactures*. The *Facts*

for Industry series makes available current statistics on commodity production and shipments for a large number of individual industries. The *Annual Survey of Manufactures* is based on a sample of manufacturing plants and provides annual data between Censuses of Manufactures on the key measures of manufacturing activity for industry groups and geographic divisions.

(*f*) *Population and Housing Division.* A Census of Housing has been taken along with the Census of Population since 1940. The current *Population Survey* is conducted monthly. Through enumeration of a sample of the population, information is obtained currently on employment, unemployment, hours of work, occupation, and other subjects. Estimates of population between censuses are prepared, and special censuses of local areas are taken at the request and expense of the local governments involved. Projections of the population are also prepared from time to time.

B. Private sources. The following lists represent a few of the private sources of current statistical data.

1. *Newspapers and magazines*

(*a*) *Barron's* (weekly).
(*b*) *Business Week* (weekly).
(*c*) *Commercial and Financial Chronicle* (weekly).
(*d*) *Dun's Review* (weekly).
(*e*) *U. S. News & World Report* (weekly).
(*f*) *Wall Street Journal* (daily except Saturday and Sunday).

2. *Statistical services*

The publications of the agencies listed below are normally sold on a subscription basis, and the costs are high in comparison with the sources described above. In addition to supplying statistical data, these agencies furnish analyses, forecasts, and specific recommendations with respect to business and investment problems.

(*a*) Babson's Reports, Inc.
(*b*) Brookmire Corporation.
(*c*) Standard and Poor's Corporation. This service was also listed under the comprehensive sources of long-term data; since the publications are loose-leaf and kept constantly up to date, they properly belong in this category as well.
(*d*) Moody's Investors Service.
(*e*) United Business Service.

Miscellaneous Sources of Statistical Data on Specific Subjects

The sources described in the preceding sections are general sources that everyone dealing with business data should know reasonably well. The following list represents a small sampling of additional sources, usually

giving information on special subjects. Some data from practically all these sources are included in the comprehensive sources already described, but in almost every case the primary source gives more information than is repeated in the secondary sources.

1. *Trade associations*

A listing of more than 2,000 national trade associations and related national organizations of businessmen appears in the *Directory of National Trade Associations*, which is published by the Department of Commerce.

2. *Technical periodicals*

(*a*) *American Metal Market.*
(*b*) *Chain Store Age.*
(*c*) *Engineering & Mining Journal.*
(*d*) *Engineering News-Record.*
(*e*) *Fairchild Publications.*
(*f*) *Iron Age.*
(*g*) *Oil and Gas Journal.*
(*h*) *Oil, Paint and Drug Reporter.*
(*i*) *Printers' Ink.*
(*j*) *Railway Age.*
(*k*) *Steel.*

3. *Private organizations*

(*a*) F. W. Dodge Corporation.
(*b*) Media Records, Inc.
(*c*) R. L. Polk and Co.
(*d*) Publishers' Information Bureau.

STUDY QUESTIONS

1. Assume that you have compiled a table from the *Statistical Abstract.* When you present the table in a report, the argument is advanced that the data have been taken from a secondary source and are not so accurate as if you had secured them from a primary source. Justify the use of the secondary source, explaining the advantages of using it instead of the primary source.

2. How do the problems of collecting data by a trade association differ from the problems encountered by a governmental statistical agency in collecting industry data?

3. On the basis of the information in Chapter 4, list the sources that you would consider worth investigating in order to find the following information. Indicate which sources are primary and those which are secondary.

(a) Number of turkeys sold by farmers in each state, in each month of the year.
(b) Number of producing oil wells and the number of dry holes drilled in each state in 1957.
(c) The size of the organic chemicals manufacturing industry in each state in 1957.
(d) Natural gas production by states in 1957.
(e) Number of employees in the aircraft industry.
(f) Number of accidents on scheduled airlines.
(g) Number of homes built in the United States in 1957.
(h) Changes in the average level of prices.
(i) Amount of money in circulation.
(j) Number of man-hours of work lost by strikes during 1957.
(k) Number of students registered in colleges and universities in the fall of 1957.
(l) Expected enrollment of students in colleges and universities in 1965.

4. At a congressional hearing the comptroller of a large corporation stated that his company wanted the government statistical agencies to reduce the amount of data they were collecting from business firms. He stated that it would not only save the government money but it would also reduce the cost of filling out the reports, which had become a large expense for his company. A few days later a representative from the sales department of the same company appeared at the hearing and protested against any cuts being made in the amount of information collected by the government statistical agencies.

Comment on this situation. What criteria can an appropriations committee use in deciding how much information should be collected by a government agency?

5. Do you think it is a proper function of government to collect statistical information on the economy of the country and furnish it free, or at a price that merely covers the cost of printing and distributing the publications? If you think it is a proper function of government, how would you decide what information should be collected? It is obvious that all the external data businessmen want cannot be furnished.

6. Why are published guides to statistical data important tools for the staff of a business?

7. Reports submitted to the Bureau of the Census by business organizations are confidential, and the law prohibits the use of the reports in any way that will reveal information about an individual business concern. This sometimes results in information not being published for certain small areas, since the number of business concerns is so small that publication of the data for the area would reveal information about an individual concern. Not publishing some information leaves gaps in the data on certain areas, and considerable objection to the law on nondisclosure of census data has occasionally been raised. Do you agree that the law is wise, or would you recommend it be changed to permit the publication of data even though they revealed information about individual concerns?

8. If you were requested to recommend one published source of current statistical data (monthly) on business conditions and one published source of annual data, what would be your choice for each?

PROBLEM

A local businessman, in a small town in the southeastern part of the United States, where there is a large production of timber products, wishes to secure funds to finance the construction of a plant. The businessman has described to a banker his plans for building a plant to produce charcoal that would be sold for use in barbecue pits and for other forms of outdoor cooking.

You are asked to assemble all the information you can find on the production and sale of charcoal for outdoor cooking. Outline a program of checking that you would follow to be certain that you did not overlook any information of value to the banker in making his decision on advancing money to finance the project. (It is not necessary to compile the data; a description of the sources to be used is all that is required.)

5 COLLECTION AND TABULATION OF EXTERNAL DATA

After a businessman's problem has been defined and a careful search has been made for all available external data on the subject, the question must be raised as to whether the amount of information is adequate for a decision. If it is not, the next question is whether to decide the matter on the basis of what is available or to collect more. If it is decided to collect additional external data, further analysis is necessary to determine exactly what is needed to supplement the data already assembled. Certain parts of the problem may be solved completely by the information at hand. The analysis should isolate the portions for which no satisfactory solution is yet suggested and for which additional information is needed. Making the most effective possible use of data already published and collecting additional data if the published data available are not sufficient are both important; the amount of use that will be made of each technique depends upon the specific business and industry, the type of problems to be solved, and the amount of data already collected.

COLLECTING EXTERNAL DATA

Because so much material has been gathered by agencies specializing in the collection of external data, the executive may conclude that he need not concern himself with this phase of statistical technique. Some people believe that statistical method consists chiefly in the analysis and use of data already collected. It is doubtless true that most businessmen will make more use of data collected by some specialized agency than they will of data collected by their own organizations. Nevertheless, the collecting of statistical data is so important a function that businessmen should understand something about it, because when adequate facts are not already available the only alternative is to collect them.

Collecting Data About the Consumer

Newspapers and magazines must continually furnish their subscribers with reading matter, and the success of a publication in the long run depends upon whether the editors furnish material that subscribers want to read. Very little information is available that will tell the editor whether he is publishing the things his subscribers want to read; such data almost of necessity must be collected especially for the publication interested. One

common method of doing this is to interview subscribers, turning through the magazine or newspaper with them and, for each article or advertisement, asking whether or not it was read or noticed. When enough interviews are assembled to constitute a satisfactory sample of the subscribers, the editorial staff has information that indicates what people are reading and what they are not reading.

The manager of a television station is faced with the same problem, in that he must know what people with television sets want to view. About the best method of determining this is to find out *what* they view, in the same way that the best method of finding out what people *want* to read is to learn what they *are* reading. Because such information is so important, great numbers of surveys are made asking questions about televiewing habits, such as what stations are popular, what programs are viewed, etc.

It is perhaps safe to say that every business selling a product needs information about its customers. Exceptions are insignificant and relatively unimportant. Whenever the market for the product is extensive, the facts about the customers will be numerous and in most cases will be of such a nature that they can be reduced to quantitative form. In other words, information about the market will be statistical, and collecting, summarizing, and analyzing the facts are part of the statistical method.

Fig. 5-A below attempts to illustrate what happens when a manufacturer sells his products to a mass of consumers through the orthodox channels of distribution. It shows that goods go from the manufacturer through the wholesaler and retailer to the consumer. The manufacturer attempts to create a demand for the goods by various forms of selling effort,

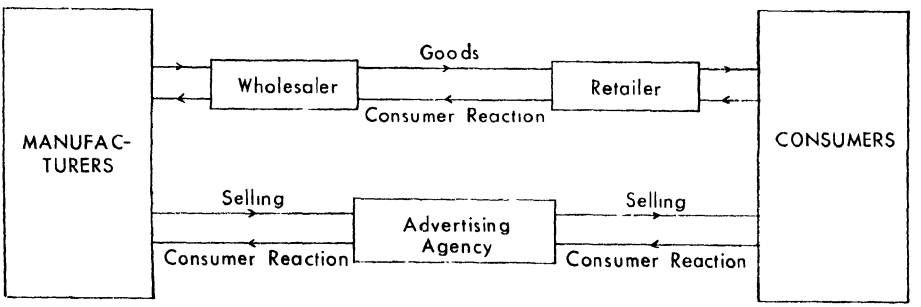


Fig. 5-A **Collecting Consumer Data Through Orthodox Channels**

usually grouped together and called advertising. This stream of selling effort goes through the advertising agency to newspapers, magazines, radio and television stations, car cards, outdoor posters, and other forms of advertising, and makes its impression on the consumer. By various means the manufacturer tries to persuade the wholesalers and retailers to push the product, and thus aid in the job of demand creation. This means that two

special devices have been set up: one to reach the consumer with sales appeal, and the other to get the goods into his hands when he decides to buy. However, no machinery is provided by this form of organization for getting information to the manufacturer from the consumer, no method by which the manufacturer can find out what is liked and what is not liked.

Since this information is so vital, there is usually an attempt to get it through the channels already set up for distributing and advertising. Fig. 5-A shows, therefore, a second line on each of its main divisions. This line represents consumer reaction traveling back to the manufacturer from the concerns in closer contact with the consumers. The retailers get customers' reactions and pass them on to the manufacturer either directly or through the wholesaler from whom they buy. The advertising agent secures information on consumer reaction, which he uses in future advertising campaigns and also passes on to the manufacturer. It will be observed that all this is a minor part of the activities of retailers, wholesalers, and advertising agencies, carried on as a sort of by-product of their chief activities.

The Consumer Research Division

In place of this uncertain method of securing facts about the consumer market, some businesses have a specialized division, such as Consumer Research, that has the sole function of collecting and analyzing information about consumers of the product. This division might be thought of as standing in relation to the other departments much as shown in Fig. 5-B.

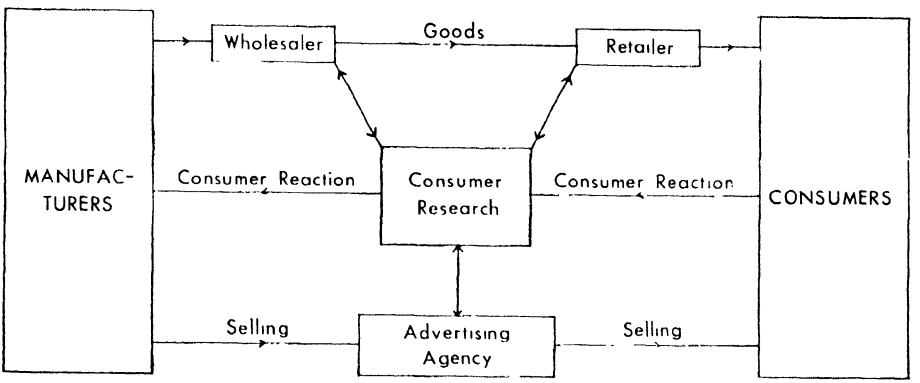


Fig. 5-B **Collecting Consumer Data Through the Consumer Research Division**

The consumer research division collects information primarily from consumers but also from retailers, wholesalers, the advertising agency, advertising media, and any other source that can supply significant data. This information is then made available to everyone concerned with the sale of the product, for it may have an effect on the form of the product, the package, the type of advertising, or the method of distribution.

An elaborate method such as this for collecting data can normally be employed only by a company with a large market and extensive distribution. But collecting original data is a technique that can be used to advantage by many different kinds of business organizations. It may be resorted to on a small scale whenever the information cannot be secured from some agency that has already collected it.

Companies employing large numbers of people with varying skills find it necessary to make periodic surveys of wage rates paid in their communities in order to be assured of being able to compete for employees. Competitive business is subject to constant change, which means that wage scales should be studied regularly to determine how they compare with those of competitors. The businessman who fails to keep his rates revised to conform with the prevailing rates in the labor market runs the risk of losing employees and fails to recruit the kind of people needed for new positions and replacements. Information about labor markets that is collected by public agencies is used to the extent that it is available, and business concerns encourage these agencies to collect the maximum amount of data that their budgets will permit. However, the need for accurate, up-to-date information is great enough to make it desirable for individual businesses to make their own collections of labor market data.

This chapter will discuss the collection of statistical data by the individual business for its own use, and the collection of statistical data by the agencies listed in Chapter 4. The problems of collection and tabulation of external data are essentially the same whether done by a business or public agency for use in carrying on its functions, or done by a statistical agency for distribution to others.

METHODS OF COLLECTING EXTERNAL DATA

The problem of collecting external data differs from that encountered in collecting internal data chiefly because the external data wanted are facts in the possession of someone not under the control of the organization collecting the data. This means that the facts must be secured either by observation or by asking questions. In each method, appropriate records are made of the observations or of the answers, and these records are the original documents from which the statistician prepares his tabulations.

If the facts are to be collected by observation, it is essential that the observer have some systematic method of recording his observations. If the facts are to be collected by asking questions, the questions are usually put on a form, called a *questionnaire* or *schedule*, that provides space for recording the answers. The proper designing of the questionnaire as one step in the planning of a statistical study is discussed on pages 91 to 94.

Observations

The scientific method requires observation of the results obtained from controlled experiments. When all the factors in an experiment are controlled, variations in the results obtained may be related to known variations in the factors. Business statisticians, however, usually must make observations in situations that cannot be controlled. Traffic studies, which provide useful information for planning streets and highways, make regular use of the observation method of collecting data. Observers may count the number of vehicles that pass a given point, or automatic counting machines may be used. Store location studies, as a measure of the potential business, usually provide for the counting of the number of persons walking past a particular location. Fashion magazine writers may observe the types of street clothing worn in order to secure information on style trends. The size of the audience viewing a given television program may be measured by a recording device, such as the audimeter, which is used by the A. C. Nielsen Company. The audimeter, when attached to individual receivers, indicates when the set is turned on and to what station it is tuned. Information as to the brands of food products used may be obtained by securing permission of housewives to take an inventory of the products on their pantry shelves.

The amount and type of data which the business statistician can obtain by observation is limited, since so many of the facts needed can be obtained only by asking questions of someone who has the desired information. When it is possible to secure the information by observation, this method is usually used in preference to asking questions. For example, the audimeter will generally give more accurate information on the listening habits of a family than could be obtained by asking what programs are listened to.

Questions

Because of the limited application of the observation method, asking for the information needed is the most widely used method of collecting external data. Any method of communication may be used in securing the information, but probably the most common questioning methods are mail, telephone, or personal messenger. The advantages of the different methods vary with circumstances. The best method to use is the one that will secure the information needed, with the greatest degree of accuracy, at the lowest cost, and often in the smallest amount of time. Using the mail is often the cheapest method, but it is notoriously weak in that it is not very successful in getting the information needed from every individual. Personal interviewers are much more successful in procuring the information, but their use is much more expensive than the mail. The telephone method

is cheaper to use than personal interviewers, but many kinds of information cannot be obtained accurately by telephone. Also, in some situations the person from whom the information is desired cannot be reached by telephone.

It has been amply demonstrated that the personal interviewer is able to get information from a much higher percentage of desired respondents than other methods of communication. However, the high cost of personal interviewing has induced statisticians to experiment with combinations of methods in an attempt to profit from the lower costs of other methods. For example, an attempt is first made to secure as many returns as possible by mail. Individuals failing to respond may then be called on the telephone and asked to send the information. If this fails to secure a reply, a personal interviewer might then be sent to secure the information.

If a report is needed periodically, such as every month, it may be possible to use a personal interviewer to secure the first report and to secure also a promise to mail the information for succeeding months. This method is used by agencies collecting information on current business conditions, and over a period of time it is learned which concerns will promptly report by mail and which ones will require a personal visit.

The best method of collecting data must be worked out for each situation by taking into consideration the specific factors in each case, for it is impossible to give rules that can be universally applied. The one test that must always be met is that the method secures the needed information. For example, if a complete census is being taken of an industry, information should be obtained from every concern in the industry. Even though it will probably not be possible to secure complete coverage, any method that fails to approach complete coverage should not be considered satisfactory.

If a sample survey is made, the desired information must be secured from every individual selected as part of the sample. It is just as important that information be secured from every individual in the originally selected sample as that every individual be included if a complete census is being taken. If any method or combination of methods fails to secure all the returns required, it should be recognized that the method is inferior. If cost considerations make it necessary to use an inferior method, it should be recognized for what it is.

Form in Which Facts Exist

The statistician collecting data from individuals or from business concerns should distinguish between facts that can be supplied from records, and facts, opinions, or estimates that are not recorded in any place. The records may be public records or they may be the private records of individuals or business concerns. Information on the public records may

ordinarily be obtained easily, although, of course, there are exceptions. It is obvious that information on the records of individual businesses will not often be made available to other concerns. On the other hand, a great deal of information, such as data on sales, production, employment, and profits may be collected from individual concerns by public agencies or trade associations. When such information is given out, it is usually with the understanding that the facts are confidential and will be released only as part of a total, in such form that no information will be shown about individual businesses.

An important part of the information released by governmental departments is collected from the individual records of business concerns, and a study of statistical methods used by such agencies would give considerable emphasis to the ways of collecting this type of information. The statistican for a trade association, a governmental department, or a state university bureau of business research will encounter a great variety of problems in getting the data wanted.

Probably most of the information collected by individual business concerns exists in the minds and memories of individuals rather than as data in recorded form. Information collected from consumers on commodities used and reasons for their use, data on reading habits and radio-listening and televiewing habits, and many other types of data collected for use by marketing executives represent facts stated verbally by some individual and existing in no formal record. Public statistical agencies are perhaps less concerned with this type of information than with data in recorded form.

The distinction between information which is secured from records and that which is not recorded in any place is important in designing a questionnaire. When the information has not been recorded, it is usually difficult to get accurate answers to any but the simplest questions; sometimes a great deal of skill is required to secure the facts. The proper wording of questions in marketing research and public-opinion research, for example, is a job for a skillful psychologist.

Questionnaire Design

To collect data that can be supplied from the records of a business is generally easier than to collect previously unrecorded data. The former type of data is usually collected by governmental or trade association statisticians. Unrecorded data is collected more often from individuals than from business concerns. The opinion polls and market studies are well-known activities that collect large amounts of unrecorded data by means of questionnaires.

The kind of questions used and the design of a questionnaire vary with the type of information that is to be recorded and whether the information is to be written on the form by the person supplying the information, or by the person collecting it. If the facts are to be supplied from records, it is necessary to know what records are kept and what information can be summarized from them. Governmental statistical agencies have made use of industry committees in order to know the kind of questions that industrial concerns could reasonably be expected to answer. When the information is not in recorded form, the questions must be phrased carefully to be certain that the desired information is secured.

Principles of questionnaire construction. A few general principles of questionnaire construction are discussed below, with particular emphasis on information secured from sources other than written records.

1. *Make the questionnaire as short as possible.* Sending out questionnaires is becoming so common that they are rated a public nuisance in many places.[1] Obviously this makes it essential to use the smallest number of questions that will give the information needed. Everything essential should be asked, but care should be taken to see that nothing nonessential is included. The greater the number of questions asked, the fewer the returns will be. When personal interviews are made, a long questionnaire takes more time and is therefore more expensive.

2. *Make the questions clear.* It would seem unnecessary to state this principle, but it is violated so often that its importance is apparently underestimated. A definite, clearly stated question has a better chance of being answered. If it is not perfectly clear and unambiguous, the answers will in most cases be worthless.

A questionnaire sent to consumers asked if they owned an ironer and, if they did, what make. The word "ironer" referred to the machine for ironing, popularly known as a "mangle"; but many people evidently interpreted "ironer" to mean an electric iron. At least, many answered "yes" and gave the make of an electric iron. In many other cases it was impossible to tell whether the person was referring to an ironer or an electric iron, so the answers were practically worthless.

One simple method of making questions clear is to indicate the form the answer is to take. For example:

Do you own a car? _____ What make? _____
(Yes or No) (Make)

It is difficult for anyone to misinterpret such a question, although of course some do.

Another method is to give all the possible answers, and ask that the correct one be checked. For example:

Do you read the food advertisements in magazines?

always_____; occasionally_____; seldom_____; never_____.
(Please check)

When care is taken to include all possible answers, this method is useful. It also helps when the answers are being summarized.

3. *Whenever possible confine the questions to facts, and facts that people can reasonably be expected to know.* Questions asking reasons why a certain article was bought and reasons for liking or not liking certain things are difficult to answer correctly. Sometimes a person thinks he knows, but does not; on other occasions he may know he does not know, but since an answer is insisted upon he gives one. Questions requiring a knowledge of easily ascertained facts are much more likely to get significant answers.

Questionnaires sent to business organizations frequently ask for detailed information which would require a great amount of time and effort to compile. If such facts are to be secured, it is usually necessary for someone in authority to be convinced that it is worth the time required to assemble them.

Questions that people cannot reasonably be expected to answer are those asking for personal or confidential information. Probably no one would like to have a stranger come to the door and ask the amount of the family income. Most people do not object to telling the kind of soap preferred by different members of the family, but giving information on the amount of the family income is an entirely different matter. Business executives may be expected to refuse to tell the amount of their total net sales, the amount of sales of certain products, the amount of the weekly payroll, the amount of net profit earned, and a great many similar facts.

Governments can require individuals and business organizations to report personal and confidential information; but when an individual is collecting data, they must be secured on a purely voluntary basis. In fact, many governmental agencies prefer to collect information on a voluntary basis.

4. *Avoid asking questions in a manner that will influence the answers.* Since an investigation is conducted to secure facts about a particular problem, it is important that answers give information that is as accurate as possible. In other words, all leading questions should be avoided. An example of a leading question is "Why do you prefer Our Electric Iron?" A better approach would be to ask first "Which electric iron do you prefer?" and then to follow up with "Why do you prefer it?"

It is unfortunately true that much information is collected to prove a particular case. When this is the purpose, it is difficult to avoid asking the questions in such a way that the desired answer is secured. When using information collected by someone else, the first question that should be asked is who made the collection. If this person has something to gain or lose from the results, the data should be used with caution. This does not mean that people are naturally dishonest; but when an individual will gain or lose according to the way certain facts show up in the returns from a questionnaire, he will unconsciously try to make the facts appear as favorable as possible.

Pretesting questionnaires. Because it is so difficult to anticipate all the problems that may arise in designing a questionnaire, it is usual to pretest the form with a small number of actual interviews or a small mailing before starting the survey. In either case it is possible to get enough replies to determine the strengths and weaknesses of the questionnaire before a full-scale survey is made.

Some typical examples of questionnaires are shown on pages 95 to 97. Fig. 5-C is a form used by the Bureau of Labor Statistics and the Bureau of Business Research at The University of Texas to collect data on new housing authorized. The form is filled out monthly by a city official, who performs a considerable amount of summarizing of individual building permits. Fig. 5-D is a typical questionnaire used in the study of market problems to collect data from consumers. Fig. 5-E is used to collect the total amount of retail sales each month from a panel of stores.

Sampling

Instead of collecting complete data with respect to a particular problem, it is possible to limit the collection to a portion, or a sample, of the whole group. Then the characteristics of the larger group may be estimated from the data collected from the sample. It is a rather general practice in collecting internal data to record all of the items and to base summaries on this complete compilation. The practice of collecting external data in this same manner has become increasingly expensive as the scope of business operations has increased and the amount of statistical information desired has become more extensive. Thus, the development of sampling methods that produce approximate data which are sufficiently accurate came as a result of the steadily increasing demand for more information.

Sampling methods of collecting data may be applied to either internal or external data. Some progress is being made in the substitution of samples for complete records of some kinds of internal data, but since costs are generally much greater in making complete compilations of external

B. L. S. 404 B
(Rev. 1-1-55)
U. S. DEPARTMENT OF LABOR
BUREAU OF LABOR STATISTICS
WASHINGTON 25, D. C.

FOR BLS OFFICE USE ONLY

KARDEX		EDIT		ADD		POST	
In	Tab	Made	Ver	Made	Ver	Mach	Hand

COOPERATIVE REPORT BUILDING PERMITS ISSUED and LOCAL PUBLIC CONSTRUCTION

Form approved
Budget Bureau No. 44-R778.5

THE UNIVERSITY OF TEXAS
College of Business Administration
Bureau of Business Research
Austin 12, Texas

If possible, please return one copy on or before the 8th of the month to above address.

Please report permits issued in month of—

If no permits were issued this month enter "X" here → ☐ and return this form

(If above mailing address is incorrect or zone number omitted, please indicate change)

Item No.	NEW CONSTRUCTION	Number of Buildings	Estimated Cost (Omit cents)	Number of Housekeeping Units	Code
1. One-family houses	(a) Detached				001
	(b) Semidetached and row				001
2. Two-family buildings					002
3. Apartment buildings	(a) Three- and four-family				003
	(b) Five or more family				004
	Leave blank →				094
4. Hotels, motels, tourist cabins, and seasonal camps				X X X X	006
	Leave blank →				031
5. Garages (affiliated with private residences)				X X X X	013
6. Gasoline and service stations				X X X X	014

7. **All other buildings and structures not reported above** Include privately and new publicly owned structures (State, county and municipal) started within your municipality even though no permit was required and report contract amount or best available estimated cost. Exclude streets, roads, and sewer and water lines.

Kind of Structure	Owner (Private, State, County, City, etc.) If estimated cost is $100,000 or more, list separately and give owner's full name and address	Number of Structures	Estimated Cost (Omit cents)	Code (Leave blank)
	Leave blank →			024

EXAMPLES FOR ITEM 7

Elementary school	City Board of Education, 27 West Oak Street	1	$150,000	X X X X
Stores	Private	2	$80,000	X X X X
Shirt factory addition	Acme Shirt Co., 2716 Temple Road	1	$100,000	X X X X

ADDITIONS, ALTERATIONS, AND REPAIRS (Enter single projects costing $100,000 or more under Item 7 above, enter all others under items 8 and 9 below.)	Number of Permits	Estimated Cost (Omit cents)	Code
8. To homes and apartments			025
9. To other buildings			027
Leave blank →			028

Leave this line blank: 30 _____ 1-4 _____ 6-23 _____

Name _____ Title _____

U. S. GOVERNMENT PRINTING OFFICE 16—80764-2

Fig. 5-C **Cooperative Report on New Housing Authorized**

Page 1

FAMILY INFORMATION—to be asked of a responsible adult family member

1. Including yourself, how many people are living in this household at present? . _____
(Include only people related by blood, marriage or adoption)

LIST EACH HOUSEHOLD MEMBER

Name	Relationship to Head	Sex	Age	Check Here If Employed Full Time	Interviewed
1. (Head)					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					

2. (a) How much did you and your family spend in stores on foods and groceries during the past seven days? . . . . . $_____
(b) Approximately what is the cash value of home grown food used at home during the past seven days? . . . . . $_____
(c) Approximately how much was spent by family members for meals bought outside the home during the past seven days? . . . . . $_____

3. How much did you and your family spend on cosmetic and drug items during the past seven days? . . . . . $_____

FOR OFFICE USE

1-	5-	9-	13-	17-	21-	25-	29-	33-	37-
2-	6-	10-	14-	18-	22-	26-	30-	34-	38-
3-	7-	11-	15-	19-	23-	27-	31-	35-	39-
4-	8-	12-	16-	20-	24-	28-	32-	36-	40-

Page 2

URBAN AND RURAL NON-FARM FAMILIES

4. (a) Do you own or rent your home 'apartment? . . . . Own_____ Rent_____
(b) In normal times approximately what would a home apartment such as this one rent for by the month in this community? . . . Monthly Rent $_____

Check Type of Structure
One family house . . _____
Two, three or four family . _____
More than four family . _____

5. Do you have a telephone in the family name? . . . . . Yes _____ No_____

6. Do you have a mechanical refrigerator? . . . Yes _____ No _____

7. Do you have electricity in your home? . Yes_____ No_____

8. Do you or someone in your family own an automobile? . . Yes_____ No_____
(If "Yes") Make and year of latest model? Make_____ Year_____
Number of cars owned_____

SHOW FAMILY INCOME CARD Into which of these groups would you estimate the combined family income falls?
Group No._____

9. Economic Status A_____ B_____ C_____ D_____

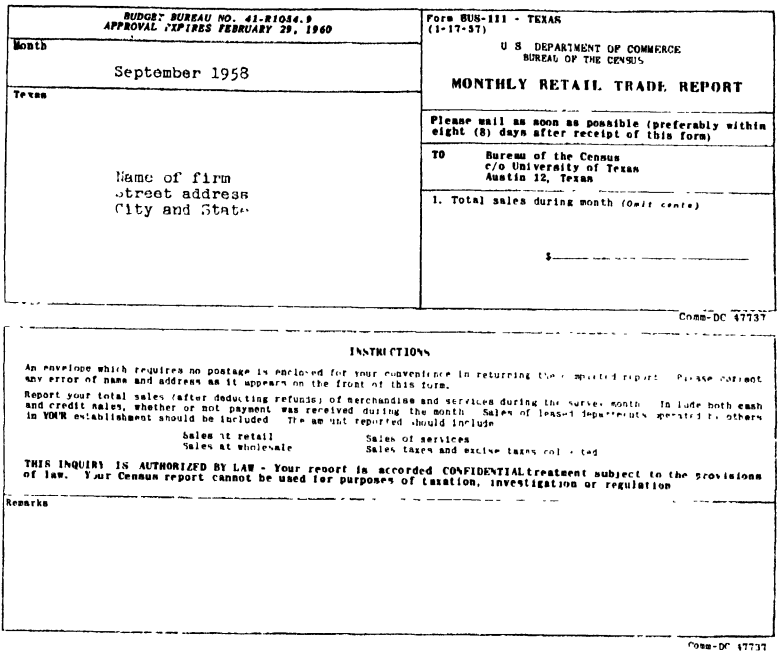
Enumeration District No._____City or Town_____State_____

INTERVIEWER_____ DATE _____

FOR OFFICE USE

41-	45-	49-	53-	57-	61-	65-	69-	73-	77-
42-	46-	50-	54-	58-	62-	66-	70-	74-	78-
43-	47-	51-	55-	59-	63-	67-	71-	75-	79-
44-	48-	52-	56-	60-	64-	68-	72-	76-	80-

Fig. 5-D **Marketing Research Questionnaire**



BUDGET BUREAU NO. 41-R1044.6
APPROVAL EXPIRES FEBRUARY 29, 1960

Form BUS-111 - TEXAS
(1-17-57)

U S DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

MONTHLY RETAIL TRADE REPORT

Month

September 1958

Texas

Name of firm
Street address
City and State

Please mail as soon as possible (preferably within eight (8) days after receipt of this form)

TO Bureau of the Census
c/o University of Texas
Austin 12, Texas

1. Total sales during month (Omit cents)

$_____

Comm-DC 47737

INSTRUCTIONS

An envelope which requires no postage is enclosed for your convenience in returning this [illegible] report. Please correct any error of name and address as it appears on the front of this form.

Report your total sales (after deducting refunds) of merchandise and services during the survey month. Include both cash and credit sales, whether or not payment was received during the month. Sales of leased departments operated by others in YOUR establishment should be included. The amount reported should include

Sales at retail
Sales at wholesale
Sales of services
Sales taxes and excise taxes collected

THIS INQUIRY IS AUTHORIZED BY LAW - Your report is accorded CONFIDENTIAL treatment subject to the provisions of law. Your Census report cannot be used for purposes of taxation, investigation or regulation

Remarks

Comm-DC 47737

Fig. 5-E **Monthly Retail Trade Report**

data, it is not surprising that the greatest use of sampling has occurred in the collection of external data. It is also true that many kinds of accounting data require that complete records be kept, such as records of cash, accounts receivable, and accounts payable. However, it is also true that many kinds of internal data would be just as useful if they were only approximate; the continued use of complete records probably is often a result of habit rather than a careful appraisal of the needs for accuracy.

The problems of designing a sample of units to collect that will be representative of the whole group of items, and the related problems of determining the degree of accuracy with which the sample measures the characteristics of the larger group, are discussed in Chapter 10. The collection of sample data is not significantly different from the collection of complete data; the only difference between the two types of data collection is the number of individual units covered in the collection of the data.

TABULATION OF EXTERNAL DATA

The tabulation of external data is not fundamentally different from the tabulation of internal data. All the methods described in Chapter 3 may be used, although certain differences in treatment do exist. Internal data are usually recorded continuously with summaries made periodically, such as weekly or monthly. Statistical surveys of external data, however, are

usually summarized by accumulating replies until they have all been received and then tabulating directly from the questionnaires rather than entering the data in a record from which summaries are later derived. When records are maintained as the information accumulates (the typical method with accounting and internal statistical data), these records usually duplicate to a certain extent the original business papers from which the information was taken. In other words, internal statistical records, like accounting records, become a fairly complete substitute for the original documents. In nearly all tabulations of external data, however, the summaries are made directly from the questionnaires. This distinction is not always true, since with some machine processes the summary of internal data may be made directly from such business papers as invoices; and with some tabulations of external data, a complete duplicate record is made.

It is important to remember that the plan for tabulating data should be drawn up at the time the survey is planned. The tabulations required will be a factor in the design of the questionnaire and in the selection of the sample. It is desirable to outline just what tables are to be prepared and how the tabulations are to be made, and then to determine the form of the questionnaire and the sample needed to secure the required accuracy.

The following pages contain brief comments on the effectiveness of tabulating external data by means of the methods discussed in Chapter 3 for internal data.

Multiple-Column Record

The multiple-column work sheet can be used to classify external data in exactly the same way it can be used to record internal data. The amounts to be totaled by classes can be distributed to columns according to classification. This method, however, is not used nearly so often for tabulating statistical data collected by a survey as for internal data recorded currently. When tabulating survey data, all the information is normally in hand before starting to classify it. The totals for the classes can be accumulated directly from the schedules, which can be sorted into the desired classes before adding the individual amounts. This saves writing the amounts in the columnar record. But when internal data are accumulated currently throughout the month, it is generally worth while to enter the data as they are received on the multiple-column record.

Tally Sheet

The tally sheet, which was described in Chapter 3 as basically the same kind of record as the multiple-column record, is universally used as a hand method when the tabulation involves counting the number of items in each class. A great many tabulations of external data do involve such counting.

Table 5-1 gives the gross margin for each one of 238 lumber yards, and Table 2-8 (page 30) gives a classification of these percentages according to

Table 5-1

GROSS MARGINS OF 238 LUMBER YARDS, 1956

RATIO OF GROSS PROFIT TO NET SALES
(PER CENT)

Firm no.	Ratio	Firm no.	Ratio	Firm no.	Ratio	Firm no.	Ratio	Firm no.	Ratio	Firm no.	Ratio
1A	21.5	257B	23.7	475C	29.8	738C	20.9	925B	20.4	1099B	18.5
14C	19.6	264C	21.9	488C	17.5	740C	22.7	930A	25.3	1102D	19.0
16C	21.4	270D	21.6	489D	18.8	741C	28.6	931A	23.1	1118A	27.1
21B	31.4	274C	33.1	512C	24.9	744D	19.9	935B	30.0	1121C	23.2
23B	24.0	280D	24.1	520A	26.5	745C	17.0	943C	32.1	1122A	26.3
24B	27.4	281A	28.7	524A	17.9	746C	19.3	949B	35.9	1148A	28.5
28D	30.5	282A	23.5	525C	17.9	754B	20.3	950B	25.1	1153A	31.5
30B	29.8	288C	31.2	526A	18.3	755F	16.1	951D	22.8	1155A	14.9
43B	27.4	293F	22.5	536B	25.8	756E	16.2	958B	21.0	1156A	17.1
47D	20.6	294D	21.5	537A	26.0	758A	31.9	959A	21.1	1157C	22.6
48D	27.6	295E	11.2	542A	24.6	762B	20.7	966A	25.8	1158D	15.5
55B	24.1	296D	16.6	555B	23.8	764B	28.0	971B	23.1	1167C	19.0
65C	23.0	297E	24.7	557C	19.1	772B	25.8	973B	28.8	1169C	31.7
72C	24.4	303C	21.7	560C	24.3	776C	20.9	974B	31.5	1170A	25.3
80B	22.7	306C	23.9	563E	21.0	783A	30.9	975B	17.6	1176A	24.0
104D	38.2	309D	27.4	569D	13.7	785B	29.9	976D	13.2	1177A	21.2
108C	24.6	310F	24.2	574D	18.9	801A	17.3	989A	24.8	1185B	21.2
114A	24.0	314E	22.1	575D	18.2	802B	17.4	997C	23.3	1186E	18.2
122F	20.5	315D	22.0	582C	29.1	804B	21.0	1000A	18.7	1193C	25.4
133A	15.5	326D	25.2	585C	24.0	807A	25.6	1009C	23.1	1194C	26.4
134B	25.4	337C	21.5	595B	22.9	809A	18.9	1023B	17.8	1197C	30.8
143C	21.7	341E	19.4	596E	23.0	814B	28.6	1024B	26.7	1198A	21.3
147B	29.6	345B	22.2	601A	15.8	822C	22.9	1029B	25.7	1200D	20.9
163B	29.5	347A	11.9	607A	21.3	828B	23.6	1030B	27.2	1201D	21.3
165B	27.3	358A	16.5	614A	25.7	832A	31.7	1031C	18.6	1203D	33.1
166D	22.6	370A	24.8	617D	14.4	839A	23.6	1037C	25.1	1206C	20.2
168C	22.1	382A	20.4	619F	19.6	844C	22.7	1043C	31.8	1210D	18.9
171C	15.0	387C	25.4	623F	13.5	848B	26.3	1054D	26.1	1211B	25.9
174C	21.9	393D	24.5	632E	18.4	850B	24.1	1059D	18.0	1218C	21.6
178A	31.1	398C	29.6	642E	16.9	852D	19.5	1060B	19.4	1230B	13.8
182A	23.2	400C	27.3	644A	23.5	856B	16.6	1062A	22.1	1231A	18.8
190A	27.0	406A	20.1	648C	18.9	860E	24.8	1065C	21.6	1233A	29.8
191C	23.7	417D	14.5	651B	33.5	873D	19.3	1066B	27.8	1237B	22.9
192B	22.6	419B	24.6	656A	32.4	874D	20.8	1068A	19.8	1241C	27.7
194D	21.7	425A	29.1	689B	13.2	878C	25.1	1078B	15.8	1243A	27.1
215B	23.2	441D	28.7	693C	18.8	879A	17.6	1083C	30.3	1248A	20 5
225B	13.8	446E	16.8	702B	30.7	884C	28.5	1087B	22.8	1253C	21.8
228B	19.6	452A	18.2	703C	20.8	888B	31.9	1089A	36.8	1259A	27.0
240B	21.6	453A	25.6	712D	19.7	901C	24.1	1097B	18.3	...	...
247B	24.8	465C	22.6	719B	24.5	916C	27.4	1098B	20.3	...	...

SOURCE: Reports on file in the Bureau of Business Research, The University of Texas.

number of yards. The following tally sheet was used in making this classification. The problem of defining the classes for this kind of tabulation (a frequency distribution) was discussed in Chapter 2 on page 25.

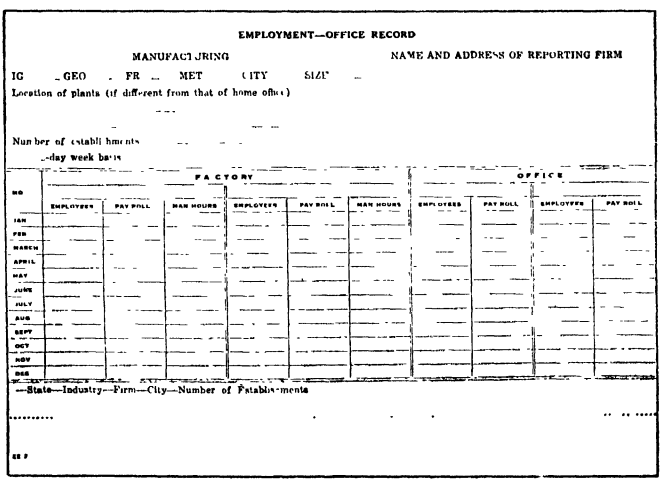
Ratio of Gross Profit to Net Sales
238 Lumber Yards
1956

Ratio	Tally	No. of yards
10.0 — 12.4	//	2
12.5 — 14.9	𝍸 ////	9
15.0 — 17.4	𝍸 𝍸 𝍸 /	16
17.5 — 19.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 //	37
20.0 — 22.4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ///	43
22.5 — 24.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 //	52
25.0 — 27.4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	35
27.5 — 29.9	𝍸 𝍸 𝍸 𝍸	20
30.0 — 32.4	𝍸 𝍸 𝍸 ///	18
32.5 — 34.9	///	3
35.0 — 37.4	//	2
37.5 — 39.9	/	1
Total		238

Fig. 5-F **Tally Sheet**

Ledger Sheet

When external data are collected regularly from a number of reporting firms, the agency commonly uses a ledger sheet to keep the record by establishments. The following ledger sheet is used by the Bureau of Business


EMPLOYMENT—OFFICE RECORD

MANUFACTURING NAME AND ADDRESS OF REPORTING FIRM

IG GEO FR MET CITY SIZE

Location of plants (if different from that of home office)

Number of establishments

-day week basis

	FACTORY						OFFICE			
MO	EMPLOYEES	PAY ROLL	MAN HOURS	EMPLOYEES	PAY ROLL	MAN HOURS	EMPLOYEES	PAY ROLL	EMPLOYEES	PAY ROLL
JAN										
FEB										
MARCH										
APRIL										
MAY										
JUNE										
JULY										
AUG										
SEPT										
OCT										
NOV										
DEC										

—State—Industry—Firm—City—Number of Establishments

Fig. 5-G **Ledger Sheet**

Research of The University of Texas to record the employment and payroll data reported monthly by individual business concerns. A separate card is set up for each firm and when the data are received by the Bureau each month, the information is posted to the card as a permanent office record for the establishment.

Unit Records

Questionnaires may be used as unit records and sorted into groups for counting or for adding amounts from them by groups. Some of the best-known mechanical tabulation devices, however, use the unit record as a basis, and the importance of this method is largely in the application of machine methods. A more extensive discussion follows on the next page.

Exhaust Method

This method is used for external data on questionnaires in the same manner as for internal data, although it is not widely used for either type.

Pegboard and Summary Strips

This method can be used effectively in totaling amounts on questionnaires. The individual questionnaire is used as a strip, and the totals for the group are entered on a blank questionnaire. When the questionnaires are to be sorted into numerous groups for cross tabulation, this method is an economical means of accumulating the totals desired.

Multiple-Register Machines

The multiple-register machines are used for the same type of questionnaire as the pegboard. These machines permit entering several figures from the schedule, one in each register of the machine, before proceeding to the next one. This method normally will be faster than leafing through the papers many times, and it is unnecessary to fasten the forms on a special board, as must be carefully done with the pegboard. The machine is more frequently purchased for use on internal records than for tabulations of external data; however, for surveys that involve adding figures from a large number of questionnaires, multiple-register machines are efficient.

Posting Machines

Except for unusual situations, posting machines are not adapted for tabulations of external data but are used chiefly for accounting records and internal statistics.

Machine Tabulation of Unit Records

Needle-sorting devices. This method can be used for external data as easily as for internal data. It provides rapid sorting, which makes it useful when numerous cross classifications are needed. Sorting, however, is not usually the most time-consuming step in tabulation, and the needle-sorted cards must be counted or amounts from them must be added in the same manner as if questionnaires were used. This factor definitely limits the usefulness of the method for large-scale collections of data.

Punched-card tabulation. From 1890 until the years immediately following World War II, punched-card equipment was the most highly mechanized device for tabulating both internal and external data. Today electronic data processing has taken over the larger jobs that deal with tabulation of external data. A prime example of such a job is the tabulation of data for the U. S. Census. However, punched-card tabulation is still a major technique that is used to secure many large collections of external data, and it is not likely that this technique will be displaced immediately.

In the years following 1880 the Bureau of the Census was engaged in the problem of tabulating the results of the Census taken in 1880 as required by law. The tabulation was not completed until 1887, only three years before the eleventh Census was due. It was easy to foresee that the time was not far distant when the volume of work would make it impossible to publish the results of one census before time for taking the next one.

During the summer of 1889 a commission of three experts, appointed by the Superintendent of the Census, made a series of tests of all existing mechanical systems of tabulating data. As a result of this study, the system devised by Dr. Herman Hollerith, a statistician in the Bureau, was officially adopted for compiling census returns. The 1890 Census was tabulated in about two and one half years, whereas tabulation of the 1880 Census had taken seven and one half years.

In 1896 Dr. Hollerith incorporated his business under the name of the Tabulating Machine Company; later, through merger, it became a part of International Business Machines Corporation.

Electronic data processing. The developments in data processing that have taken place since the end of World War II have made it possible to tabulate both the regular census reports and current census data much more speedily. Punched-card tabulating machines enabled the Bureau of the Census and other statistical agencies to process the growing volume of statistical data faster than had been previously possible for a much smaller volume. Electronic data processing is now making possible the compilation of a still greater volume of statistical data in less time, and the use of electronic data processing by both public and private statistical agencies seems

certain to grow. The rise in population and the expanding economy are creating a demand for more statistical data and, at the same time, the amount of data in each study, such as a census, continues to grow larger. All these factors seem to insure that the use of electronic data processing for tabulating external data will continue to increase in importance.

STUDY QUESTIONS

1. Does the collection of external data offer any problems that are basically different from those encountered in collecting internal data? Discuss.

2. Does the tabulation of external data offer any problems that are basically different from those encountered in tabulating internal data? Discuss.

3. What are the advantages of collecting external data by observation rather than by asking questions? What are the advantages of asking questions rather than making observations? Which of the two methods is more widely used in collecting external data for business? Why?

4. What factors determine whether personal interviewers, mail, or the telephone should be used in a survey?

5. What are the chief obstacles encountered in collecting information that has not been recorded but must be supplied from memory? Give an example of this type of information.

6. What are the chief obstacles encountered in collecting information that is in written form in the files of a business concern? Give an example of this type of information.

7. In studies of consumer buying habits and related subjects, it is often helpful to seek the advice of a psychologist in designing a questionnaire that will be used to collect data. On the other hand, it is much less likely that a psychologist would be employed to design a schedule that would be used to collect data from a business concern. What characteristics of the two situations mentioned would influence the usefulness of a psychologist in the design of the questionnaire?

PROBLEMS

1. Assume that you are employed in a large city by a dairy that has been manufacturing *Mellorine* — a frozen dessert made of vegetable fat instead of butterfat. *Mellorine* has been sold, at a substantial reduction in price, in competition with your brand of ice cream. You have analyzed sales figures for the past three years and find that the percentage of *Mellorine* sales has increased steadily. You have tabulated the sales data by areas of the city and find that the increase in the percentage of *Mellorine* sales has been greatest in the areas with the lowest income. It is felt that the popularity of *Mellorine* in low-income areas can easily be attributed to the fact that the price of *Mellorine* is about 40 per cent below the price of the best grade of ice cream that you market.

Questions have regularly arisen regarding the consumer acceptance of *Mellorine* — such questions as whether the increase in sales has been due entirely to price, and whether consumers can tell any difference between *Mellorine* and other brands of ice cream even if they know what the difference is. The label explains what *Mellorine* is made of, but there is some question as to whether consumers read the labels. Finally, a decision is made to collect information about the consumer acceptance of *Mellorine*, in the hope that questions such as those given above can be answered.

(a) You are instructed to design a questionnaire that will supply management with the information needed about the sale of *Mellorine* and ice cream. It is considered important that each product be fitted to its proper market, but management does not know whether it possesses all the facts. The purpose of your survey is to supply all the facts that can be collected from consumers about *Mellorine* and its acceptance.
(b) Draw up a plan for tabulating by hand the information secured by your questionnaire.
(c) Design a machine tabulating card to be used in tabulating the data collected.
(d) Set up tables, according to form, with stubs and captions (at least for the classes you are reasonably certain will be included in the replies from the consumers).

2. The following table shows the ratio of credit sales to net sales for 112 department stores in the United States in 1956.

(a) Arrange these ratios in a simple frequency distribution.
(b) Explain briefly what the frequency distribution shows about credit sales in relation to total sales of these department stores.
(c) Derive a cumulative frequency distribution from the data on a "less than" basis.
(d) Derive a cumulative frequency distribution from the same data on a "more than" basis.
(e) *Retain this solution for use in Problem 5 of Chapter 7.*

RATIO OF CREDIT SALES TO NET SALES IN 112 DEPARTMENT STORES, 1956

78.5	45.4	80.9	39.7	55.1	50.2	70.4	68.3
76.5	61.2	75.1	59.7	58.5	64.0	56.9	94.9
55.1	78.8	55.7	69.2	30.9	34.5	66.3	57.6
56.6	58.2	64.1	54.7	62.8	57.5	59.7	56.4
20.8	75.2	54.9	91.1	49.1	35.4	53.3	65.2
56.9	73.0	44.0	47.4	58.4	40.1	64.7	56.5
40.5	60.0	47.6	72.5	63.2	51.2	69.2	52.3
58.4	59.1	59.2	54.8	37.3	49.6	65.7	60.5
44.2	84.4	70.2	71.2	59.4	57.4	61.8	63.7
62.2	59.7	58.4	48.2	73.6	48.0	60.2	41.2
75.8	52.0	67.1	51.0	87.3	44.5	58.6	59.4
31.5	77.7	53.2	58.9	79.0	67.8	74.8	63.3
75.1	56.9	69.3	59.2	71.6	71.2	52.9	30.7
65.7	66.5	61.7	70.5	59.7	57.9	48.9	30.5

SOURCE: Reports on file in the Bureau of Business Research, The University of Texas.

Part

II STATISTICAL DESCRIPTION

chapter 6 ♦ Statistical Analysis

chapter 7 ♦ Charts

chapter 8 ♦ Averages

chapter 9 ♦ Dispersion

6 STATISTICAL ANALYSIS

Statistical analysis is concerned with the application of technical operations and theoretical considerations to the compilation, presentation, discussion, and interpretation of numerical data. The purpose of statistical analysis is to get as much significant information as possible out of a given amount of numerical data. Any operation performed on the data for the purpose of obtaining more information is called *statistical analysis*. The amount of analysis required varies greatly with different situations and with different kinds of data. Detailed analysis is usually applied to frequency distributions and time series more than to other types of classifications, although it would be a mistake to conclude that classifications based on differences of kind and geographical divisions are not important to business executives. Furnishing information to executives in a form not needing further analysis is one of the most important functions of many statisticians' jobs. The use and interpretation of information in this form is one of the most familiar contacts many executives have with business statistics.

For example, an investigation may be made to determine the reasons that people are not buying a particular product. When the reasons have been tabulated, the classification probably gives the sales executive the information he wants. Further analysis is not needed; interpretation by the executive is the next step. In other words, he must try to translate the information into changes in policy that will stimulate sales.

PURPOSE OF STATISTICAL ANALYSIS

The aim of all statistical analysis in business is to put into the most significant form the information on a problem that will help in the solution of that problem. Enough analysis of the data should always be performed as soon as they have been classified so that there is no need for additional analysis. But if further analysis will add to the usefulness of the data, it is desirable that the proper methods be used. Statistical analysis is a technical matter, but its methods have many important applications to the problems of business management; and an intelligent use of the results obtained requires that executives understand statistical methods and the significance of the results.

It is not possible to make a clear-cut break between the collection of data and their analysis. When collecting data for a particular problem, the analysis starts with the first tabulations; but when published data are used, the first tabulations have already been performed and the user's analysis must start with whatever classification he is able to get. The problems of tabulating both internal and external data have been discussed in previous chapters, with emphasis on the methods of performing the different operations. When classification of data was discussed in Chapter 2, it was indicated that classification represents a fundamental phase of statistical analysis. This operation in some form is almost always applied to data, the chief exceptions occurring when the number of individual facts is very small.

RATIOS, PERCENTAGES, AND RELATIVES

Since much analysis of statistical data involves comparisons between classes, the use of ratios, percentages, and relatives is an important part of the methods of statistical analysis. These devices are simple yet effective, and an understanding of the underlying principles is a requirement for anyone who expects to analyze or use statistical data in any way.

Ratios

A *ratio* is a fraction used to show the magnitude of one quantity relative to the magnitude of another. A common use of ratios in business is in the analysis of accounting statements, but this type of analysis will be considered a part of accounting methods and will not be discussed here. Statistical ratios are the same in principle, however, and are different only in that they are applied to different kinds of numerical data.

A ratio can consist of any two numbers. For example, the 1954 cottonseed production was 875,000 tons in Mississippi and 1,808,000 tons in Texas. These facts might be expressed by saying that the ratio of the 1954 cottonseed production in Texas to production in Mississippi was 1,808,000 to 875,000. It is more likely, however, that the ratio would be stated as approximately 2 to 1; or, it might also be stated as 2.07 to 1.00; or as 207 to 100. All these ratios give practically the same information, but some show it more clearly than others.

The advantage of using the ratio of 2 to 1 is that small numbers are easier to compare than large numbers. The use of the ratio 2.07 to 1.00 or 207 to 100 involves a larger number of digits, but the fact that one of the numbers is 1.00 or 100 makes the comparison fairly easy. The most effective ratios are usually those in which one of the numbers is some power of 10. The numbers one and one hundred are probably most commonly

used as one number in the ratio, although other powers of 10 are also used on occasions. *An important principle to remember is that whenever a ratio is to be used in further calculations it should be in the form of a ratio to one.*

A basic problem in computing a ratio is to decide which of the two numbers should be made a power of 10. If one is interested in comparing Texas production with Mississippi production, it is easier if the smaller value, Mississippi production, is assigned a value of 1.00 (or 100). It would be stated that the ratio of Texas to Mississippi production was 2.07 to 1.00. If one is interested in comparing Mississippi production with Texas production, however, the ratio would be stated as .48 to 1.00. The decision as to which number should be given the value of 1.00 is based on the comparison that is wanted; there is no way of stating which should be made to equal 1.00 until it is decided which number is to be compared with the other.

Percentages

When one number in a ratio is 100, as the ratio 48 to 100, it is called a *percentage* (from the Latin *per centum*, per hundred). It could be said that in 1954 Mississippi cottonseed production was 48 per cent of Texas production. Or if the comparison was Texas with Mississippi, one would say that Texas cottonseed production was 207 per cent of Mississippi production. The percentage form is the most generally used of all the forms in which a ratio may be expressed.

Computation of ratios and percentages. In computing the ratio of Texas cottonseed production to Mississippi production, the following operations were performed.

$$\text{Ratio} = \frac{1{,}808{,}000}{875{,}000} = \frac{2.07}{1}, \text{ which is the ratio 2.07 to 1}$$

The ratio of Mississippi production to Texas production is computed as follows:

$$\text{Ratio} = \frac{875{,}000}{1{,}808{,}000} = \frac{.48}{1}, \text{ which is the ratio .48 to 1}$$

If it is desired to compute the ratio to 100, the above results are simply multiplied by 100, and become 207 to 100 and 48 to 100. Another way of writing a ratio to 100 is to use the term per cent as, for example, in the following statement: Mississippi production was 48 per cent of Texas production of cottonseed in 1954.

The above computations follow the general rule that the computation of a ratio to 1 is performed *by dividing the number to be compared by the number with which it is to be compared.* The number used as the basis of

the comparison is called the *base*, and is always the denominator of the ratio. It should always be remembered that a ratio is a fraction. When the comparison is Mississippi with Texas production, the ratio is $\frac{875,000}{1,808,000}$ or $\frac{.48}{1}$, and Texas production is the base. When the comparison is Texas with Mississippi production, the ratio is $\frac{1,808,000}{875,000}$ or $\frac{2.07}{1}$, and Mississippi production is the base.

When the ratio is a percentage, which means the base is made to equal 100, both the numerator and the denominator of the resulting fraction are multiplied by 100. The ratio of Texas to Mississippi production then becomes $\frac{207}{100}$ or 207 to 100, which may be stated as a percentage by saying that Texas production was 207 per cent of Mississippi production. If the base is to be made equal to ten, one thousand, ten thousand, or any other power of ten, the same computation rule is followed and both the numerator and the denominator of the fraction are multiplied by the number the base is to equal. When a ratio has been multiplied by some number such as 100, this must be taken into account if the ratio is used in further calculations. This fact is the reason for the principle stated on page 107 that ratios to be used in further calculations should use 1.00 as the base.

There are two common methods of stating the relationship of two numbers in the percentage form. The percentage may be computed from the two numbers and thus express the relationship between them, or it may be the relationship between one of the numbers and the difference between the two numbers. For example, 1954 cottonseed production in Mississippi was 875,000 tons and in Texas was 1,808,000 tons. Texas production was computed on page 107 to be 207 per cent of Mississippi production. Another way of showing the relationship would be to say that Texas production was 933,000 tons more than Mississippi production and that this was 107 per cent *more than* Mississippi production. The computation of this percentage is as follows:

$$\text{Percentage} = \frac{1,808,000 - 875,000}{875,000} \times 100 = \frac{933,000}{875,000} \times 100$$

$$= 1.07 \times 100 = 107 \text{ per cent}$$

This is merely another way of expressing the same information, since 107 per cent more than Mississippi production is the same as 207 per cent of Mississippi production.

In similar manner, it would be computed that Mississippi production was 52 per cent less than Texas production.

$$\text{Percentage} = \frac{875{,}000 - 1{,}808{,}000}{1{,}808{,}000} \times 100 = \frac{-933{,}000}{1{,}808{,}000} \times 100$$
$$= -.52 \times 100 = -52 \text{ per cent}$$

On page 107 it was found that Mississippi production was 48 per cent of Texas production. This means that Texas production is taken as 100, and 48 is 52 less than 100. The minus sign indicates that it is 52 per cent *less than the base*; when the sign is positive, it means *more than the base*. In other words, a ratio may be expressed as a given percentage of the base or as a given percentage more than or less than the base.

Determination of the base for ratios. Probably the most difficult part of a ratio analysis for most people is the decision as to which number should be used as the base. This was mentioned briefly on page 107, but it will be discussed further here even at the risk of repetition. The only rule that will invariably indicate the base is that *the number with which the comparison is being made* is the base (denominator). For two numbers it is always possible to compute two ratios, but the ratio that is wanted depends on the comparison to be made.

An example of this is a well-known percentage used by retailers to represent gross profit. The modern practice is to express the gross profit (the difference between cost price and selling price) as a percentage of the selling price. An item selling for $5 and costing $3 would have a gross profit of $2, or 40 per cent of the selling price.

$$\text{Percentage} = \frac{\$2}{\$5} \times 100 = 40 \text{ per cent}$$

The department store executive looks at this ratio and knows that he has 40 per cent of every sales dollar to cover operating expenses and to yield a profit. Some people, however, think of the gross profit as 66⅔ per cent of the cost of the item.

$$\text{Percentage} = \frac{\$2}{\$3} \times 100 = 66\tfrac{2}{3} \text{ per cent}$$

This percentage means that the profit was 66⅔ per cent of the cost of the item. When thinking of the rate of gross profit on the investment, this percentage is significant, but it does not give the same information as that given by the percentage of gross profit to selling price.

When referring to percentage of gross profit, it is necessary to state clearly what is used as the base. It may be stated in the above example that gross profit is 40 per cent of selling price or 66⅔ per cent of cost price.

It is not enough to state simply that gross profit is 40 per cent or 66⅔ per cent. In other words, a statement of the base with which the comparison is being made is an essential part of the statement of any percentage.

One further caution in the choice of the base is that the number used as the base should not be abnormally small. Preferably it should be more than 100, since the percentage analysis reduces the data to a "per one hundred" basis.

When the proposal to permit women to vote was being debated in one state, a speaker opposed to woman's suffrage stated that in another state the number of women in the penitentiary increased 100 per cent after women were given the vote. An investigation of the facts disclosed that his statement was correct; the number had increased from one to two. It is not valid statistical analysis, however, to report the percentage without calling attention to the numbers on which it is based when the numbers used are very small.

Ratio of parts to the whole. An important form of analysis is the use of percentages to express the relationship of each class of data to the total of all the classes. These are referred to as *distributive percentages.* Table 6-1 shows the uses of energy in the United States in 1953, classified according to end-uses. The use of energy in highway vehicles was 5,277 trillion British thermal units (B.T.U.), out of a total of 38,472 trillion B.T.U. used in the United States. When energy use in highway vehicles is expressed as a ratio to the total use of energy in the United States, it is found to be in the ratio of 13.7 to 100.00; or it may be said that the use of energy in high-

Table 6-1

END-USES OF ENERGY IN THE UNITED STATES, 1953

(TRILLIONS OF B.T.U.)

End-use	Trillions of B.T.U.	Percentage of total
Industrial and miscellaneous	16,424	42.7
Space heating	7,410	19.3
Production of electricity and gas	6,590	17.1
Highway vehicles	5,277	13.7
Railroads	1,399	3.6
Ships	832	2.2
Aviation	540	1.4
Total	38,472	100.0

SOURCE: "Natural-Gas Statistics" by Richard C. Henshaw, Jr., Bureau of Business Research, The University of Texas, 1955, p. 100.

way vehicles in 1953 was 13.7 per cent of the total use of energy in the United States. The table gives the distributive percentages, which show the relationship of the energy consumed in each end-use to the total energy consumed in the United States.

The advantage of using distributive percentages rather than the actual amounts may be readily seen by once again referring to Table 6-1. The comparison of the amount in a class with 100 is much easier than with the large number 38,472. Whenever the analysis of the data requires a comparison of the individual classes with the total of all classes, a percentage distribution is most useful.

In a percentage distribution the total is, of course, the base, and the individual numbers are each expressed as a percentage of this base. This means that each of the numbers is divided by the total, and the resulting quotients are multiplied by 100. The calculation of the percentage that highway vehicle use is of the total is as follows:

$$\text{Highway vehicles percentage} = \frac{5{,}277}{38{,}472} \times 100 = .137 \times 100 = 13.7 \text{ per cent}$$

In dividing several numbers by the same number, as in the computation of a percentage distribution, one can save time by multiplying each number by the reciprocal of the divisor. A *reciprocal* of a number is defined as the quotient of one (unity) divided by that number. In the above example, the reciprocal of 38,472 is $1 \div 38{,}472$, or .00002599. The percentage distribution for highway vehicles is then calculated as follows:

$$5{,}277 \times .00002599 = .137, \text{ or } 13.7 \text{ per cent}$$

The percentage distribution for each of the other end-uses would be determined in a like manner.[1]

Another important use of percentage distributions is the comparison of two sets of data classified according to the same classes. Table 6-2 gives a classification of weekly salaries of women calculating machine operators in New Orleans and St. Louis. There were 349 women in the New Orleans sample and 1,043 from St. Louis. This study was based on samples rather than on complete enumerations. Since the number of women from New Orleans was so much smaller than the number from St. Louis, it is not proper to compare the number making a given salary in New Orleans with the number making the same salary in St. Louis. For example, the number of women making between $32.50 and $34.99 was 47 in New Orleans and 132 in St. Louis; but the significance of this fact is not apparent from this simple comparison, since there are almost three times as many women in

[1]When using a Monroe, Marchant, or Friden calculator, the reciprocal of the divisor is set up in the machine and successive multiplications are made without changing the reciprocal. Multiplying by the reciprocal of a number is the same as dividing by that number.

Table 6-2

WEEKLY SALARIES OF WOMEN CALCULATING MACHINE OPERATORS IN NEW ORLEANS AND ST. LOUIS

Salary (Dollars)	Number of employees		Percentage of total	
	New Orleans	St. Louis	New Orleans	St. Louis
22.50 — 24.99	2	...	.6	...
25.00 — 27.49	24	3	6.9	.3
27.50 — 29.99	23	28	6.6	2.7
30.00 — 32.49	59	52	16.9	5.0
32.50 — 34.99	47	132	13.5	12.7
35.00 — 37.49	74	165	21.1	15.8
37.50 — 39.99	42	165	12.0	15.8
40.00 — 42.49	36	159	10.3	15.2
42.50 — 44.99	27	150	7.7	14.4
45.00 — 47.49	11	86	3.2	8.2
47.50 — 49.99	1	67	.3	6.4
50.00 — 52.49	3	15	.9	1.4
52.50 — 54.99	..	9	.	.9
55.00 — 57.49	. .	1	...	.1
57.50 — 59.99	.	9	. .	.9
60.00 — 62.49	...	2	.	.2
Total.....	349	1,043	100.0	100.0

SOURCE: Bureau of Labor Statistics, Bulletin 960-1.

the St. Louis sample as in the New Orleans sample. However, when the number in each salary group is expressed as a percentage of the total for each city, it is valid to compare the percentage of New Orleans employees with the percentage of St. Louis employees in the same class.

When data are classified according to given classes for a number of years, one may want to compare the relationship of a certain class to the total of all classes for several years. Table 6-3 shows the factory sales of passenger cars in the United States for the years 1948 and 1955 classified according to body styles. Since the number of passenger cars sold was larger in 1955 than in 1948, it may be that the number of 4-door sedans sold increased only in proportion to the total. To find out whether or not this is true requires an analysis of the relationship of 4-door sedan sales

to the total for the two years. The two columns with the caption "Percentage of total" show the relationship of each body style to the total sales for the year. These columns show that in 1955 the sale of 4-door sedans was less in relation to the total than in 1948. A comparison of the percentages in the last two columns gives a summary of the shift in consumer demand for different body styles.

Table 6-3

FACTORY SALES OF PASSENGER CARS IN THE UNITED STATES: 1948 AND 1955

CLASSIFIED BY BODY STYLE

Body style	Thousands of vehicles		Percentage of total	
	1948	1955	1948	1955
4-door sedans	1,969	3,043	50.4	38.5
2-door sedans	1,618	1,691	41.4	21.4
4-door hard-tops		1,667	..	21.0
2-door hard-tops		502	..	6.3
Convertibles (except hard-tops)	197	240	5.0	3.0
Station wagons	102	759	2.6	9.6
Chassis	23	18	.6	.2
Total	3,909	7,920	100.0	100.0

SOURCE: Automobile Manufacturers Association, *Automobile Facts and Figures*, 1956.

Forcing percentages to total one hundred. When a number of items are expressed as percentages of their total, these percentages obviously should add to 100. But since the percentages are "rounded" whenever they do not come out even, it is seldom that they total exactly 100. The following table shows the percentage distribution of passenger car production among the principal automobile manufacturers in 1955. The last digit has been rounded off to the nearest one tenth of 1 per cent, and the percentages total 99.9. The question of how to show this total is answered differently by statisticians.

Table 6-4

PASSENGER CAR PRODUCTION BY PRINCIPAL COMPANIES IN THE UNITED STATES, 1955

Manufacturer	Number of cars (Thousands)	Percentage of total
General Motors	3,990.0	50.2
Ford Motor Company	2,240.7	28.2
Chrysler Corporation	1,361.8	17.1
Studebaker-Packard	182.1	2.3
American Motors	161.8	2.0
Kaiser Motors	5.8	.1
Total	7,942.2	99.9

SOURCE: *Standard and Poor's Industry Surveys*, June 7, 1956, p. A113.

There are three possible courses: (1) the total may be written 99.9, representing the actual total of the percentages; (2) the total may be written 100.0, even though the column actually totals 99.9; or (3) the figures may be forced to total 100.0 by increasing one of the percentages to bring the sum to exactly 100.0. The total is just as likely to be larger than 100.0 as smaller, and it sometimes is as much as two over or under in the last digit.

Three of the various methods that may be used to force the total are:

1. Increase or decrease the largest percentage by enough to bring the sum of the percentages to 100.
2. Increase or decrease the percentage that would involve the smallest amount of change.
3. Increase or decrease the percentage that would involve the smallest relative change.

If the first method is used, the percentage for General Motors is increased to 50.3. If the second method is used, it is necessary to know the second and the third decimals which were used in rounding the percentages. The following list shows the percentages carried to three decimals, before they were rounded to one decimal place.

50.238
28.213
17.146
2.293
2.037
.073

The largest amount that was dropped occurred when 17.146 was rounded to 17.1. If 17.146 were raised to 17.2, there would be less error than if any other amount were raised. However, the relative error is greater than when 50.238 is raised to 50.3. The relative error introduced in rounding is more significant than the amount of error. Changing the percentage that involves the smallest relative error can be defended more easily than any other change.

When the largest percentage is changed, regardless of the amount of change that is involved, it frequently follows that the relative error is the least; although, this is not always true. In the above example, raising 50.238 to 50.3 results in the smallest relative error, and so the first and third methods of forcing the total are the same. The outstanding advantage of forcing the total by changing the largest percentage is the ease with which it can be done. Raising the percentage that gives the smallest amount of error, or the one that gives the smallest relative error, involves additional calculation. Changing the largest percentage is the most commonly used method and the one that most statistical departments consider standard practice.

The question of whether the total should be forced is perhaps more important than the question of method. The argument usually advanced in favor of making the percentages total 100 is that the person using the figures is likely to conclude that they are incorrect if their total is not exactly 100. This fact has led most statistical organizations to adopt the rule that distributive percentages must be made to total 100. However, when complex cross tabulations are presented in a table, it is difficult to adjust all of the percentages. In such a situation it is wise to adopt the practice of rounding each percentage to the nearest value and ignore the fact that some of the totals are not exactly 100.

Percentage change. In comparing numbers relating to different periods of time, one may wish to place emphasis upon the percentage change rather than the absolute amount of change.

When comparing changes in prices for several items from one year to another, the absolute amount of change for each item is not usually so significant as the relative change in price. A relative change shows the relationship between the absolute change in price and the price with which a comparison is being made.

According to Table 6-5, there is an absolute change of 4.5 cents in the price of grapes and an absolute change of 12.1 cents in the price of coffee. But note that the price of grapes is less than one fourth the price of coffee. Consequently, little significance can be attached to a comparison of the

absolute price changes. When these price changes are expressed as percentage changes, however, the difficulty of dealing with different prices is overcome. Only when the prices are about the same can the amount of change for two items be compared, and it is obvious that prices will usually not be even approximately the same.

Table 6-5

RETAIL PRICES OF SELECTED FOODS

NOVEMBER, 1954, AND NOVEMBER, 1955, AND PERCENTAGE CHANGE

Commodity	Unit	Nov. 1954 cents	Nov. 1955 cents	Change in price (Cents)	Percentage change
Round steak....	lb.	92.4	88.1	− 4.3	− 4.7
Frozen peas....	10 oz.	19.3	21.4	+ 2.1	+10.9
Grapes.....	lb.	23.1	18.6	− 4.5	−19.5
Coffee.....	lb.	105.1	93.0	−12.1	−11.5
Eggs.....	doz.	66.4	58.1	− 8.3	−12.5

SOURCE: *Monthly Labor Review*, Vol. 79, No. 1 (January, 1956), p. 131.

The normal comparison between two periods of time is to compare the later with the earlier, or to compare the amount of change with the earlier period as the base. When the amount of change is compared with the more recent period, absurd results may be obtained, unless extreme care is maintained in expressing them. For example, assume that accidents were reduced from 400 in 1956 to 100 in 1957, or a reduction of 300. If 300 is expressed as a percentage of the more recent figure, the following percentage is obtained:

$$\text{Percentage} = \frac{-300}{100} \times 100 = -3.00 \times 100 = -300 \text{ per cent}$$

If it is then stated that accidents were reduced 300 per cent in 1957, it is implied that the 400 accidents in 1956 were reduced 300 per cent. But 300 per cent of 400 is 1,200, which is obviously impossible. In a statement of the percentage increase or decrease, such as is commonly made, it is always understood that the earlier period is the base.

A correct statement of the above facts would be that 1956 accidents were 300 per cent more than 1957 accidents. This avoids stating that there was a change from one number to the other, for it is commonly considered that time runs from the earlier year to the more recent, and one would not say that accidents changed from 100 in 1957 to 400 in 1956.

Another source of confusion in percentages of change grows out of the use of a different base without realizing a change has been made. For example, assume that the employees of a concern were given a 10 per cent reduction in salary when business conditions were unfavorable. Then, when conditions improved, it was desired to restore the cut. A salary of $250 per month cut 10 per cent was reduced $25 to $225. But a 10 per cent increase later amounted to only an increase of $22.50, making the monthly salary only $247.50. The obvious reason a 10 per cent cut was not restored by a 10 per cent increase was that a smaller base was used in computing the increase than was used in computing the cut.

Another principle to remember is that zero cannot be used as the base of a percentage, since division by zero is not a permissible operation. It was stated on page 110 that a number less than 100 should not be used as the base of a percentage; to use zero as a base is an extreme violation of this principle. It is also impossible to compare plus and minus numbers by means of percentages. When either of these situations exists, the percentage change cannot be computed. Table 6-6 shows a percentage column with two percentages not computed; there is nothing else that can be done.

Table 6-6

NET PROFITS OF FOUR CORPORATIONS, 1956, 1957, AND PERCENTAGE CHANGE

Company	1956 Profits	1957 Profits	Percentage change
A.....	$2,645,820	$3,845,769	45
B.....	1,786,416	1,531,231	— 14
C.....	342,862	(d)415,214	..
D.....	.....	(#)784,326	..

d Loss.
First year of operation.

Relatives

Fixed-base percentage relatives of time series. A time series may be simplified by reducing the data for the various intervals to ratios of the data for one period. This period becomes the base and has a value of some multiple of 10. The ratios are called *relatives*; and if the base is given a value of 100, they are *percentage relatives.* Numbers that are expressed as relatives are also known as *index numbers,* or percentages of a selected base number. The construction of index numbers is treated in detail in

Chapter 16. An example of percentage relatives is given in Table 6-7: the second column shows steel ingot production since 1926; the third

Table 6-7

STEEL INGOT PRODUCTION: 1926 TO 1956

Year	Production (Thousands of tons)	Relatives 1926 = 100
1926	54,072	100
1927	50,317	93
1928	57,721	107
1929	63,198	117
1930	45,581	84
1931	29,057	54
1932	15,322	28
1933	26,020	48
1934	29,181	54
1935	38,183	71
1936	53,499	99
1937	56,636	105
1938	31,752	59
1939	52,798	98
1940	66,982	124
1941	82,837	153
1942	86,030	159
1943	88,873	164
1944	89,642	166
1945	79,702	147
1946	66,603	123
1947	84,894	157
1948	88,640	164
1949	77,978	144
1950	96,836	179
1951	105,200	195
1952	93,168	172
1953	111,610	206
1954	88,312	163
1955	117,036	216
1956	115,258	213

SOURCE: American Iron and Steel Institute, *Annual Statistical Report*, 1955 (for years 1926-1955); *Survey of Current Business*, March, 1957 (for 1956).

column shows the same information expressed as percentages of the 1926 production, which means that the 1926 production is 100 per cent and the other years are percentages either larger or smaller than 100.

Such a series of relatives reflects the changes from year to year in a manner that is easier to understand than the original data because the comparison is with 100 in each case. When the comparison is between two years, neither of which is the base year, the relatives are usually easier to compare than the actual data, although not so easy as one year with the base year.

Sometimes no year may be considered normal; and instead of using a year as the base, an average of several years may be used. The favorite base year during and immediately after World War I was 1913, the last prewar year. For the decade of the twenties and the first part of the thirties, 1926 was a common base year, although the average of the three years 1923–1925 was also used in many cases. By the end of the thirties, many series of relatives had been shifted to the five-year average, 1935–1939.

After World War II, government statistical agencies set out to revise their statistical series; and one of the steps in this revision was shifting the index numbers to a postwar base. The average of the three years 1947–1949 was selected as the base period for government index numbers.

When two series are expressed as relatives with respect to the same period, the two may be compared more easily than if they were related to different base periods. For this reason, it is of some value to use the same base period as well-known published series. Any period, however, may be used if the problems of analyzing the data seem to justify it.

Link relatives and chain relatives. Instead of comparing each period with a fixed period, the comparison wanted may be that of each period with the preceding period. Such relatives are referred to as *link relatives.* Column (3) in Table 6-8 gives the production of steel ingots for each year expressed as a ratio to the preceding year. For example, production in 1943 was to 1942 as 1.033 is to 1.000, computed as follows:

$$\text{Ratio} = \frac{88{,}873{,}000}{86{,}030{,}000} = 1.033$$

In the same manner the production for 1942 was found to be related to 1941 as 1.039 is to 1.000; in other words, the production for each year is successively the numerator and then the denominator of the ratio.

Since the relatives computed in Table 6-7 and the link relatives computed in Column (3) of Table 6-8 are from the same data, there is a fixed relationship between the two. From the link relatives it is possible to compute the fixed-base relatives that were computed directly from the data. Such relatives are referred to as *chain relatives* to distinguish them from the fixed-base relatives computed directly from the original data. The chain relatives are shown in Column (4) of Table 6-8 expressed

as ratios to 1.000. In Column (5) of the same table they are expressed as ratios to 100. A comparison of this column with the fixed-base relatives computed in Table 6-7 will demonstrate the fact that they are the same values, although occasionally errors of rounding will make a slight difference in the amounts.

Table 6-8

COMPUTATION OF CHAIN RELATIVES, 1926 = 100

STEEL INGOT PRODUCTION: 1926 TO 1956

Year (1)	Steel ingot prod. (Thousand tons) (2)	Link relatives (Previous year = 1.000) (3)	Chain relatives	
			1926 = 1.000 (4)	1926 = 100 (5)
1926	54,072	....	1.000	100.0
1927	50,317	.931	.931	93.1
1928	57,721	1.147	1.068	106.8
1929	63,198	1.095	1.169	116.9
1930	45,581	.721	.843	84.3
1931	29,057	.637	.537	53.7
1932	15,322	.527	.283	28.3
1933	26,020	1.698	.481	48.1
1934	29,181	1.121	.539	53.9
1935	38,183	1.308	.705	70.5
1936	53,499	1.401	.988	98.8
1937	56,636	1.059	1.046	104.6
1938	31,752	.561	.587	58.7
1939	52,798	1.663	.976	97.6
1940	66,982	1.269	1.239	123.9
1941	82,837	1.237	1.533	153.3
1942	86,030	1.039	1.593	159.3
1943	88,873	1.033	1.646	164.6
1944	89,642	1.009	1.661	166.1
1945	79,702	.889	1.477	147.7
1946	66,603	.836	1.235	123.5
1947	84,894	1.275	1.575	157.5
1948	88,640	1.044	1.644	164.4
1949	77,978	.880	1.447	144.7
1950	96,836	1.242	1.797	179.7
1951	105,200	1.086	1.952	195.2
1952	93,168	.886	1.729	172.9
1953	111,610	1.198	2.071	207.1
1954	88,312	.791	1.638	163.8
1955	117,036	1.325	2.170	217.0
1956	115,258	.985	2.137	213.7

SOURCE: American Iron and Steel Institute, *Annual Statistical Report*, 1955 (for years 1926-1955); *Survey of Current Business*, March, 1957 (for 1956).

The computations are carried through with ratios to 1.000 instead of percentages. This is in accordance with the statement on page 107 that ratios to be used in further computations should be to the base 1.00 instead of 100 or some other power of ten. In Table 6-8 the chain relative in Column (4) for 1926 is given a value of 1.000 since this year is to be the base. The link relative for 1927 is .931 in relation to 1926, and so this is also the chain relative for 1927. The computation of the chain relative for 1928 is as follows: the 1928 link relative is 1.147, which means the ratio of 1928 to 1927 is 1.147 to 1. But 1927 in relation to 1926 is .931 to 1, so the ratio of 1928 to 1926 is .931 $\times$ 1.147 or 1.068. This is the chain relative for 1928 with 1926 equal to 1.000.

The chain relative for 1929 is computed by multiplying the link relative for 1929 by the chain relative for 1928. Each chain relative is computed by this same formula: namely, the link relative for a given year times the chain relative for the preceding year equals the chain relative for the given year. The chain relatives were carried out one more decimal to reduce the error due to rounding, since each chain relative was used in the computation of the next. The chain relatives contain some rounding error even when carried out one more decimal, and a number of them differ slightly from the correct fixed-base relatives.

That the chain relatives and the fixed-base relatives are the same mathematically can be demonstrated by showing all the steps in the computation of the chain relatives for the first three years.

$$\text{Chain relative 1927} = \frac{1927}{1926} \quad \begin{array}{l}\text{(This is the same as the}\\ \text{link relative for 1927.)}\end{array}$$

$$\text{Chain relative 1928} = \frac{1928}{1927} \times \frac{1927}{1926} = \frac{1928}{1926}$$

$$\text{Chain relative 1929} = \frac{1929}{1928} \times \frac{1928}{1927} \times \frac{1927}{1926} = \frac{1929}{1926}$$

For a series such as steel ingot production (Table 6-8), there is no reason for computing the chain relative when the fixed-base relative can be computed directly from the data. The chain relative does have a very useful application in statistical analysis, however, when only the ratio of one period to the previous period is known. It frequently happens that a comparable series does not exist over a large number of periods. Such a situation may arise when a statistical agency is compiling a series from firms reporting at fixed periods, such as monthly. Assume that retail sales of independent stores are reported by a group of stores that is almost the same every month,

but usually not exactly the same. While it would be desirable to have every store report every month, some reports simply do not come in. And there is also the fact that new stores need to be added to the list, and stores stop reporting for one reason or another. If the sales of the stores reporting were totaled each month, these totals would not be comparable from one month to the next because the stores included would not be exactly the same.

In this situation stores are requested always to report the amount of their sales for the current period and for the preceding period. When the sales for the two periods are added, the two totals are comparable, since they are based on the identical stores for both periods. It is, therefore, possible to compute the ratio of the current period sales to the preceding period. This is a link relative for the current period. If a link relative is secured for each month, the resulting series shows the changes from month to month in retail sales. The chain relatives will show changes in retail sales in the same way the chain relatives in Table 6-8 show the changes in steel ingot production.

The situation described above can occur only when sample data are being compiled. If the concerns reporting in each period constitute a representative sample, the link relatives correctly express the relationship of each month to the preceding month, within the limits of error due to the fact that a sample has been used. The chain relatives derived from the link relatives are therefore as accurate as the sample data on which they are based.

Ratios between unlike quantities. In the ratios discussed previously, the comparisons have been made between two amounts that are stated in the same units. When expressing the classes of a distribution as percentages of the total of all classes, the amounts are in the same units or they could not have been added to secure a total. When comparisons between two periods were made, it was between the same kinds of figures in the two periods.

The ratios discussed in this section express a relationship between two amounts not in the same units. These ratios are extremely useful, and many business ratios are of this type. The per capita ratios are an important example. If the retail sales in Texas are expressed as a ratio to the number of people in the state, it would be stated as a given number of dollars per capita. The computation of this ratio for 1954 sales in relation to the estimate of the population of Texas made by the Bureau of the Census is as follows:

$$\frac{\text{1954 sales}}{\text{population}} = \frac{\$9{,}020{,}658{,}000}{8{,}251{,}000} = \$1{,}093 \text{ per capita}$$

Ratios of this type are expressed in the units occurring in the numerator of the fraction, in this case, retail sales in dollars. This amount is followed by the phrase "per capita"; every ratio of this type contains such a phrase that states the unit in which the denominator is expressed. Whenever the denominator represents population, the phrase is "per capita," which means "per person." If the ratio were retail sales in relation to families, it would be computed by dividing the total retail sales by the number of families.

Since so many business data classified on a geographic basis are related to the number of people in the area, per capita ratios are widely used by business statisticians. In computing per capita ratios it is essential that the data in the numerator of the ratio and the population in the denominator be related to the same geographic areas and the same time period. The 1954 sales in the above ratio are compared with the population of Texas on July 1, 1954. It would probably be better if the sales for 1954 could be compared with the average population in Texas during the year 1954, since the population did not stay the same throughout the year. However, the population for July 1 was the only estimate available, and this estimate would be nearly enough comparable to the sales figure for the year.

A few commonly used ratios of this type are described below. Death rates are computed by finding the ratio between the number of deaths and the population, and are ordinarily expressed as the number of deaths per 1,000 population. Birth rates are expressed in the same way. In 1953 the number of births in the United States was 25.1 per 1,000 population. Specific causes of death are expressed in relation to 100,000. Deaths in 1953 from diseases of the heart were 357.8 per 100,000 population; while in 1920, deaths from the same cause were 159.6 per 100,000 population. Deaths from influenza and pneumonia in 1953 were 33.1 per 100,000, and 207.3 in 1920. Deaths per 100,000 population from motor vehicle accidents were 24.2 and 26.7 in 1953 and 1920 respectively.

The population per square mile is a measure of the density of population in different geographic divisions of the country. According to the 1950 Census of Population there were 147.5 persons per square mile in the New England states and 5.9 per square mile in the Mountain states.

The ratio of crop production to the number of acres shows the yield per acre. For example, in 1954 wheat production was 18.1 bushels per acre and corn production was 37.1 bushels per acre.

Airplane accidents are reported as a ratio to the number of miles flown in the same period. There were 88 fatalities from accidents on domestic airlines in 1945 and 103 fatalities in 1953; however, the ratios were .42 fatalities per 1,000,000 revenue miles flown in 1945 and .19 per 1,000,000 revenue miles flown in 1953. The accident ratios may also be expressed in

the reverse form by stating the number of miles flown per accident. There were 40 accidents in 1945, and the ratio of 5,234,222 revenue miles flown per accident was published as the accident ratio, compared with 14,504,263 revenue miles flown per accident in 1953.

Since a ratio may express the relationship of either number to the other, it is frequently a matter of usage as to which way the ratio is stated. In the example of airplane accidents, both methods are used; in many situations one method is customarily used in preference to the other.

STUDY QUESTIONS

1. Is the classification of statistical data a separate operation from the analysis of data, or should classification be considered one step in analysis? Discuss.

2. Ratios are most commonly related to some power of 10, but there is a considerable variation in the power of 10 used. For example, in Problem 7 on page 126, ratios are based on 10,000 vehicles, 100,000 population, and 100,000,000 vehicle-miles. What basic principle should govern the selection of the power of 10 that is to be used as the base for a ratio?

3. Which value is the base in the following ratios?

(a) Freight revenue per ton mile.
(b) Revenue ton miles per mile of road.
(c) Percentage of total sales spent for advertising.
(d) Death rate per 1,000 population.
(e) Per capita personal income.
(f) Number of physicians per 100,000 population.
(g) Wind velocity in miles per hour.
(h) Labor force participation rates (proportion of each age group in the labor force).
(i) Monthly labor turnover rates per 100 employees.
(j) Severity rates based on the average number of days' disability for each million employee-hours worked.

4. Most business organizations prefer to force distributive percentages to make them total 100. Under what circumstances do you consider this rule undesirable?

5. Probably the most commonly used rule for forcing distributive percentages to total 100 is to change the largest percentage in the distribution. This rule has the advantage of being simple and specific, but do you think that it is a logical method of forcing percentages to total 100?

6. Why is it not possible to compute the percentage change between a profit in one year and a loss the following year?

7. For a business that had no sales in 1956, how is the percentage increase in sales between 1956 and 1957 calculated?

PROBLEMS

1. During a period of declining business a manufacturer of electric mixers cut the price of his mixers 10 per cent. The demand improved the next year and the price of the mixer was increased 10 per cent.

(a) When the price cut was put into effect, what was the new selling price of an electric mixer that had sold for $39.95?

(b) What was the selling price after the increase was put into effect?

(c) Why did the 10 per cent increase in price not restore all of the 10 per cent cut of the previous year?

2. Business failures reported by *Dun's Statistical Review* increased from 10,969 in 1955 to 12,686 in 1956. What was the percentage increase?

3. The United States Patent Office reports that 83,266 patent applications were filed in 1955 and 80,035 in 1956. What was the percentage decrease?

4. A store's gross profit in 1955 was 24 per cent of net sales, but in 1956 gross profit was 30 per cent of net sales. Was the percentage increase in the gross profit 6 per cent or 25 per cent? Explain.

5. The following table gives estimated advertising expenditures by media for 1950 and 1956.

(a) Compute the percentage distribution of expenditures by media for each of the two years.

(b) Write a brief summary of the shifts that have taken place in the distribution of advertising expenditures between 1950 and 1956.

ADVERTISING EXPENDITURES IN THE UNITED STATES, BY MEDIA: 1950 AND 1956

(MILLIONS OF DOLLARS)

Medium	1950	1956
Newspapers	2,076	3,305
Radio	605	565
Television	171	1,255
Magazines	515	782
Business papers	251	480
Direct mail	803	1,400
Outdoor	143	200
Miscellaneous	1,146	1,995
Total	5,710	9,982

SOURCE: *Statistical Abstract of the United States, 1957.*

6. The following table shows the national income and population of the United States for the years 1950 to 1956.

(a) Compute the per capita national income for each year.

(b) Explain briefly what is shown by the per capita figures with respect to the changes in national income during the period.

(c) Compute the fixed-base relatives for national income and for population, using 1950 as the base.

(d) What do the two relatives for 1956 show about the increase in national income and population since 1950? Does this agree with your conclusions in (b)?

NATIONAL INCOME AND POPULATION OF THE UNITED STATES: 1950 TO 1956

Year	National income (Millions of dollars)	Population[1] (Thousands)
1950.....	239,956	151,683
1951.....	277,041	154,360
1952.....	290,177	157,028
1953.....	302,129	159,636
1954.....	298,335	162,417
1955.....	324,048	165,271
1956.....	342,400	168,091

[1]Total population including Armed Forces overseas, as of July 1.

SOURCE: *Statistical Abstract of the United States, 1957.*

7. The following table gives the number of traffic deaths and traffic-death rates for the years 1950 to 1955. The death rates are shown on three bases: deaths per 100,000 population, deaths per 10,000 motor vehicles registered, and deaths per 100 million vehicle-miles.

(a) Write a brief statement telling what each of these series shows concerning the increasing number of traffic deaths.

(b) Which series do you think gives the most accurate measure of the change in the number of accidents?

TRAFFIC DEATHS AND DEATH RATES: 1950 TO 1955

Year	Traffic deaths (Number)	Traffic deaths per 100,000 population	Traffic deaths per 10,000 motor vehicles	Traffic deaths per 100,000,000 vehicle-miles
1950.....	34,763	23.0	7.2	7.6
1951.....	36,996	24.1	7.2	7.5
1952.....	37,794	24.3	7.2	7.4
1953.....	37,955	24.0	6.8	7.0
1954.....	35,586	22.1	6.1	6.3
1955.....	38,300	23.3	6.2	6.4

SOURCE: *Automobile Facts and Figures, 1956.*

8. The net national debt of $225.4 billion at the end of 1956 was much larger than in 1939. However, a simple comparison of the size of the debt at different times may not be the most precise method of determining what has happened. The table below gives information from which the following ratios may be computed as part of an analysis of changes in the public debt.

(a) Compute the percentage increase in the public debt between 1939 and 1956.

(b) Compute the per capita debt for each year. What was the percentage change in the per capita public debt between 1939 and 1956?

(c) Compute the ratio of the public debt to national income for each year. What was the percentage change in this ratio?

(d) Write a brief statement that summarizes the change in the public debt between 1939 and 1956, making use of the above computations.

NET PUBLIC DEBT, POPULATION, AND NATIONAL INCOME: 1939 AND 1956

	1939	1956
Net public debt (billion dollars)[1].....	42.6	225.4
Population (millions)[2].....	131.0	168.1
National income (billion dollars)[3]...	72.8	342.4

[1]As of December 31.
[2]As of June 30.
[3]Total for year.

SOURCE: *Statistical Abstract of the United States, 1957.*

7 CHARTS

Since all types of statistical classifications are capable of being expressed graphically, this phase of statistical analysis is properly considered highly important. Most graphs go beyond merely presenting the data; there is usually some purpose behind the construction of a chart. A chart is inferior to a table of figures as a method of *presenting the data*, since one can get only approximate quantities from the chart; but its outstanding advantage is that it *emphasizes certain facts or relationships more than figures do.* The chart is not a substitute for a table of figures; it is ordinarily a form of analysis of the figures.

SIMPLE BAR CHARTS

A basic form of graphic presentation is the *bar chart*, which can be employed whether the classification is one of kind, differences of degree, geographical, or temporal. The basic principle is to set up a scale and to draw bars of the correct length, according to the scale, to represent the different classes.

Charts 7-1, 7-2, 7-3, and 7-4 illustrate the following important points to consider in the construction of all bar charts:

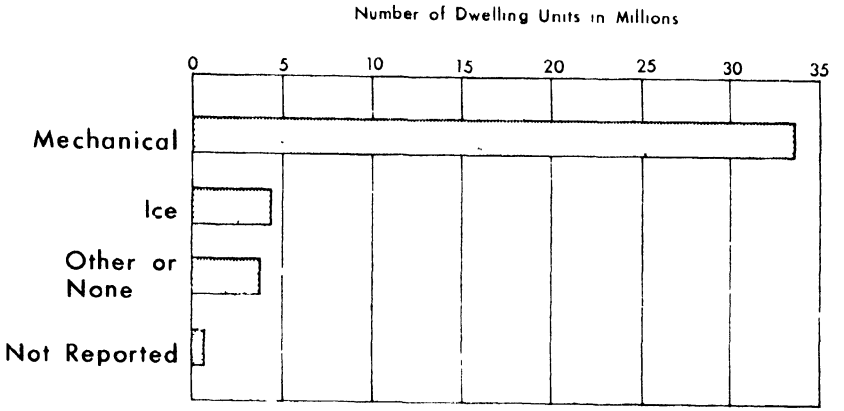
1. *Title.* The title should tell exactly what information the chart contains.
2. *Scale.* A scale must be provided to allow the reader to interpret the significance of the length of the bars. Guide lines should be drawn across the chart at major points of the scale to aid in estimating the lengths of the bars.
3. *Arrangement.* The principles stated in Chapter 2 on arrangement of data in tabular form apply to the arrangement of bars.
4. *Class designations.* The classes represented by the bars should be clearly indicated at the left of the bars for horizontal bar charts, and at the bottom of vertical bars.
5. *Source.* The source of the chart, if it was previously published, or the source of the data from which the chart was constructed, should be indicated.
6. *Mechanical details.* The bars should be constructed of heavier lines than any other lines on the chart. The space between the bars should be less than the width of the bars. The whole chart should be either approximately square or slightly wider than high.

Classifications of Kind

Table 2-1 (page 20) gives the number of occupied dwelling units in the United States classified according to kind of refrigeration reported. The figures in the table give the number of homes in each class; the length of the bars in Chart 7-1 presents this same information graphically.

Chart 7-1

REFRIGERATION IN UNITED STATES DWELLING UNITS, 1950



Source of data Table 2-1

Geographical Divisions

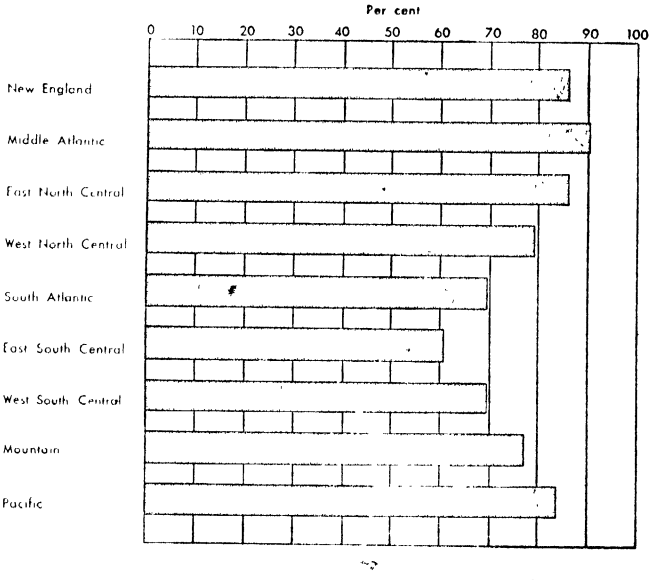
Table 2-4 (page 24) gives the number of dwelling units in the United States and the percentage of those with mechanical refrigeration, by geographical divisions; Chart 7-2 shows this percentage distribution graphically. Bar charts of geographical distributions are usually constructed with horizontal bars.

There is a definite limit to the number of geographical divisions that can be shown advantageously in this manner. An attempt to show all the states in the United States, for example, would require too many bars; this is a very real limitation. But if the number of divisions is small enough, the bar chart is a very effective method of presenting geographical distributions. It is important to notice that this chart shows only the percentage of occupied dwellings with mechanical refrigerators, by regions; it gives no information as to the size of the various geographical divisions of the country. The chart shows that the percentage of dwellings with refrigerators is greater in New England than in the Mountain states; it does not show that New England is much smaller than the other sections of the country.

Chart 7-2

PERCENTAGE OF OCCUPIED DWELLINGS WITH MECHANICAL REFRIGERATORS, 1950

By Geographical Divisions



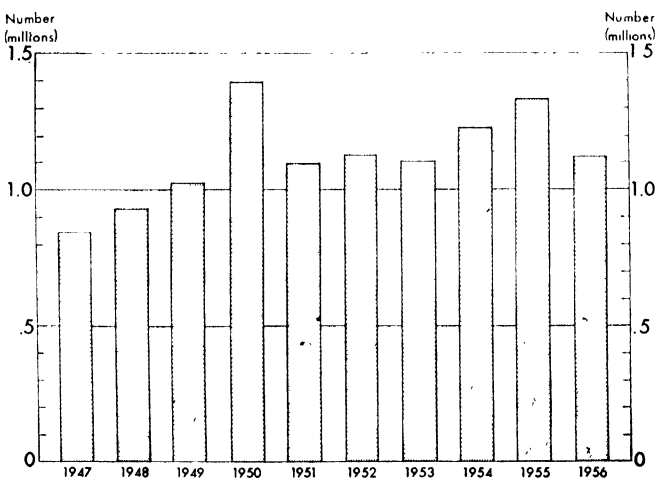
Source of data Table 2 4

Time Series

Table 2-6 (page 25) gives the number of new dwelling units started in nonfarm areas each year from 1947 to 1956; Chart 7-3 shows this information graphically. Vertical bars are commonly used in charts of time series.

Chart 7-3

NEW PERMANENT DWELLING UNITS STARTED IN NONFARM AREAS, 1947-1956



Source of data Table 2-6

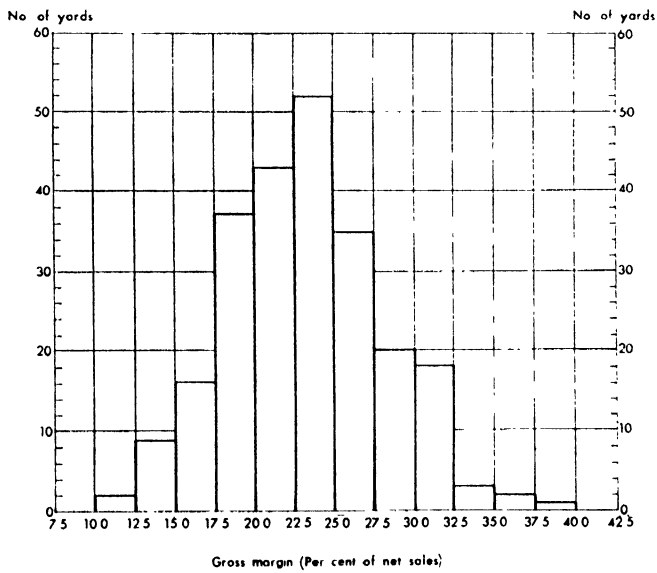
When the number of time units is small, and when it is important to stress the size of the different years individually, the bar chart is very effective for time series. There is the same limitation on the number of periods that can be shown effectively as in the case of trying to show too many divisions in a geographical distribution. It is possible to show a large number of years or months in a bar chart, but in most cases a line chart would be superior. Examples of the line chart for time series will be found on pages 140 to 149, where the subject is discussed completely. But in many circumstances a chart similar to Chart 7-3 is an extremely effective and simple method of showing the series.

Frequency Distribution

Chart 7-4 shows graphically the frequency distribution in Table 2-8 (page 30). Charts of frequency distributions always use vertical bars, which are usually identified by a scale on the X axis that labels the class boundaries. (Refer to page 136 for a discussion of the X axis.) Since a bar is drawn to represent each class, this scale serves to identify the bars. It is particularly important in this type of chart that both the vertical scale and the horizontal scale be carefully labeled.

Chart 7-4

GROSS MARGINS OF 238 LUMBER YARDS, 1956

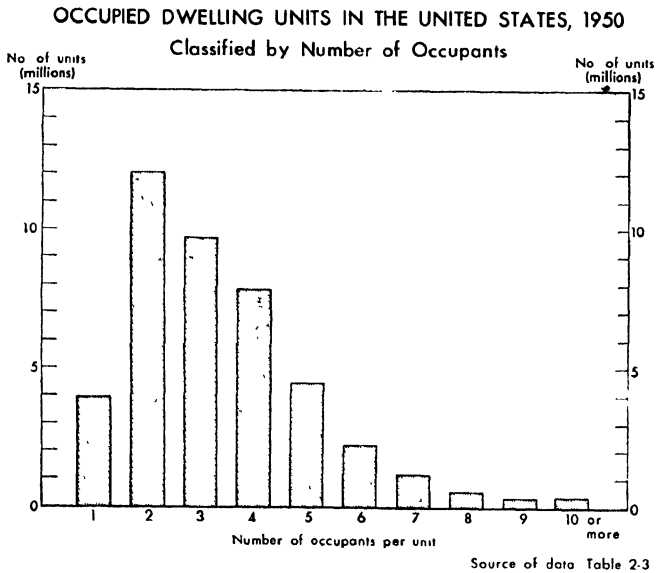


Source of data Table 2-8

Table 2-8 is an example of a continuous series of data and for this reason there is no space between the bars in Chart 7-4. Since the data are a continuous series, the class limits must have no gaps between them; another class begins where the preceding class stops. A bar chart of continuous data is usually called a *histogram.*

Discrete or discontinuous data were described on page 29, and an example was given in Table 2-3 (page 23). Classifications of discrete data may be shown by a bar chart similar to those described in the preceding pages; however, in constructing a bar chart depicting discrete data, space is left between the bars. Chart 7-5 shows a classification of discrete

Chart 7-5



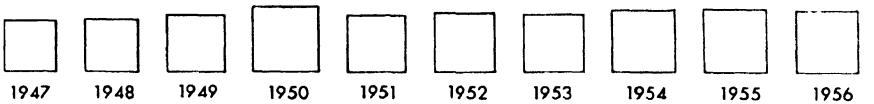
data, which is secured from Table 2-3. It is constructed in the same way as the chart of continuous data, except for the space between the bars and the fact that the scale on the X axis is placed slightly differently in relation to the bars.

TWO- AND THREE-DIMENSION CHARTS

Sometimes magnitudes are compared by making areas or volumes in proportion to the different classes. For example, instead of showing bars of different length for Chart 7-3 it would be possible to show squares of different areas, such as Chart 7-6.

Chart 7-6

NEW PERMANENT DWELLING UNITS STARTED IN NONFARM AREAS
1947-1956



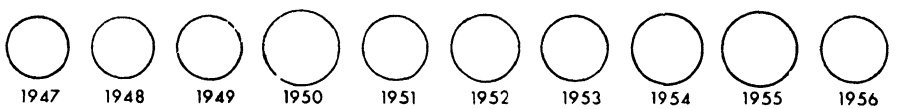
□ = 90,000 dwelling units

Source of data: Table 2-6

Or, instead of squares, circles might be used in the same manner, as depicted in Chart 7-7 below.

Chart 7-7

NEW PERMANENT DWELLING UNITS STARTED IN NONFARM AREAS
1947-1956



○ = 90,000 dwelling units

Source of data. Table 2-6

Since the areas in Charts 7-6 and 7-7 are in direct relation to the magnitudes being compared, either of the charts shown is an accurate picture of the data. But a considerable amount of experimental evidence has rather definitely established the fact that the person looking at Charts 7-6 and 7-7 does not get so accurate an impression of the magnitudes in each class as when looking at a bar chart. When bars of the same width are used, as in the bar charts in this chapter, the reader must make only one comparison, that is, between the lengths of the various bars. But in comparing areas, two dimensions of each figure have been changed and the comparisons are more difficult.

It is even more difficult to compare accurately the *volumes* of several figures than to compare the lengths of different bars. For this reason spheres and cubes are always to be avoided as methods of showing data graphically.

PICTOGRAMS

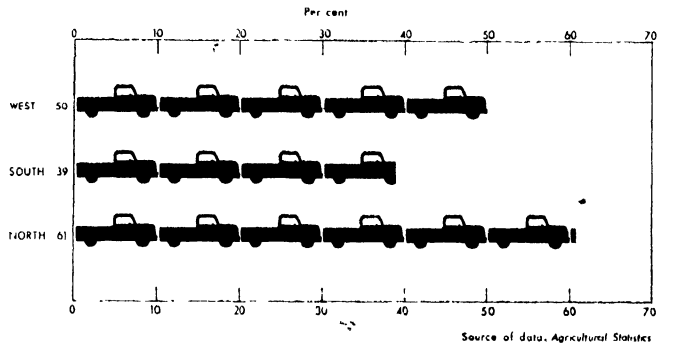
Instead of using bars of different lengths, or figures with different areas, differences in magnitudes may be shown by a *pictogram*, which is a number of picture symbols of the same size — each symbol representing a given

magnitude. This method of presenting statistical data has a particular appeal when an effort is being made to stimulate interest in the information being presented.

Chart 7-8 is a good example of this method. In this case the bars used in the comparison are composed of small pictures, but the chart is essentially a bar chart. The length of each bar represents the magnitude of a class.

Chart 7-8

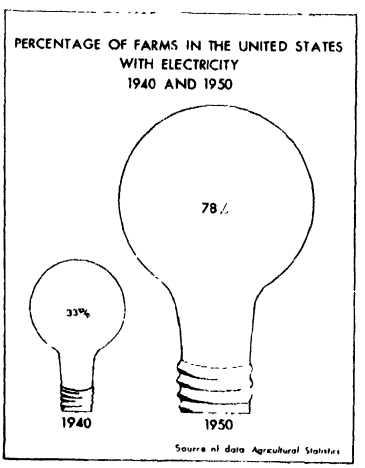
PERCENTAGE OF FARMS IN THE U.S. WITH MOTOR TRUCKS, BY REGIONS, 1954



Source of data, Agricultural Statistics

Any devices that will make charts more attractive are to be welcomed, but it is important that the pictorial charts give as accurate impressions of the data as the less picturesque bar charts and line charts. It is not sound practice to sacrifice accuracy for a form that is more effective in getting attention.

Chart 7-9



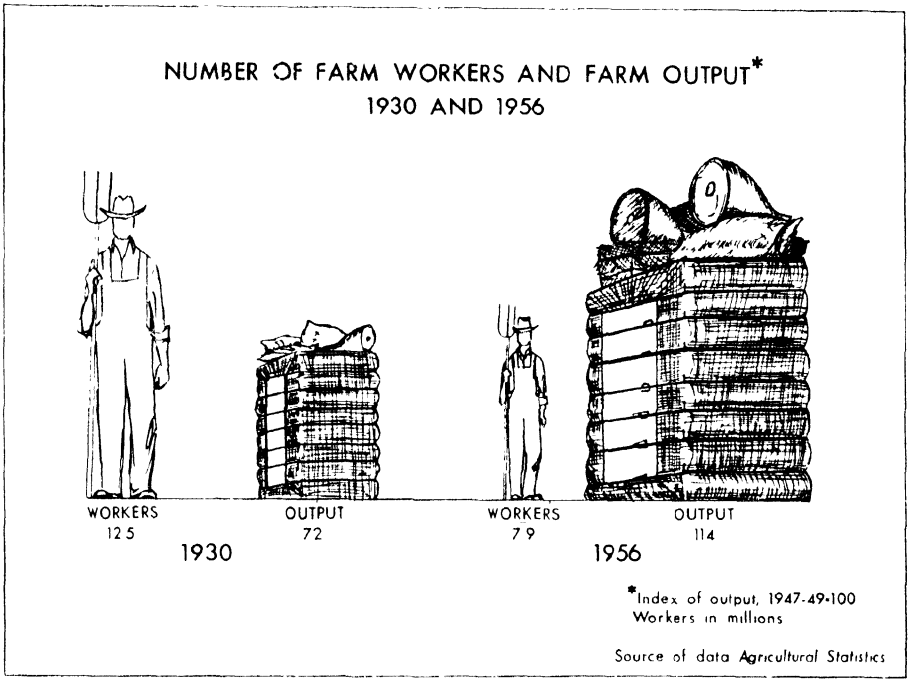
Probably the most common way in which accuracy is sacrificed is in making a two-dimension comparison of the pictures when only a one-dimension comparison was intended. Chart 7-9 is an example of this situation.

The percentage of farms with electricity in 1950 was 2.4 times the percentage of farms with electricity in 1940, so the light globe representing 1950 was drawn with all the dimensions 2.4 times the dimensions used for the 1940 globe. This is a common mistake made in drawing area charts. It is incorrect to assume that making the dimensions of the 1950 globe

2.4 times those of the 1940 globe will make the globe representing 1950 appear to be 2.4 times the size of the one representing 1940. What actually happened is that the areas of the two globes differ as the *square* of the linear dimensions; thus, the area of the 1950 globe is 5.8 times the area of the 1940 globe.

Chart 7-10 is a somewhat more complicated example of an area chart in which the areas are not in proportion to the data they are to represent. The figures representing the number of farm workers in the two years are drawn with their heights in proportion to the number of workers (12.5 and 7.9 millions). This means that each line in the figure representing 1956 is 63 per cent of the corresponding line in the 1930 figure. But the resulting area comparison makes the area of the 1956 figure 40 per cent of the area of the 1956 figure, since the areas vary as the square of the linear dimensions ($.63^2 = .40$). The pile of products representing

Chart 7-10



farm output distorts the correct relationship in the same way. The index of 114 in 1956 is 1.58 times the index of 72 for 1930 (or 158 per cent of 1930), but making the dimensions in this ratio results in an area for 1956 that is 2.5 times the area for 1930. Charts 7-9 and 7-10 were drawn in this way deliberately to illustrate areas not in proportion to the data they are to represent, a mistake that is made with surprising frequency in published charts.

LINE CHARTS

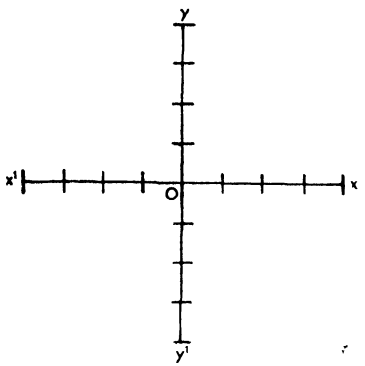


Fig. 7-A **Cartesian Coordinates**

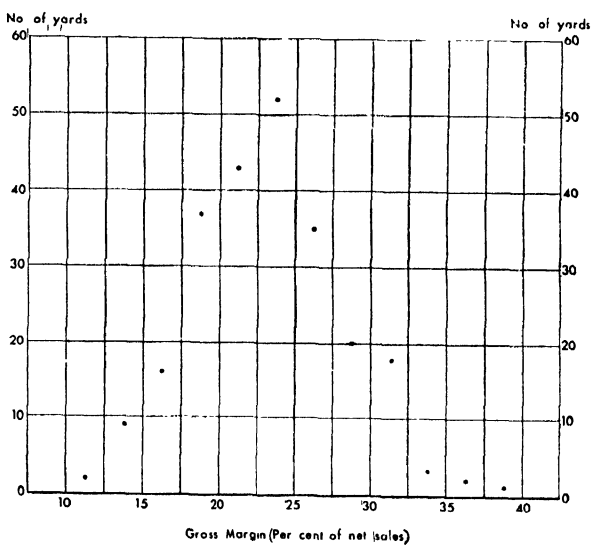
Most statistical classifications can be shown graphically by a bar chart, but some classifications may also be shown on a *line chart*. In constructing a line chart, points are located on a grid in relation to two scales that intersect at right angles; these scales are known as *Cartesian coordinates*. The plotted points are then connected by straight lines. Since the two scales are quantitative, it is obvious that only classifications that are in serial order can be plotted in this manner.

Frequency Polygon

In Chart 7-11 the ratio of gross profit to net sales for 238 lumber yards, shown in Table 2-8 (page 30), is plotted on the Cartesian coordinates. The class intervals are set up as the *X* axis, and the frequency of

Chart 7-11

GROSS MARGINS OF 238 LUMBER YARDS, 1956



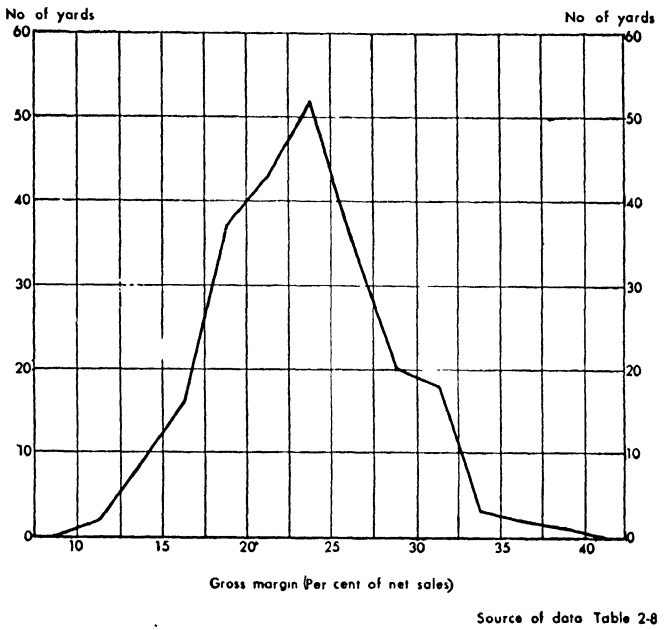
Source of data Table 2-8

each class is then plotted by locating a point in respect to the scale on the *Y* axis and the class intervals on the *X* axis. The points on the *X* axis are located at the mid-points of the class intervals, which would be at the middle of the bars of a histogram drawn on the same grid.

The distances between the plotted points and the base line show the number of frequencies in each class, just as the lengths of the bars show this in the histogram. The plotted points are somewhat inconspicuous, and so for emphasis they are connected by a line. This gives the line chart shown in Chart 7-12. This line chart is known as a *frequency polygon*, and serves

Chart 7-12

GROSS MARGINS OF 238 LUMBER YARDS, 1956



Source of data Table 2-8

essentially the same purpose as the histogram. Comparison with Chart 7-4 will quickly demonstrate that the histogram and the frequency polygon are closely related. It will be noticed that a histogram emphasizes differences between classes within a distribution, while a polygon emphasizes the spread of the distribution.

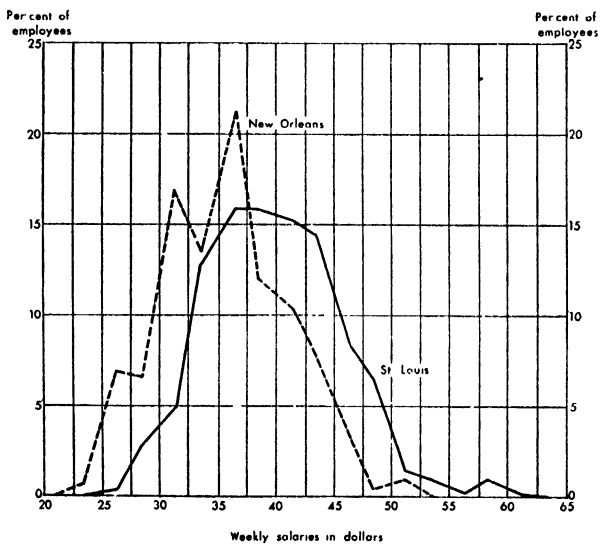
The line connecting the points should always be considerably heavier than the guide lines on the chart. The base line should be somewhat heavier than the guide lines. The significant distance on the chart is the distance between the base line and the line connecting the plotted points, and so these two lines should be the heaviest lines on the chart. The guide lines appearing on the chart should be lighter, since they are of less importance than the other two lines.

The frequency polygon gets its name from the fact that the data line connecting with the base line forms a closed figure, that is, a polygon. The data line always comes down to the base line at the right side and at the left side of the chart, because for every frequency distribution there are finally reached two classes for which there are no frequencies. One more class should be shown beyond the last one in which there is a frequency at both ends of the distribution. One exception to this is when the first class starts with zero. A class below this one would have minus values, and the interpretation of this would be meaningless in most cases. The answer to this is to bring the plotted line down to the base line at zero.

Instead of plotting the *number* of frequencies, it is possible to construct a frequency polygon with the frequencies expressed as *percentages of the total number of items.* This is treated simply as any other frequency distribution in drawing a frequency polygon, except that the Y axis is labeled "per cent of" instead of "number of." When two or more frequency distributions containing a different number of items are to be compared graphically, it can be done effectively by plotting the distributions on the same chart, after expressing the frequencies for each distribution as percentages of the total number in the distribution. Chart 7-13 is based on the percentage distributions in Table 6-2 (page 112), and compares the weekly salaries of women calculating machine operators in New Orleans and St. Louis.

Chart 7-13

WEEKLY SALARIES OF WOMEN CALCULATING MACHINE OPERATORS IN NEW ORLEANS AND ST LOUIS



Source of data Table 6-2

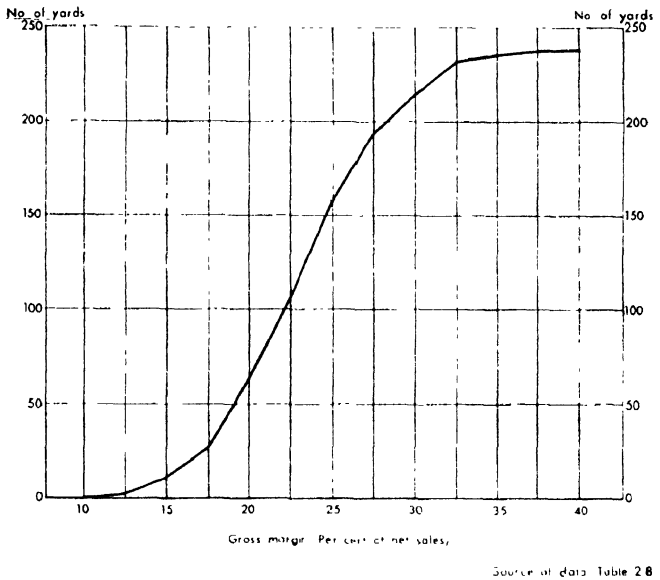
Ogive

The cumulative frequency distribution was described on pages 31 and 32. It is possible to represent the cumulative frequency distribution by a line chart, called an *ogive*. On some occasions this chart may be more useful than the data themselves. Chart 7-14 shows the number of lumber yards having gross margins less than certain indicated percentages of total net sales. It was constructed by plotting on each vertical line the number of

Chart 7-14

GROSS MARGINS OF 238 LUMBER YARDS, 1956

Number of Yards with Less Than Indicated Gross Margin



Source of data: Table 2.8

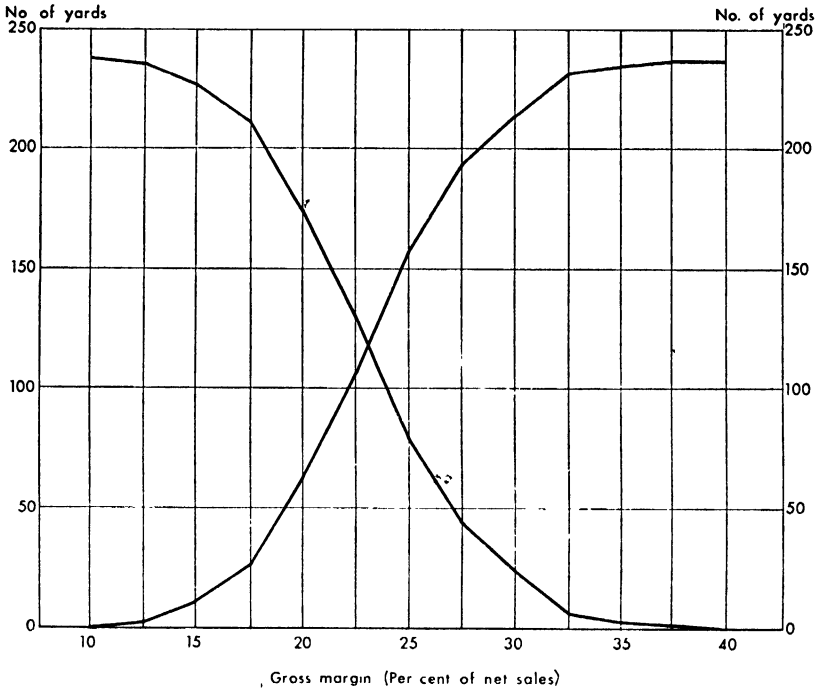
lumber yards having a gross margin less than the percentage represented by the line. These plotted points were then connected by straight lines. Sometimes the points are connected by a curve, but straight lines are used more frequently. This subject will be discussed further in Chapter 8.

Instead of basing the chart on the number of lumber yards having a gross margin less than the indicated percentages, it is possible to construct a chart showing the number of lumber yards having a gross margin more than a given percentage. It is also possible to plot two curves on the same graph, as in Chart 7-15, one curve showing the number of yards having a gross margin less than the indicated percentage of net sales and the other showing the number of yards having a gross margin of a given percentage or more. The two curves cross at the mid-point of the distribution.

Chart 7-15

GROSS MARGINS OF 238 LUMBER YARDS, 1956

Number of Yards with More Than and Less Than Indicated Gross Margin



Source of data Table 2-8

Time Series

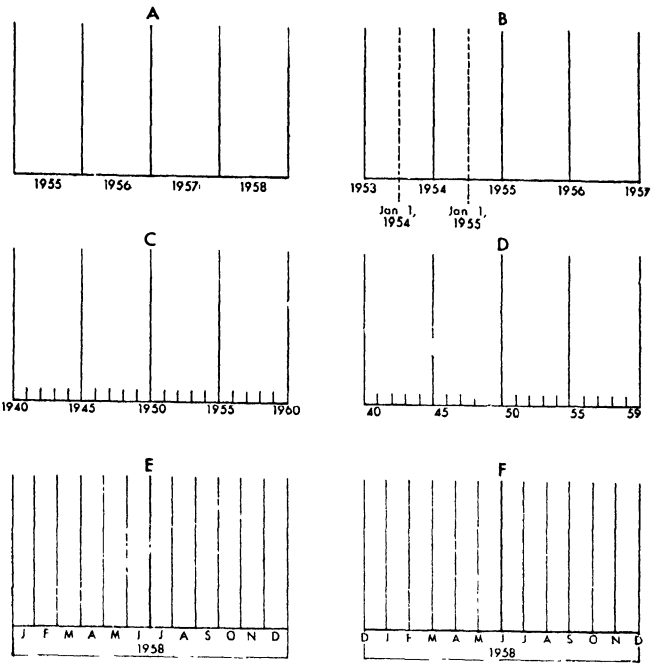
In plotting time series as line charts, the time classification is always laid out on the horizontal axis with the data scale on the vertical axis. The time scale on the horizontal axis is a continuous scale covering the period of the data. The vertical rulings always indicate specific dates, but the number of rulings varies considerably for different charts.

When graph paper is used, each time period may be bounded by guide lines. When the whole chart is constructed by a draftsman, it is common to use fewer guide lines. When there are more plotted points than vertical lines, it is common to use short lines at the bottom of the chart to indicate the frequency of the plotted points.

Several different methods of designating the time scale are given in Chart 7-16. The interpretation of these scales varies somewhat with the data being plotted, but in general the scale designation is written in the

Chart 7-16

EXAMPLES OF TIME SCALES



middle of the time period. In Example A, the middle of each space is July 1 of the year indicated. In Example B, the vertical line above the year is July 1 of that year. Thus, January 1, 1954, is between the first and the second solid vertical lines, and January 1, 1955, is between the second and third. The space between the dotted lines represents the whole year 1954.

In Example C, every fifth year is represented by a guide line, and the intervening years, by short lines. In this example the guide lines and the short lines represent the middle of the years.

Example D shows another time scale with a guide line every five years, and short lines to mark the intervening years. The scale designations are written in the middle of the spaces as in Example A, in this case labeling every fifth space. This example also illustrates the practice of omitting the first two digits and writing the year 1940 as '40 when space is limited.

Example E is designed in the same manner for monthly data as Example A for annual data in that the scale designations are written directly below the spaces rather than below the vertical lines. For monthly data this seems the better method, but some printed graph paper is laid out as in Example F. This design is not completely logical, since assuming that the vertical lines represent the middle of the time interval results in the heavy lines that divide the years falling on the middle of December instead of December 31.

Locating point data and period data on the time scale (X axis). Data classified on the basis of time periods may be either point data or period data. *Point data* are measurements of some characteristic *at a given point of time*, such as prices, inventories, or number of employees. *Period data* are accumulations of amounts *over a period of time*, such as the number of cars sold, tons of coal mined, or amount paid out in wages during a week, month, or year. Table 2-5 (page 24) is an example of point data, and Table 2-6 (page 25) is an example of period data. This distinction is met in accounting, where balance sheet information is point data, since the balance sheet represents the status of the business at a given date. Information from the profit and loss statement is period data, since this statement covers operations over a period of time.

When point data are taken at frequent intervals, it may be desirable to combine these numerous individual values. For example, the price of wheat in Chicago for the month of January might be the price at a given date in the month, or the average of the prices of wheat taken at different dates during the month. If the price of wheat at the close of the day were compiled for each day of January, an average of these prices could be used as the January price of wheat. The individual prices would be point data, and the average of the daily prices would be period data.

The location of point data on the time scale is simple and logical. If inventory data as of January 1 of each year are to be plotted, a point is located at the beginning of the space designated to represent each year and the points are connected by straight lines. If data represented the number of employees on the payroll on the fifteenth of each month, the points would be plotted in the middle of the space representing each month. In every case, point data are located according to this rule.

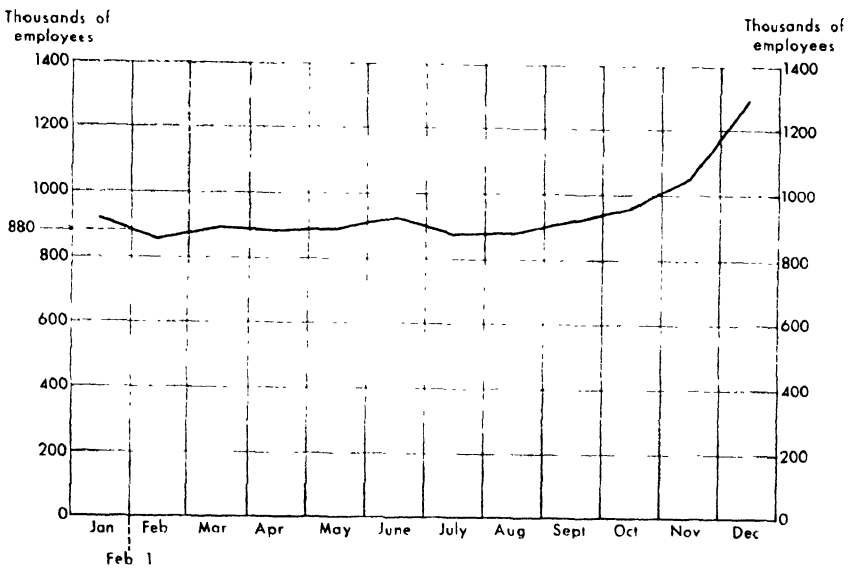
The plotting of period data for a line chart is slightly different, since the line chart must represent the data for a period by the location of one point. It is standard practice to plot the point for a period in the *middle of the space representing that period.* If the vertical guide lines are erected at the middle of the time period as in Example B on page 141, the points are plotted on the vertical guide lines. If the guide lines are at the beginning and end of the period, as in Example A, the data are plotted between the lines. If care is taken to designate the beginning and end of the time periods, no trouble is experienced with this type of chart. But it must always be remembered that period data are plotted in the middle of the space representing the period.

Interpolating on point data and period data charts. Closely related to the problem of plotting data is the use of a graph to estimate intermediate values between the plotted points; this is referred to as *interpolation.*

Values between the plotted points may be estimated from point data charts by erecting a perpendicular at any desired location on the time scale, and reading from the *Y* scale the value at which this line intersects the data line. The following chart shows the number of employees on the payroll on the fifteenth of the month, with a point plotted in the middle of the space representing the month. The lines connecting the plotted points may be taken as an estimate of the manner in which the number of employees has changed. For example, the number of employees on February 1 might be estimated as about 880. This amount is read from the *Y* scale as the distance of the plotted line above the base line at February 1 on the time scale of Chart 7-17.

Chart 7-17

EMPLOYEES IN DEPARTMENT STORES
1956



Source of data *Monthly Labor Review*, March, 1957

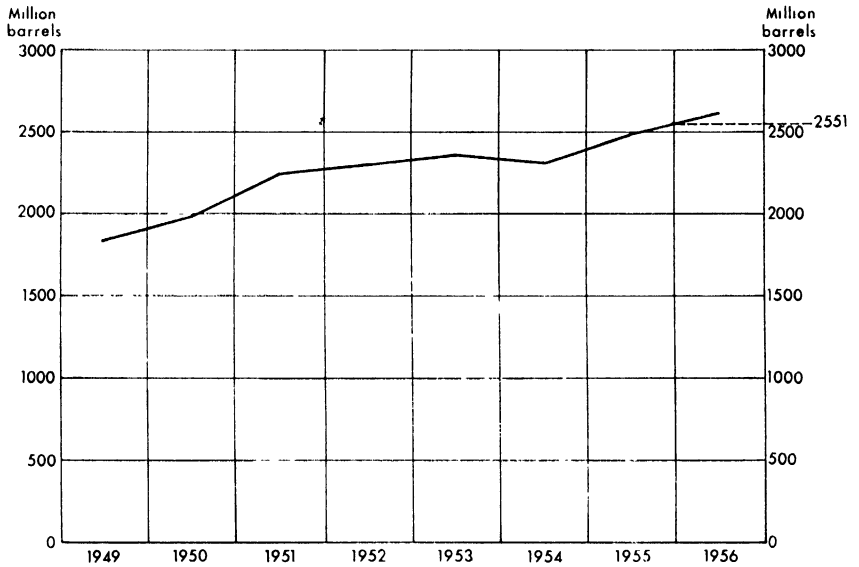
This method of estimating amounts between two points is *straight line interpolation*, and such an estimate is based on the assumption that the number of employees changed at a constant amount between the two dates. If this is not true, the estimate will not be correct; but in the absence of any other information, this method is commonly used. If information is available on how the change occurred, these facts should be used instead of straight line interpolation.

Interpolating on a graph of period data is done in the same manner, although it appears to be somewhat different. In Chart 7-18, each plotted

point represents the amount of petroleum production in a year. The data line at January 1, 1956, has a value of 2,551 million barrels, which is an estimate of annual production. Since the data were plotted in the middle of the time scale for a year, the production of 2,551 million barrels *represents*

Chart 7-18

PETROLEUM PRODUCTION IN THE UNITED STATES
1949-1956



Source of data *Petroleum Facts and Figures*

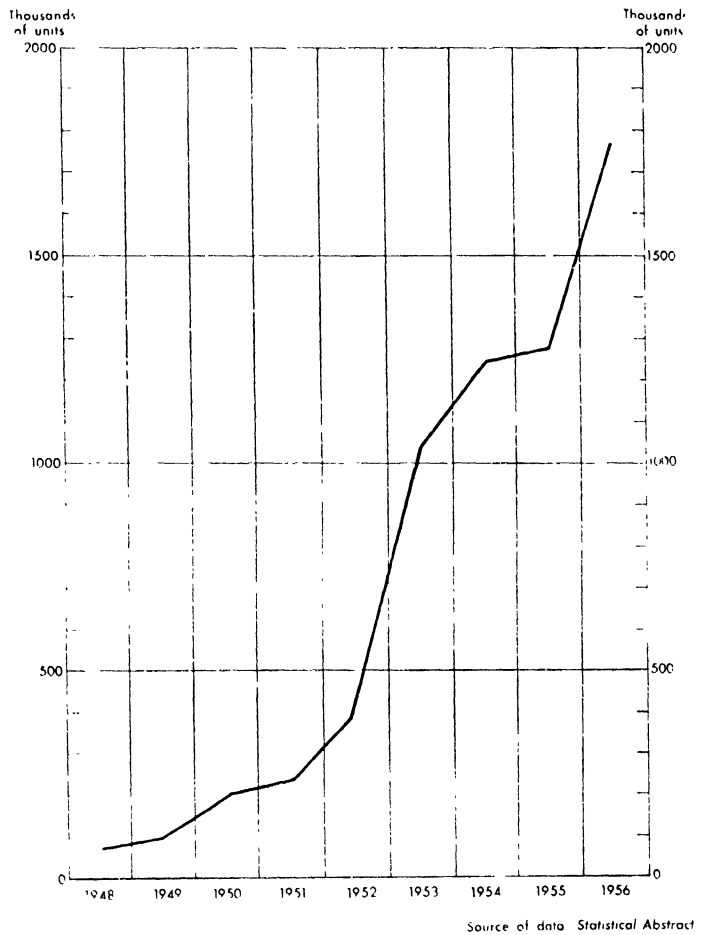
estimated annual production for the period of a year for which January 1, 1956, is the middle. This would be the twelve-month period from July 1, 1955, to June 30, 1956. In the same manner the amount of production for any twelve months of the period covered can be estimated. It is important to remember that any data estimated from this chart will be annual production, and that the point will be located in the middle of the twelve-month period. If the chart were based on average monthly production for each year, the estimates from the chart would be average monthly production for a twelve-month period. All interpolation on charts of period data is carried out according to these principles.

Cumulative time series are point data, since they reflect the total of some value through a given date. Interpolation on a cumulative chart is carried out by erecting a perpendicular on the time scale at the desired point, and reading the amount on the *Y* scale where the perpendicular intersects the data line. This amount will be an estimate of the cumulative amount *to the date* located on the *X* axis.

Y axis. On the vertical, or *Y* axis, the scale, as used in Chart 7-19, is constructed by letting a given space represent a certain number of air conditioners, and the *amount* by which sales increased or decreased can be

Chart 7-19

SALES OF ROOM AIR CONDITIONING UNITS
1948-1956



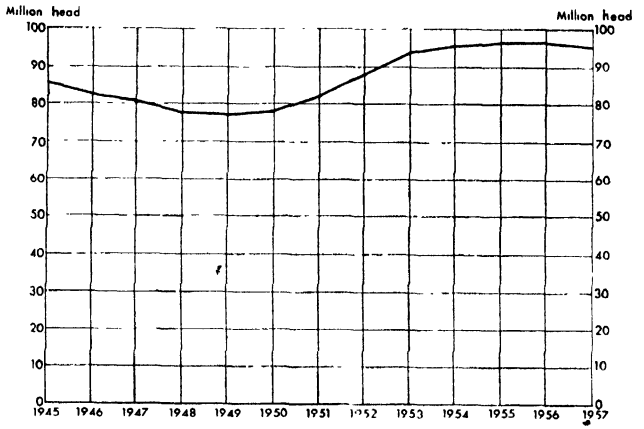
Source of data: Statistical Abstract

determined from the changes in the position of the plotted line.[1] The main purpose of this chart is to show the size of the figures plotted, since the distance from the base line to the plotted points is proportional to the size of the figures. This distance is the most important part of the chart, and

[1]Another common type of chart is the *semilogarithmic* or *ratio chart*, on which the *Y* axis is laid out according to the logarithms of the numbers on the scale. The purpose of the chart is to show the rate of change from one period of time to another. Since the semilogarithmic chart is a specialized type of chart for time series, it will be fully discussed in Chapter 11.

Chart 7-20

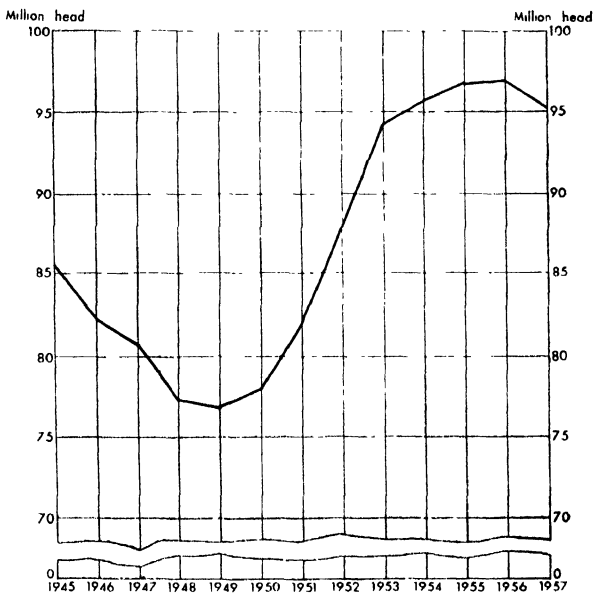
NUMBER OF CATTLE ON FARMS
January 1 of Each Year, 1945-1957



Source of data *The Livestock and Meat Situation* August, 1957

Chart 7-21

NUMBER OF CATTLE ON FARMS
January 1 of Each Year, 1945-1957



Source of data *The Livestock and Meat Situation*, August, 1957

it is essential that the complete scale from the base line be shown. Usually the base line is zero, but occasionally a base other than zero is used.

Sometimes when the base line is zero and the plotted points on the chart are near the top of the grid, there is a temptation to omit part of the scale. Chart 7-20 shows an example of this type of chart, with the plotted line taking up only a small part of the scale. Chart 7-21 shows how the fluctuations can be emphasized by omitting all of the scale except the part in which the points are plotted. Chart 7-20 gives the more accurate impression of the facts. Chart 7-21 gives the impression that the number of cattle fluctuated more than it did. When it is indicated that part of the scale is omitted, the reader is warned; but the fact still remains that the chief function of a chart is to convey a correct impression.

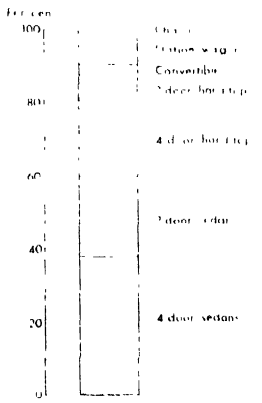
The rule on showing zero is, ordinarily, that either zero should be shown on the scale when it is the base line, or the omission of part of the scale should be plainly indicated. It is perhaps safer to state as a general rule that *all the scale above the base line on a chart should be shown unless there is some special purpose to be accomplished by omitting part of the scale, and then only if its omission is clearly indicated.*

COMPONENT PARTS CHARTS

On page 110 the use of distributive percentages to compare the parts of a distribution with the total was described. This same comparison can be made graphically by means of a component parts chart. There are two satisfactory methods of portraying such a comparison: (1) a *bar chart* divided in the ratio of the different components, and (2) the *circle and sector*, or *pie chart*.

Chart 7-22

PERCENTAGE DISTRIBUTION OF FACTORY SALES OF PASSENGER CARS IN THE UNITED STATES BY BODY STYLES, 1955



Source of data: Table 6-3

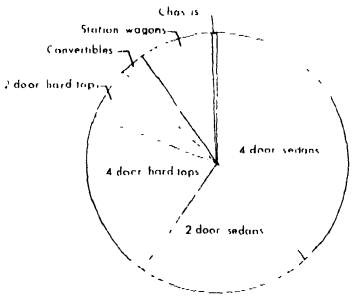
Chart 7-22 illustrates the divided bar chart. The total length of the bar represents the total factory sales of passenger cars in the United States in 1955, and the subdivisions of the bar are based on the percentage distribution in Table 6-3 (page 113).

Although a horizontal bar can be used in the same way as a vertical bar, it is less convenient to label the component parts.

The pie chart, shown in Chart 7-23, is constructed by dividing a circle into segments in proportion to the size of the classes. The circular pie chart can be constructed with a compass, and a protractor can be used for marking off the sections. It is possible to buy a protractor graduated from 0 to 100, but most protractors are constructed to divide a circle into 360 degrees. To use this type, it is necessary to multiply the percentage distribution by 3.6 before plotting the sectors on the circle. It is also possible to buy printed forms with a circle divided into 100 equal parts, ready for plotting the data.

Chart 7-23

PERCENTAGE DISTRIBUTION OF FACTORY SALES OF PASSENGER CARS IN THE UNITED STATES BY BODY STYLES, 1955

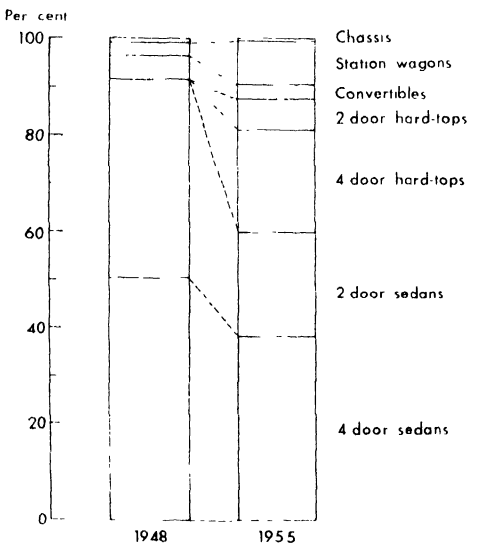


Source of data Table 6 3

It frequently happens that one wants to make comparisons graphically between component parts and the total over a period of time. Chart 7-24 shows the percentage distribution of passenger car sales in the United States by body styles for the years 1948 and 1955. This chart is plotted from the data in Table 6-3 and shows graphically the change in consumer preference of body styles.

Chart 7-24

PERCENTAGE DISTRIBUTION OF FACTORY SALES OF PASSENGER CARS IN THE UNITED STATES BY BODY STYLES, 1948 AND 1955

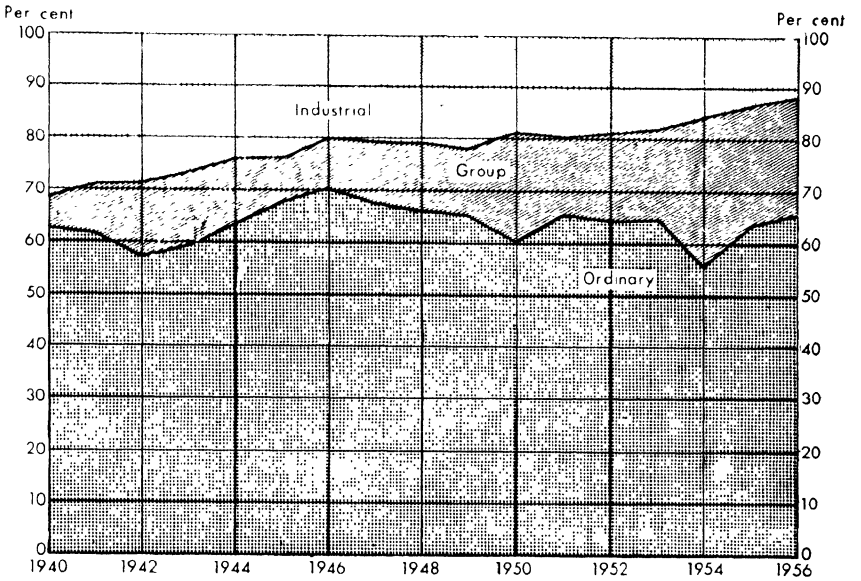


Source of data Table 6 3

Chart 7-25 shows the distribution of life insurance purchases, by kinds of insurance, from United States companies for the years 1940 to 1956. This chart shows graphically the shift that has taken place in the type of insurance bought. The changing width of the sections of the chart shows that purchases of group insurance have been gaining over the purchases of ordinary insurance and even more over those of industrial insurance.

Chart 7-25

PERCENTAGE DISTRIBUTION OF LIFE INSURANCE PURCHASES IN THE UNITED STATES BY TYPE OF INSURANCE, 1940-1956



Source of data *Life Insurance Fact Book*

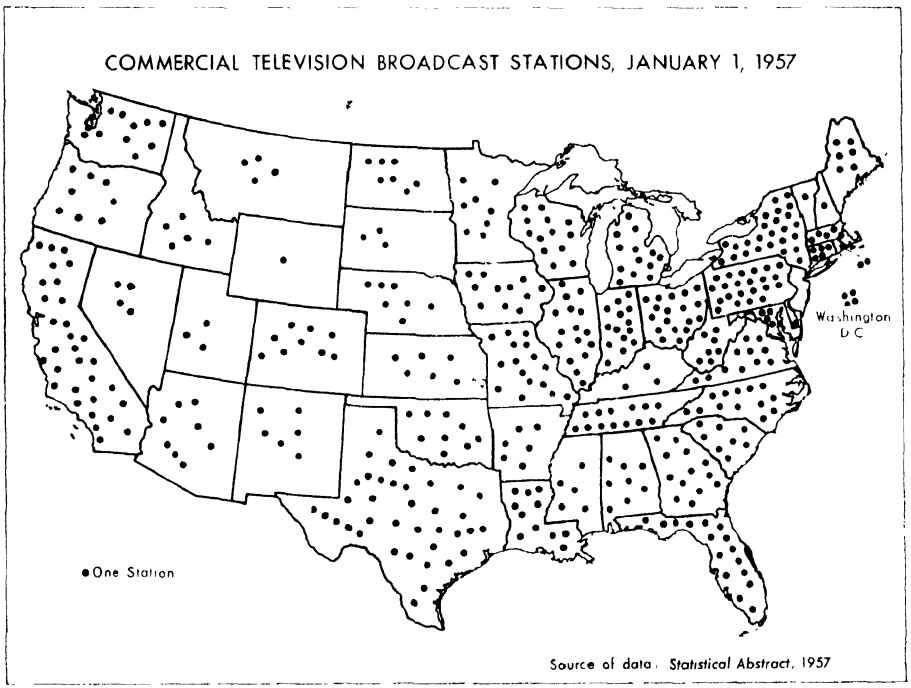
It is possible to use the pie chart to show the percentage distribution for each year; but since no new principles are involved, this chart is not illustrated. It is easier to compare the various segments of the bars than to compare the sectors of the circles; so, the bar chart is usually the type used to compare a number of years.

STATISTICAL MAPS

Chart 7-2 (page 130) illustrates how a geographical distribution may be presented by means of a bar chart. The greatest limitation of this method is that it does not present a very effective chart when the number of

geographical divisions is large. Another method of showing the same information is to use the *large-dot* map. In Chart 7-26 it is possible to count the dots in each state and arrive at a close approximation of the total. This chart gives essentially the same information as a bar chart would give if a bar were drawn for each state. It does have the added advantage, however, of showing the area of each state, as well as the number of television stations in each state.

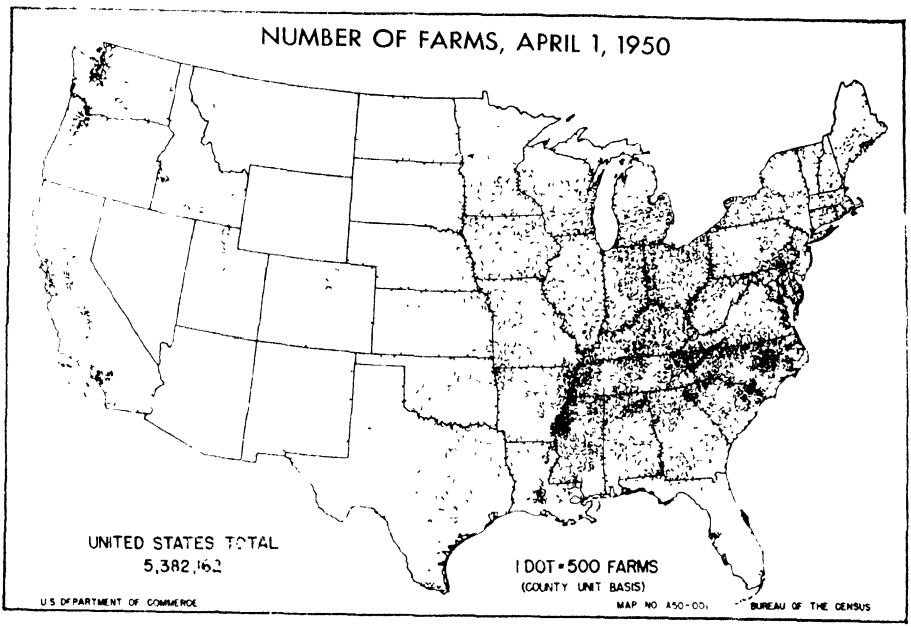
Chart 7-26



There is frequently another important factor to be considered in the analysis of geographical distributions. Different geographical divisions are not the same size; and the analysis wanted may not pertain merely to the number of units of the variable in each state, but perhaps to the number in relation to the size of the state. In other words, the concentration of population, stores, farms, or manufacturing plants in relation to the size of the area, or some other factor, might represent the information desired.

Density is best shown by the *small-dot* map, such as Chart 7-27. It is best to construct small-dot maps by using small areas for the location of the dots, since the individual dots are not located accurately within the area. For example, if a map were being made to show the concentration of farms in the state of Iowa, one dot might be used to represent 50 farms; but the location of a dot would not indicate the location of any individual farms.

Chart 7-27



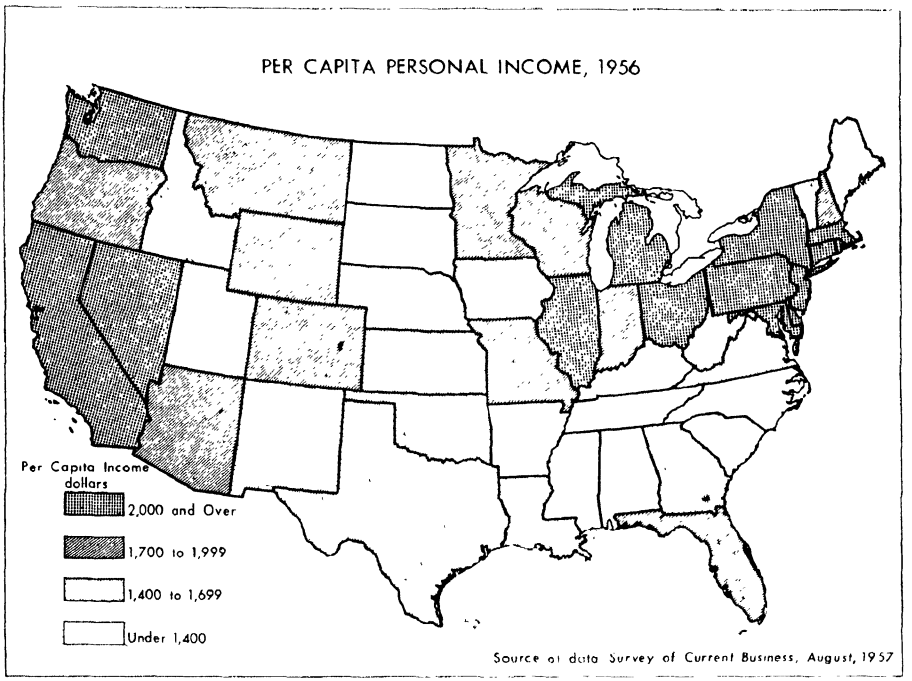
SOURCE: Bureau of the Census, *U.S. Census of Agriculture: 1950*, Vol. V, *Special Reports* Part 6, Agriculture 1950—A Graphic Summary.

If the map were divided into counties, the number of dots for a given county should be spread evenly throughout the county. It is practically impossible to get any impression from this chart as to the number of values of the variable, either in total or in the different areas. But it is easy to make comparisons concerning the density of the different areas. It is also possible to make comparisons of density on different dot maps if the same scales have been used.

The bar chart and the large-dot map, showing the amount of the values of the variable in each geographical area, and the small-dot map, showing the density for each area, are important in different situations. Such a situation exists when it is desired to show a ratio for various geographic divisions. The small-dot map shows a ratio, but it can be used to show only the one ratio, the variable compared with the size of the area. If any other ratio is to be shown, another type of chart must be used.

Chart 7-28 is a crosshatched map that shows the ratio of personal income in each state to the population of the state. It was pointed out in Chapter 6 that ratios related to population figures are expressed as per capita figures; thus, this chart depicts the amount of personal income per capita in each state. The crosshatched map is constructed by states

Chart 7-28



according to the ratio (in this chart the ratio is the amount of personal income per person), and crosshatching the states to show the grouping. The states in which the ratio is highest are crosshatched the heaviest, while progressively lighter crosshatching is used for the smaller ratios.

Different colors instead of light and dark crosshatching could be used. An advantage of light and dark crosshatching, however, is that the heavier patterns indicate the higher ratios; while different colors do not carry this suggestion. When only a few copies of a chart are wanted, it is easier to show the different groups by means of color rather than to crosshatch them; but when charts are to be printed, the crosshatching can be made by mechanical means more cheaply than colors can be printed. It is possible to buy various patterns of crosshatching, which can be cut out and mounted on the map. Then, this copy is usually photographed because an original copy is not very satisfactory to use in a report.

While it is not particularly pertinent at this point, it should be mentioned that color may be used in a chart whenever it makes the presentation of the data more effective. In preparing a few copies by hand, color is often used to improve the appearance. When charts are printed, the use of color becomes expensive; but it is being used rather frequently in spite of its increased cost in comparison with black-and-white reproduction.

MECHANICAL FEATURES OF CHART CONSTRUCTION

The drawing of finished charts is a problem for a trained draftsman. The work is slow and expensive but is worth while when it is important that the charts present an attractive appearance, as in printed work.

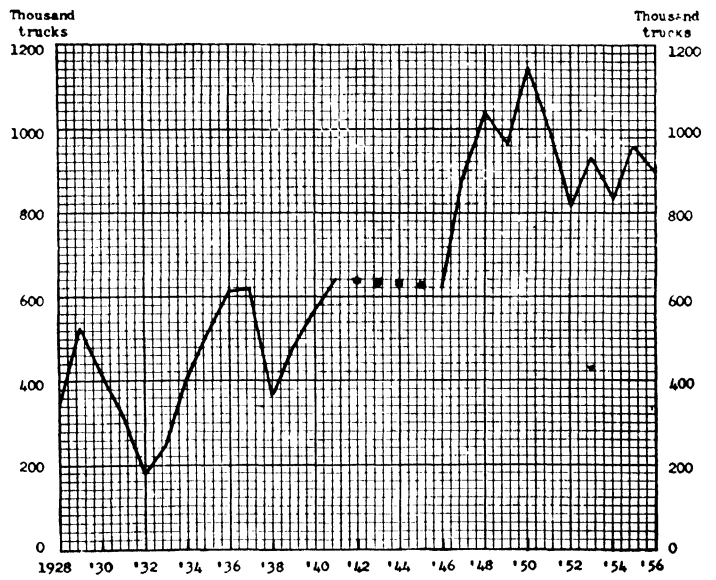
Many charts, however, can just as well be prepared on ruled graph paper, of which many varieties are sold. It normally is printed with many guide lines, so that only the data lines need to be drawn. While these guide lines aid in the construction of a chart, they may detract from the appearance of the finished product.

The construction of statistical charts has become important enough for some schools that teach drafting to give courses in statistical drafting. A statistical draftsman may be well versed in the principles of graphic analysis of statistical data, or he may be skilled only in drawing and lettering a chart that has been designed by someone familiar with such analysis. A common practice is for a statistician to sketch the necessary charts on graph paper and then have the final charts drawn by the draftsman in finished form.

Chart 7-29 is an example of a chart drawn on graph paper, with the lettering done by typewriter. From this chart a draftsman prepared Chart 11-1 on page 247. Most of the charts prepared by a business concern for its own use and most of the charts prepared by students in business statistics courses resemble Chart 7-29.

Chart 7-29

NEW TRUCK REGISTRATIONS IN THE UNITED STATES, 1928-1956



*All motor vehicles rationed

Source of data: Table 11-1

STUDY QUESTIONS

1. What are the advantages of presenting statistical data graphically? What are the advantages of the tabular form of presentation?

2. For a given frequency distribution, how would you decide whether to draw a bar chart (histogram) or a line chart (frequency polygon)?

3. How would you decide whether to draw a bar chart or a line chart of a time series?

4. If you wanted to show the number of fatal highway accidents in comparison with the vehicle-miles in each state, how would you show the relationship on a statistical map?

5. How would you show on a statistical map the concentration of the circulation of a farm magazine in certain areas of the country?

6. Why is it more difficult for a person to compare two areas than to compare the length of two bars?

7. Give five examples of point data that might be used by a businessman in making decisions. Give five examples of period data relating to the same general subjects represented by the point data.

8. When point data for a number of dates have been averaged, do you consider the average to be point data or period data? (An example of this situation would be an average of the number of employees on the payroll of a business on the fifteenth of each month for a particular year.)

9. A qualitative classification based on differences of kind may be presented graphically as a bar chart, representing each class by a bar drawn so that its length represents the size of the class. This same classification may also be presented as a pie chart or a divided bar chart, making each division of the circle or bar proportionate to the size of the class it represents. Explain the difference in the purpose of the two types of charts.

PROBLEMS

1. The following table shows the acres of commercial forest land by regions in the United States. Show this geographical distribution by means of a bar chart.

COMMERCIAL FOREST LAND BY REGIONS, 1953
(AS OF JANUARY 1)

Region	Acres (Thousands)
The North	174,041
The South	193,288
The West	117,011
Total	484,340

SOURCE: Department of Agriculture, Forest Service, *Timber Resource Review* (preliminary review draft subject to revision), 1955.

2. The following table shows the number of homes in the United States owning certain appliances in January, 1957. Present this data in a bar chart.

NUMBER OF HOMES IN THE UNITED STATES OWNING SELECTED APPLIANCES
JANUARY, 1957

Appliance	Number of homes (Thousands)
Air conditioners, room	3,600
Bed coverings, electric	6,261
Cleaners, vacuum	31,600
Coffee makers, automatic	18,425
Food mixers	20,748
Radios	47,700
Refrigerators	45,500
Television sets	38,400
Toasters	36,538
Washing machines	41,175

SOURCE: *Statistical Abstract of the United States, 1957.*

3. From the following data construct a bar chart showing the consumption of wool in manufactures from 1946 to 1956.

WOOL CONSUMED IN MANUFACTURES IN THE UNITED STATES: 1946 TO 1956

Year	Wool consumed (Million lbs.)
1946	737.5
1947	698.3
1948	693.1
1949	500.4
1950	634.8
1951	484.1
1952	466.4
1953	494.0
1954	380.8
1955	413.8
1956	439.4

SOURCE: *Statistical Abstract of the United States, 1957.*

4. Refer to the table in Problem 4, page 68, Chapter 3. If the solution to Problem 4 has not been retained, it will be first necessary to complete step (a) below.

(a) Set up a tally sheet and classify the employees according to age.
(b) Construct a histogram of the distribution.
(c) Construct a frequency polygon of the distribution.

5. Refer to the table in Problem 2, page 104, Chapter 5. If the solution to Problem 2 has not been retained, it will be first necessary to complete steps (a), (b), and (c) below.

(a) Arrange the ratios in a simple frequency distribution.
(b) Derive a cumulative frequency distribution from the data on a "less than" basis.
(c) Derive a cumulative frequency distribution from the same data on a "more than" basis.
(d) Draw two ogives on the same chart, one on a "less than" basis and one on a "more than" basis.

6. From the data given below construct a line chart showing the amount of life insurance in force in the United States at the end of each year from 1946 to 1956.

(a) Is the line chart an example of point data or period data?
(b) Explain the significance of the values represented by the portion of the line between the plotted points.

LIFE INSURANCE IN FORCE IN THE UNITED STATES: 1946 TO 1956

(As of December 31)

Year	Millions of dollars
1946	170,066
1947	186,035
1948	201,208
1949	213,672
1950	234,168
1951	253,140
1952	276,591
1953	304,259
1954	333,719
1955	372,332
1956	412,630

Source: *Life Insurance Fact Book, 1957.*

7. From the data given below construct a line chart showing the amount of life insurance purchases in the United States by years from 1946 to 1956.

(a) Is the line chart an example of point data or period data?
(b) Explain the significance of the values represented by the portion of the line between the plotted points.

LIFE INSURANCE PURCHASES IN THE UNITED STATES: 1946 TO 1956

Year	Millions of dollars
1946	21,702
1947	22,425
1948	22,449
1949	22,550
1950	28,796
1951	27,467
1952	31,539
1953	36,238
1954	45,446
1955	48,427
1956	55,313

SOURCE: *Life Insurance Fact Book, 1957.*

8. The following data show the distribution of expenditures for research and development in the United States in 1956. Construct a pie chart from these data.

DISTRIBUTION OF RESEARCH AND DEVELOPMENT EXPENDITURES AMONG SPENDING AGENCIES, 1956

Agency using funds	Expenditures	
	Amount (Millions of dollars)	Percentage distribution
Government	1,110	17
Industry	4,920	76
University	470	7
Total	6,500	100

SOURCE: *Statistical Abstract of the United States, 1957.*

9. The following table shows the distribution of domestic intercity freight traffic classified by type of transportation for the years 1939 and 1955. Construct a component bar chart with two bars, showing the shift in the type of transportation between 1939 and 1955.

VOLUME OF DOMESTIC INTERCITY FREIGHT TRAFFIC BY TYPE OF TRANSPORTATION: 1939 AND 1955

Type of transportation	Traffic volume			
	Millions of ton-miles		Percentage distribution	
	1939	1955	1939	1955
Railways	370,150	631,385	64.39	49.41
Motor vehicles	52,821	226,188	9.19	17.70
Inland waterways	96,249	216,508	16.74	16.94
Oil pipelines	55,602	203,244	9.67	15.91
Airways	12	481	.002	.04
Total	574,834	1,277,806	100.00	100.00

SOURCE: *Statistical Abstract of the United States, 1957.*

10. The following table gives the number of life insurance companies in each state. Construct a large-dot map showing the number of life insurance companies in the United States, by states.

NUMBER OF LIFE INSURANCE COMPANIES IN THE UNITED STATES
(JUNE 30, 1956)

State	No.	State	No.	State	No.
Alabama	48	Maine	2	Oklahoma	22
Arizona	32	Maryland	14	Oregon	3
Arkansas	21	Massachusetts	10	Pennsylvania	26
California	20	Michigan	11	Rhode Island	2
Colorado	17	Minnesota	14	South Carolina	39
Connecticut	7	Mississippi	16	South Dakota	7
Delaware	12	Missouri	17	Tennessee	17
Dist. of Columbia	14	Montana	1	Texas	348
Florida	25	Nebraska	19	Utah	8
Georgia	32	Nevada	1	Vermont	1
Idaho	2	New Hampshire	2	Virginia	19
Illinois	36	New Jersey	6	Washington	13
Indiana	38	New Mexico	2	West Virginia	2
Iowa	15	New York	27	Wisconsin	9
Kansas	12	North Carolina	21	Wyoming	1
Kentucky	12	North Dakota	6		
Louisiana	102	Ohio	13	Total U.S.	1,144

SOURCE: *Life Insurance Fact Book, 1957.*

11. Construct a crosshatched map showing the percentage increase in personal income per capita for each state between 1940 and 1955, using the data in the following table.

PERCENTAGE INCREASE IN PER CAPITA PERSONAL INCOME BY STATES AND REGIONS: 1940 TO 1955

State and region	Percentage increase	State and region	Percentage increase
Continental United States	**210**	**Southeast**	**276**
New England	**176**	Virginia	229
Maine	205	West Virginia	216
New Hampshire	199	Kentucky	287
Vermont	203	Tennessee	271
Massachusetts	167	North Carolina	277
Rhode Island	163	South Carolina	261
Connecticut	173	Georgia	292
Mideast	**172**	Florida	222
New York	160	Alabama	319
New Jersey	181	Mississippi	334
Pennsylvania	194	Louisiana	267
Delaware	150	Arkansas	315
Maryland	180	**Southwest**	**278**
District of Columbia	99	Oklahoma	304
Great Lakes	**212**	Texas	274
Michigan	214	New Mexico	281
Ohio	210	Arizona	217
Indiana	242	**Rocky Mountain**	**218**
Illinois	199	Montana	224
Wisconsin	220	Idaho	215
Plains	**241**	Wyoming	188
Minnesota	221	Colorado	223
Iowa	215	Utah	219
Missouri	244	**Far West**	**179**
North Dakota	292	Washington	200
South Dakota	247	Oregon	194
Nebraska	251	Nevada	178
Kansas	287	California	170

SOURCE: *Personal Income by States since 1929.*

8 AVERAGES

The fact has already been emphasized that in dealing with masses of individual observations some method of summarizing and describing them is necessary. Classification represents a basic device to bring order to a mass of data. Graphic presentation is another device used to make data comprehensible, and can nearly always be employed profitably. When interest in the data centers on differences of kind or of geographical location, it is frequently sufficient to classify them and present them graphically. Further analysis ordinarily is not necessary.

In classifications involving a time element, more analysis is usually necessary to show all the significant elements. Since changing conditions are so important to business executives, the analysis of time series is particularly important and is discussed in considerable detail in later chapters.

USE OF AVERAGES

In summarizing data, *averages* represent one of the most important devices. By the use of averages, complex groups of observations are summarized by one figure. Averages have been put to more than one use; the most common from the standpoint of statistics will be discussed briefly.

A Common Denominator

One of the simplest uses of averages is found in preparing data to be used in time series. The *Survey of Current Business* reports the production of shoes during each month of 1955, and the average number produced each month during 1955. The total production for the twelve months of 1955 was 581,970 pairs, and the average production each month was 48,498 pairs (total production of 581,970 divided by 12). This average monthly production in 1955 can be compared with the production of the individual months, or it can be compared with the average monthly production of 43,666 pairs in 1954. In other words, monthly production and annual production are put on a comparable basis. This means that two annual figures (average monthly production for each of two years) can be compared with each other, or that either annual figure can be compared with any monthly figure. The advantages of this method of presenting statistical data make it a widely used device for presenting time series.

Table 8-1 shows the production of wheat in the United States by five-year averages from 1901 to 1950 and annually from 1951 to 1956. Using the five-year average rather than the five-year total production reduces the size of the figures and also makes it easier to get an accurate impression of the size of the wheat crop, since the annual production figure is the one by which most people would judge the size of the crop. No one would be likely to develop the habit of thinking in terms of the total production of five years. An even more important reason for using annual average production is to make comparison easy with the last five years. Since it is assumed that the major interest in the wheat crop would be in the most recent years, wheat production after 1950 is given on an annual basis. But since the earlier production was given in average yearly production, each of the recent years could be compared with any earlier five-year period. Thus, one use of averages is to reduce data to a common denominator for comparative purposes.

Table 8-1

PRODUCTION OF WHEAT IN THE UNITED STATES: 1901 TO 1956

Yearly average or year	Production (Thou. bu.)
1901–1905	674,843
1906–1910	664,299
1911–1915	801,080
1916–1920	790,773
1921–1925	787,082
1926–1930	866,470
1931–1935	680,868
1936–1940	795,913
1941–1945	984,580
1946–1950	1,184,749
1951	980,810
1952	1,298,957
1953	1,169,484
1954	969,781
1955	934,731
1956	997,207

SOURCE: "The Wheat Situation," April 26, 1957, Agricultural Marketing Service, U. S. Department of Agriculture.

Use of an average as a common denominator has wide applications in many fields, and many ratios and percentages are calculated for this purpose. The batting averages computed for baseball players are well-known examples. The batting average is merely the ratio between the number of hits and the number of times at bat. To compare the number of hits made by players during the season would result in little significant information, since each one of the players might have been at bat a different number of times.

Data on consumption of many commodities in different years are expressed in relation to the number of people in the country in the different years. The purpose here is merely to reduce the data to a more nearly comparable basis, and per capita figures make allowance for the differences in population. Such figures are also useful to make allowance for differences in population between different geographical divisions. To compare the total wealth of different states is less significant for many purposes than to compare the per capita wealth, which makes allowance for differences in population.

A Measure of Typical Size

In addition to being used as a common denominator, an average can be used as a measure of typical size. In summarizing and analyzing numerical data, it may be desirable to find one value of the variable that may be considered typical of the whole mass of values. Many times the word average is used in the sense of typical, as when one says that a certain employee's ability is "about average."

In fact, most people do not distinguish clearly the sense in which they are using the word average. When a sales manager says the average monthly sales for 1957 were $450,000, he probably means that the total sales were $5,400,000 and the average monthly sales were one twelfth of this amount. He may not have in mind that $450,000 is typical of a month's sales, for there may be a decided variation between sales in different months of the year due to seasonal factors in the business. Perhaps a more precise way of making the statement would be to say that the sales for 1957 were $450,000 per month. At least this would emphasize the fact that he had in mind merely the total for the year divided by twelve.

If a sales manager were asked the average age of his salesmen, he might reply on the basis of the age of the majority of his men, rather than in terms of the total of all their ages divided by the number of men. In answering this question he would probably be thinking of the age that is typical of his group of salesmen, and might describe it as the *most usual age.*

In using averages as typical of a whole group of individual items, there are very definite precautions that must be taken. In general, these consist of being sure the average is actually typical of the group it is taken to represent. Many figures such as per capita data are not representative averages, and should not be used as such.

A popular magazine appealing primarily to homeowners published a table giving the cost of operating water heaters of various types. For example, it was stated that 2,500 cubic feet of gas was the amount of fuel used in a month by the average family with an automatic type heater. The cost of operating this heater for a month was given as $2 when natural gas was used, and $2.50 when manufactured gas was used.

The reaction of the readers of the magazine was not at all desirable from the standpoint of the editor. Some wrote that the amount of gas consumed was given as far too much, and others wrote that the amount was not nearly enough. In general, a very definite impression was given that the results were not accepted by the readers.

The editor was quite worried about the whole matter because he believed the figures were correct. The averages had been computed correctly, and he believed the data collected had represented an adequate and representative sample. The difficulty was that there was so much diversity among the individual figures averaged that it was practically impossible to find any one value typical of the whole group. To state it technically, there was a great deal of dispersion in the data averaged. This resulted basically from the lack of homogeneity of the data; too many different kinds of families, and too many different gas rates were represented. If the data had been broken up into groups and averages computed for the different groups, each average would have been more representative of its group than the one average was of all the families included.

When using an average as typical of all the values of a variable, it is highly important that the data be homogeneous for the purpose at hand. The use of homogeneous data will usually result in the dispersion of the individual items being small enough so that an average that is really typical can be computed. The dispersion of the data is an important characteristic of the data in itself, and is discussed in detail in Chapter 9.

DEFINITION AND COMPUTATION OF THE PRINCIPAL AVERAGES

The three most common averages in general use are the *arithmetic mean*, the *median*, and the *mode*. The arithmetic mean is the best known, and the one most widely used in statistical work. The *geometric mean* and *harmonic mean*, although not in such general use, have important special uses.

Arithmetic Mean

The best known of the averages is the arithmetic mean, usually referred to by the public as the *average*. It is also called the *common average* and the *mean*. When one of these terms is used without qualification, the arithmetic mean is ordinarily meant. While it would be more precise always to use the full name *arithmetic mean*, this average is so commonly referred to by the shorter terms that there is little likelihood of confusion.

The arithmetic mean of a series of values of a variable X_1, X_2, X_3, . . . X_N is the sum of the values divided by the number of values. Expressed as an equation this operation may be written

$$\bar{X} = \frac{\Sigma X}{N} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

In this case, $\bar{X}$ (which may be read "bar X," "X bar," or "the mean of X") is the symbol that will be used for the mean; Σ is the Greek letter *Sigma* which means "the sum of" or "the summation of"; X represents the value of an individual variable; and N is the number of values of the variable X. The above formula would be read as follows: The mean is equal to the sum of the values divided by the number of values.

The total of the variables is available in many cases; thus, one does not have to perform any addition. In fact, sometimes the total is available but the individual items are not listed. Data collected by the Department of Agriculture estimated the cash receipts of farmers from crops, livestock, and livestock products in 1950 to be $28,328,000,000. The 1950 Census of Agriculture reported the number of farms in the United States to be 5,382,162. The mean value of cash receipts per farm from farm products can be computed as follows:

$$\bar{X} = \frac{\$28,328,000,000}{5,382,162} = \$5,263$$

The information in Table 8-2 was taken from a study of the buying habits of farmers in the Middle West. One of the questions asked was the distance regularly traveled to buy gasoline and oil. The answers given by 260 farmers are shown in Table 8-2. The arithmetic mean of the distances traveled by these 260 farmers is computed by totaling the 260 individual values of the variable and dividing by 260. This operation gives the following value for the arithmetic mean:

$$\bar{X} = \frac{1,724.5}{260} = 6.63 \text{ miles}$$

Table 8-2

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

(MILES, ROUNDED TO THE NEAREST TENTH)

3.5	5.5	1.0	6.3	3.2	2.0	11.4	3.2	5.0	2.0
6.0	4.5	15.3	3.9	5.8	2.7	5.1	3.0	11.9	13.8
10.0	.8	1.0	1.3	13.6	4.0	9.6	3.2	2.0	10.5
2.0	5.8	5.6	12.8	4.5	9.5	7.5	11.6	5.0	3.0
11.1	6.5	1.5	5.2	3.7	2.8	10.0	3.2	3.3	17.8
4.1	12.1	5.7	3.0	7.0	1.0	4.0	11.3	4.5	5.0
18.7	4.0	8.0	6.7	10.4	10.8	8.3	9.4	6.0	9.8
.5	3.0	4.5	2.0	4.5	4.5	9.0	1.3	6.5	6.0
3.0	9.7	17.4	8.6	16.0	2.3	1.5	16.6	9.0	2.8
2.0	4.0	5.0	4.0	1.0	12.7	4.5	4.5	4.8	11.0
9.4	4.5	2.2	3.1	12.0	4.8	3.0	5.7	15.0	8.4
7.5	10.4	9.3	10.8	4.0	5.4	7.3	5.2	1.5	.8
3.8	4.0	2.0	5.0	5.8	5.6	1.3	14.7	2.5	6.3
4.5	2.3	12.0	3.5	4.0	17.0	7.0	5.5	13.5	13.2
3.0	6.5	5.5	4.8	6.0	5.0	14.6	6.5	4.7	2.4
2.5	8.8	1.9	14.8	14.5	4.8	8.0	1.8	7.0	4.7
15.1	6.5	5.5	4.0	7.0	3.5	4.0	13.1	2.8	1.0
7.6	5.8	10.1	6.5	5.9	6.5	2.8	3.3	10.5	4.5
7.0	15.8	4.5	3.5	5.5	14.1	13.6	7.7	2.0	11.5
9.8	4.0	6.5	8.0	3.5	4.5	6.5	6.0	7.2	6.3
8.3	5.0	4.9	3.1	10.5	7.5	4.0	3.8	16.2	6.0
12.9	16.5	6.7	16.8	4.5	3.0	15.7	2.5	7.0	11.0
5.7	9.5	2.8	4.2	14.1	5.0	.8	6.2	2.3	5.2
6.0	5.5	7.2	8.7	1.0	3.8	6.5	5.7	3.8	15.9
4.0	12.5	5.0	8.5	8.4	13.0	11.8	4.0	6.6	11.9
3.2	8.2	1.0	10.0	4.8	2.0	7.0	5.1	9.0	3.0

SOURCE: Schedules on file in the Bureau of Business Research, The University of Texas.

Long method of computing the mean. Table 8-3 is a frequency distribution tabulated from the individual mileages given in Table 8-2. In this case both the individual values of the variable and the frequency distribution are available, but in many situations only a frequency distribution is published. In such situations, when the individual items are not listed, it is important to have a method of finding the mean. The method of making an approximation of the mean from a frequency distribution is accurate enough so that it is generally used even if all the values of the variable are known, since there is a considerable saving of time in making the computations from data grouped in a frequency distribution.

The long method of computing the mean assumes that the mid-value of each class is the average value of all the items falling in that class. For

example, in Table 8-3 it is assumed that the 19 farmers who traveled less than two miles traveled an average of .95 miles. The product of 19 and .95 (18.05) is an estimate of the total number of miles traveled by the 19 farmers. Each of the values in the *fm* column represents an estimate of the number of miles traveled by the farmers in that class, and the total of the *fm* column is an estimate of the total number of miles traveled by the 260 farmers. Using the following formula, this total, 1,749 miles, divided by 260 gives a mean of 6.73 miles:

$$\bar{X} = \frac{\Sigma fm}{N} \qquad (2)$$

$$\bar{X} = \frac{1,749}{260} = 6.73 \text{ miles}$$

The difference between the value of the mean secured by dividing the actual total of miles traveled by the 260 farmers and the one obtained by formula (2) is not great. With a rather large number of frequencies and with a class interval that is moderately small, the long method of computing the mean can be used with confidence. The grouping errors in the individual classes are compensating if a distribution does not depart too much from symmetry.

Table 8-3

COMPUTATION OF THE MEAN (LONG METHOD)

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

Miles traveled (Class interval)	Miles traveled (Mid-values) *m*	Number of farmers *f*	*fm*
0.0– 1.9	.95	19	18.05
2.0– 3.9	2.95	52	153.40
4.0– 5.9	4.95	70	346.50
6.0– 7.9	6.95	39	271.05
8.0– 9.9	8.95	24	214.80
10.0–11.9	10.95	21	229.95
12.0–13.9	12.95	14	181.30
14.0–15.9	14.95	12	179.40
16.0–17.9	16.95	8	135.60
18.0–19.9	18.95	1	18.95
Total.....	...	260	1,749.00

SOURCE: Table 8-2.

$$\bar{X} = \frac{\Sigma fm}{N}$$

$$\bar{X} = \frac{1,749}{260} = 6.73 \text{ miles}$$

Short method of computing the mean. The procedure described above is ordinarily abbreviated in practice by computing the difference (d) between the mean and the mid-value of some arbitrarily chosen class interval, instead of the absolute value of the mean. In employing this method, the following symbols are used:

A, the arbitrary origin (mid-value of some arbitrarily chosen class interval)
m, the mid-value of any class

If $$d = m - A,$$

then $$\Sigma(fd) = \Sigma(fm) - \Sigma(fA)$$

or, since A is a constant, $$\Sigma(fd) = \Sigma(fm) - NA$$

and $$\frac{\Sigma(fd)}{N} = \frac{\Sigma(fm)}{N} - A$$

But $$\bar{X} = \frac{\Sigma(fm)}{N}$$

Therefore $$\bar{X} = A + \frac{\Sigma(fd)}{N} \quad \ldots\ldots\ldots\ldots \quad (3)$$

If the correction, or $$c = \frac{\Sigma(fd)}{N}, \quad \ldots\ldots\ldots\ldots \quad (4)$$

$$\bar{X} = A + c \quad \ldots\ldots\ldots\ldots \quad (5)$$

The computation of the arithmetic mean may be shortened still further if the difference (d') between the arbitrary origin and the mid-value of any class is expressed in terms of the class interval, rather than in original units. This can be done only if the class interval is uniform throughout the classification.

Using the class interval as the unit,

$d' = \frac{m - A}{i}$, where i is the number of units in the class interval.

Then $$d = id'$$

Substituting id' for d in formula (4) gives $c = \frac{\Sigma(fid')}{N}$

Since i is constant, $$c = \frac{\Sigma(fd')}{N} i \quad \ldots\ldots\ldots\ldots \quad (6)$$

Using the value for c computed by formula (6), the mean may be found by substituting in formula (5). The computation of the mean by the use of formulas (5) and (6) is illustrated in Table 8-4.

Table 8-4

COMPUTATION OF THE MEAN (SHORT METHOD)

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

Miles traveled (Class interval)	Number of farmers f	d'	fd'
0.0– 1.9	19	− 4	− 76
2.0– 3.9	52	− 3	− 156
4.0– 5.9	70	− 2	− 140
6.0– 7.9	39	− 1	− 39
8.0– 9.9	24	0	0
10.0–11.9	21	1	21
12.0–13.9	14	2	28
14.0–15.9	12	3	36
16.0–17.9	8	4	32
18.0–19.9	1	5	5
Total.....	260	...	− 289

SOURCE: Table 8-2.

$$\bar{X} = A + c$$

$$A = 8.95 \text{ miles} \qquad i = 2 \text{ miles}$$

$$c = \frac{\Sigma(fd')}{N} i = \frac{-289}{260} 2 = -2.22 \text{ miles}$$

$$\bar{X} = 8.95 + (-2.22) = 6.73 \text{ miles}$$

It makes no difference in the result which mid-value is taken as the arbitrary origin. If a different arbitrary origin is taken, all the figures in the fd' column will be changed, but the value ultimately obtained must be the same for the mean. An absolute check on the arithmetic may be effected by taking a different arbitrary origin for the deviations. To keep the values of d' as small as possible, the arbitrary origin should be chosen near the middle of the distribution.

It should be noticed that this method is feasible only when the class intervals are all the same size; with unequal classes, formula (2) may always be used, and probably is the simplest method. It should be mentioned here also that an open-end class renders a close approximation of the mean impossible, or at least uncertain, by either method described above, unless the total of the items in the open-end class is known.

Weighted arithmetic mean. In the formula used for the computation of the mean, each item was included in the total, and was included only once. A method of increasing the influence of a particular item on the average is to include this item more than once in the total. Each item might be given a different influence by including each one a different number of times when finding the sum of the items. It is then said that the items are given different weights, and the result is called a *weighted arithmetic mean.* The calculation of a weighted arithmetic mean is performed by multiplying each item to be averaged by the weight assigned it, totaling these products, and dividing this sum by the sum of all the weights used. The above operations may be summarized by the following formula:

$$\bar{X} = \frac{\Sigma(wX)}{\Sigma(w)}, \text{ where } w = \text{the weight assigned to each item, or } X \text{ value} \ldots . (7)$$

Students are familiar with the use of weighted averages to combine several grades when they are not all of the same importance. Assume that the grades representing a semester of work consist of a final examination, four one-hour examinations, and ten ten-minute tests, a total of fifteen grades. Obviously these grades should be given a different weight in computing the final average. Possibly a ten-minute test might be given a weight of 2, an hour examination a weight of 10, and the final examination a weight of 40. The procedure for compiling the average is to multiply each grade by its weight, add these products, and then divide the sum of these products by the sum of the weights. This calculation is illustrated in Table 8-5 on the next page.

The unweighted average in reality is an average giving the same weight to each item. If each item were given a weight of 1, the sum of the weights would simply be the number of items, or N. The formula for the unweighted average becomes the formula already given for the arithmetic mean, $\bar{X} = \frac{\Sigma X}{N}$.

The calculation of the arithmetic mean from a frequency distribution, illustrated in Table 8-3, is a weighted arithmetic mean of the values of the mid-points of the classes. The mid-point of a class is weighted by the number of items in the class. The sum of the products divided by the sum of the frequencies gives a weighted average of the mid-points, the weights being the frequencies.

Averaging ratios. In computing an average of ratios it is particularly important that each ratio be given the proper weight. It is not uncommon to compute ratios for a number of different classes, and then compute a

Table 8-5

COMPUTATION OF THE WEIGHTED ARITHMETIC MEAN

SEMESTER GRADES OF ONE STUDENT

Examination	Grade X	Weight w	Grade times weight wX
10-minute tests:			
No. 1	81	2	162
No. 2	75	2	150
No. 3	68	2	136
No. 4	90	2	180
No. 5	40	2	80
No. 6	71	2	142
No. 7	60	2	120
No. 8	58	2	116
No. 9	90	2	180
No. 10	73	2	146
Hour examinations:			
No. 1	75	10	750
No. 2	84	10	840
No. 3	79	10	790
No. 4	87	10	870
Final examination	91	40	3,640
Total	...	100	8,302

SOURCE: Hypothetical data.

$$\overline{X} = \frac{\Sigma(wX)}{\Sigma(w)} = \frac{8{,}302}{100} = 83$$

ratio for the total of all classes. Unless the classes are of equal importance, an unweighted average of the individual ratios will not give the correct result.

Table 8-6 shows the percentage of persons 5 to 29 years old enrolled in school in 1950 in each of the four West South Central states. Each percentage is the ratio of enrolled pupils to the total population between the ages of 5 and 29. To compute the average percentage of persons enrolled in the whole West South Central division, care must be taken in averaging these percentages. The unweighted arithmetic mean of the percentages for these four states would be 51 per cent. If there were the same number of persons in each state, this would be the correct mean. But there were 3,225,365 persons in Texas and 807,197 in Arkansas, and so it is not logical to give both states the same weight in computing the mean.

Table 8-6

PERCENTAGE OF PERSONS 5 TO 29 YEARS OLD ENROLLED IN SCHOOL, 1950

WEST SOUTH CENTRAL STATES

State	Percentage enrolled
Arkansas	53.1
Louisiana	50.6
Oklahoma	53.8
Texas	46.4

SOURCE: Bureau of the Census, *U. S. Census of Population: 1950.*

The weighted average of these percentages is computed in Table 8-7, in which the percentage of the population in each state enrolled in school is weighted by the number of persons in the state in the 5 to 29 age group. The weighted average is 49.2 per cent, and it is the correct value because the weights applied to the individual percentages are logical. It is the same percentage that is found by dividing the total number of pupils in the four states by the total population 5 to 29 years old in these four states.

Table 8-7

COMPUTATION OF THE WEIGHTED ARITHMETIC MEAN

PERCENTAGE OF PERSONS 5 TO 29 YEARS OLD ENROLLED IN SCHOOL, 1950
West South Central States

State	Percentage enrolled	Population 5–29 years old	Col. (2) times col. (3)
(1)	(2)	(3)	(4)
	X	w	wX
Arkansas	53.1	807,197	42,862,160.7
Louisiana	50.6	1,146,273	58,001,413.8
Oklahoma	53.8	912,890	49,113,482.0
Texas	46.4	3,225,365	149,656,936.0
Total	...	6,091,725	299,633,992.5

SOURCE: Bureau of the Census, *U. S. Census of Population: 1950.*

$$\overline{X} = \frac{\Sigma(wX)}{\Sigma(w)} = \frac{299,633,992.5}{6,091,725} = 49.2 \text{ per cent}$$

In this case, another method of arriving at the same result would be to compute the relationship between the number of persons 5 and 29 years old in all four states and the total number of persons enrolled in school. The total population, aged 5 to 29, of the four states was 6,091,725 and the number of persons enrolled in school was 3,001,960, or 49.2 per cent. This is the percentage found by taking a weighted average of the four state percentages.

It may be stated as a general principle that an average of ratios can best be computed by going back to the data from which the ratios were computed, combining these data, and then computing a ratio from the total data. In the above example the data from which the ratios were derived were used in weighting the ratios, and the two computations were in reality identical. Since the computations from the total data are easier to make than those in Table 8-7, and are also more easily understood, the weighting of the individual ratios should be used only when the original total data are not available.

For another example: A brick company has five trucks hauling shale from a pit two miles from the brickyard. The trucks are loaded by a power shovel, driven to the brickyard to dump their load, and returned to the pit to be loaded again. The computation below shows the time required per load for each truck on April 12, and the average of these figures.

Truck	Minutes per Load
No. 1	48.0
No. 2	40.0
No. 3	53.3
No. 4	30.0
No. 5	26.7
Total	198.0

$$\bar{X} = \frac{198.0}{5} = 39.6 \text{ minutes}$$

It is also possible to compute the average in another way. Each of the five trucks operated eight hours on April 12. According to the computation below, the five trucks hauled an average of 13 loads on that date.

Truck	No. of Loads Hauled
No. 1	10
No. 2	12
No. 3	9
No. 4	16
No. 5	18
Total	65

$$\bar{X} = \frac{65}{5} = 13 \text{ loads}$$

Since there are 480 minutes in an eight-hour day, the average length of time per load was 36.9 minutes (480 divided by 13). However, the first computation showed an average length of time required to haul one load to be 39.6 minutes. The difference between the two averages comes about through the difference in weighting. The average number of minutes per load on page 172 are ratios, and it is incorrect to give each ratio the same weight as in the first computation; for, since each of the trucks worked the same length of time, *each one hauled a different number of loads.* Truck No. 3 hauled 9 loads in an eight-hour day, but Truck No. 5 hauled 18 loads. The time per load for each truck should be weighted by the number of loads hauled, as is done in the following computation. This gives the average length of time required to haul one load as 36.9 minutes, which agrees with the average time computed by dividing 480 minutes by an average of 13 loads per day.

Truck	Minutes per Load X	No. of Loads in Eight Hours w	wX
No. 1	48.0	10	480
No. 2	40.0	12	480
No. 3	53.3	9	480
No. 4	30.0	16	480
No. 5	26.7	18	480
Total	....	65	2,400

$$\hat{X} = \frac{\Sigma(wX)}{\Sigma(w)} = \frac{2{,}400}{65} = 36.9 \text{ minutes}$$

If each truck had brought one load and we wanted to know the average time required, it would be correct to average the number of minutes giving each one equal weight, for in this case we would not be averaging ratios.

There are means other than the arithmetic mean that automatically give certain weight to the data being averaged, and are particularly useful when averaging ratios. The two most common of the special means are the *geometric* and the *harmonic.* When these special means give the values of the variable the weight desired, they can be used to average the data without having to weight the arithmetic mean as was done in the illustration just cited. However, the use of these special means is a technical statistical problem of an advanced nature. The harmonic mean will not be discussed further in this book, but a common and useful application of the geometric mean to the averaging of ratios is discussed in the following section.

Geometric Mean

The geometric mean of a series of values such as X_1, X_2, X_3, ... X_N is the Nth root of the product of the values. It may be expressed by the equation

$$G = (X_1 \cdot X_2 \cdot X_3 \ldots X_N)^{\frac{1}{N}} \quad \ldots\ldots \quad (8)$$

The simplest method of computing the geometric mean, at least when N is more than two, is by the use of logarithms.[1] In terms of logarithms, the equation is

$$\log G = \frac{\Sigma(\log X)}{N} \quad \ldots\ldots\ldots \quad (8a)$$

The geometric mean is necessarily zero if any value is zero, and may become imaginary if negative values occur. Except for these cases, the geometric mean is always determinate and is rigidly defined. It is a computed average, and therefore resembles the arithmetic mean, as opposed to the position averages such as the median and mode. The geometric is always less than the arithmetic mean of the same variables, unless all the values are the same.

Table 8-8 illustrates the computation of the geometric mean of the percentages of increase in the population of Detroit, Michigan, over six decades. The population in 1900 was 138.8 per cent of the 1890 population; the 1910 population was 163 per cent of the 1900 population; and so on. The problem is to find the average rate of increase over the six decades. The procedure is to find the logarithms of the six percentages, and then to find the *arithmetic mean of the logarithms*, which is the logarithm of the geometric mean. The average percentage that the population of one census date is of the preceding census date is 144.2. The average percentage increase, therefore, is 44.2.

The arithmetic mean of the six percentages of increase is 48.4 per cent. The geometric mean in this case is not a great deal smaller. The difference between the two means depends upon the dispersion of the distribution.

It is possible to compute the geometric mean of percentages of increase without knowing the percentage increase in the individual decades, just as it is possible to compute the arithmetic mean without knowing the individual items being averaged. Notice that we may abbreviate the above computation. Let P_0, P_1, P_2, P_3, P_4, P_5, and P_6 be the population in 1890, 1900, 1910, 1920, 1930, 1940, and 1950 respectively. The percentages in Table 8-8 may then be represented in terms of the following letters.

[1] A discussion of how to use logarithms is given in Appendix E, page 598.

$$1900: \frac{P_1}{P_0} \times 100 \qquad 1930: \frac{P_4}{P_3} \times 100$$

$$1910: \frac{P_2}{P_1} \times 100 \qquad 1940: \frac{P_5}{P_4} \times 100$$

$$1920: \frac{P_3}{P_2} \times 100 \qquad 1950: \frac{P_6}{P_5} \times 100$$

$$G = \sqrt[6]{\frac{P_1}{P_0} \times \frac{P_2}{P_1} \times \frac{P_3}{P_2} \times \frac{P_4}{P_3} \times \frac{P_5}{P_4} \times \frac{P_6}{P_5} \times (100)^6} = \sqrt[6]{\frac{P_6}{P_0}} \times 100$$

Table 8-8

COMPUTATION OF THE GEOMETRIC MEAN

PERCENTAGE INCREASE OVER PREVIOUS CENSUS, IN POPULATION OF DETROIT, MICHIGAN: 1900 TO 1950

Census date	Population	Percentage of population on previous census date X	*Log X*
1890.....	205,876	...	...
1900.....	285,704	138.8	2.14239
1910.....	465,766	163.0	2.21219
1920.....	993,678	213.3	2.32899
1930.....	1,568,662	157.9	2.19838
1940.....	1,623,452	103.5	2.01494
1950.....	1,849,568	113.9	2.05652
Total.....	...	...	12.95341

SOURCE: *Statistical Abstract of the United States, 1956*, p. 16.

$$\log G = \frac{\Sigma(\log X)}{N} = \frac{12.95341}{6} = 2.15890$$

antilog 2.15889 = 144.2

$$G = 144.2$$

Average rate of increase = 144.2 per cent − 100.0 per cent = 44.2 per cent

The population increased from 205,876 in 1890 to 1,849,568 in 1950. In other words, the 1950 population was 8.9839 times the 1890 population.

Six decades are covered, and if we want to know the average increase per decade we may find it directly by taking the sixth root of 8.9839. This is the geometric mean of the different rates of increase over the period.

$$G = \sqrt[6]{8.9839}$$

$$\log G = \frac{\log 8.9839}{6} = \frac{.95347}{6} = .15891$$

antilog .15891 = 1.442

$G = 1.442$

Average rate of increase = 1.442 − 1.000 = .442 × 100 = 44.2 per cent

It might be necessary to find the average monthly rate of increase between two census dates, such as 1940 and 1950. The population was 1,623,452 on April 1, 1940, and 1,849,568 on April 1, 1950. The 1950 population was 1.139 times the 1940 population. Since 120 months elapsed between the two census dates, the average monthly increase may be found by extracting the 120th root of 1.139, which will be the geometric mean of the monthly rates of increase.

$$\log G = \frac{\log 1.139}{120} = \frac{.05652}{120} = .000471$$

antilog .000471 = 1.0011

$G = 1.0011$

Average monthly rate of increase = 1.0011 − 1.0000 = .0011 or .11 per cent

The population of Detroit increased .11 per cent each month between the 1940 and the 1950 censuses. If we wish to estimate the population at any date between these two census dates, it can be done by use of the average monthly percentage of increase. January 1, 1945, fell 57 months after the 1940 census date. Since the population had been increasing at the rate of .11 per cent each month for these 57 months, it is possible to compute the average percentage of increase that would be attained on January 1, 1945. Each month was 1.0011 times the preceding month, and in 57 months this process would have taken place 57 times. January 1, 1945, was therefore $(1.0011)^{57}$, which equals 1.0638. The population on January 1, 1945, may be estimated at 1.0638 times the population on April 1, 1940, or 1.0638 times 1,623,452. This multiplication gives an estimated population of 1,727,028 on January 1, 1945.

Median

The *median* is found by arranging the values of the variable in an array and finding the middle value. An *array* is an arrangement in order of size from the smallest to the largest, or from largest to smallest. The following shows the location of the median of the grades of a student on five one-hour examinations.

93%
89%
76% The middle grade is 76 per cent. This is the median.
74%
50%

If there is an even number of items, there obviously is no middle value. The median is defined in this case as midway between the two middle items. This is a practical way out of a definition that is in itself indeterminate. Another case in which the median is indeterminate is illustrated by the following values. If 89 per cent is taken as the median, there is one grade

93%
91%
89%
89%
80%

smaller and two larger. It is impossible to find a value of the variable that has the same number larger as smaller; so, strictly speaking, the median is indeterminate. In such cases, especially where there is a large number of items, the value nearest the middle would be considered the median. In the above case 89 per cent would be taken as the median grade.

In actual statistical work the median is very often approximated from a frequency distribution, except when the individual values are available and the number is small. It is a burdensome job to arrange a large number of values in an array — much more work than to classify them into a frequency distribution. Also, the individual items are frequently not available, although a frequency distribution may be. For these reasons a method that will permit the approximation of the median from the frequency distribution is needed.

The location of the median is facilitated by the use of a cumulative frequency distribution, used in connection with the simple frequency distribution, as in Table 8-9. Looking down the third column in Table 8-9, it is seen that there are 71 farmers who traveled less than 4.0 miles to buy gasoline and oil. From the second column it is seen that 70 farmers traveled between 4.0 and 6.0 miles. This means that the median distance traveled is between 4.0 and 6.0 miles.

Table 8-9

LOCATION OF THE MEDIAN

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

Miles traveled (Class interval)	Number of farmers	Cumulative number of farmers
0.0– 1.9	19	19
2.0– 3.9	52	71
4.0– 5.9	70	141
6.0– 7.9	39	180
8.0– 9.9	24	204
10.0–11.9	21	225
12.0–13.9	14	239
14.0–15.9	12	251
16.0–17.9	8	259
18.0–19.9	1	260
Total	260	...

SOURCE: Table 8-2.

$$\frac{N}{2} = \frac{260}{2} = 130 \qquad 130 - 71 = 59$$

$$\text{Median} = 4.0 + \frac{59 \times 2}{70} = 4.0 + 1.69 = 5.69 \text{ miles}$$

According to the definition on page 177, the median in Table 8-9 is halfway between the 130th and 131st items. If we did not have the original data, it would be impossible to know exactly the value of these two items; but we can assume values that are not far from correct.

There are 70 items between 4.0 and 6.0 miles, and it is assumed that they are distributed equally between these two values. Obviously this assumption is probably not entirely correct, but it usually gives results that are reasonably accurate. Shown graphically, the 70 items are distributed within the class as follows.

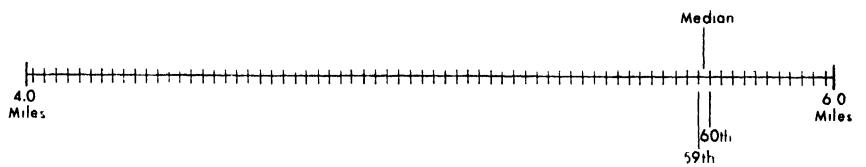


Fig. 8-A **Assumed Distribution of 70 Items Within Class Interval**

It will be observed that there are not only 70 items between 4.0 miles and 6.0 miles, but there are also 70 of the subspaces between the items. Actually there are 69 whole spaces and the half-space at each end. Since the items are all the same distance apart, the size of each space is found to be .0286 miles by dividing the number of spaces, 70, into the class interval, 2 miles.

By consulting the chart we find that the median is located 59 spaces from the lower boundary of the class, and 11 spaces from the upper boundary. Actually, the median is 10 whole spaces and two half-spaces from the upper boundary. This means that the median is 4.0 miles + (59 × .0286), or 4.0 miles + 1.69 miles, which is 5.69 miles. Or, to list all the steps, the computation of the median is as follows:

$$\text{Median} = 4.0 + \frac{59 \times 2}{70} = 4.0 + 1.69 = 5.69 \text{ miles}$$

In actually computing the median it is not necessary to construct the diagram on page 178 in order to find the values for use in the computations. In this computation 4.0 miles is the value of the lower limit of the class in which the median falls; 70 is the number of frequencies in this class; and 2 miles is the class interval. The only value that needs to be computed is the number of spaces the median falls from the limit of the class. This value can always be computed by following the steps outlined below.

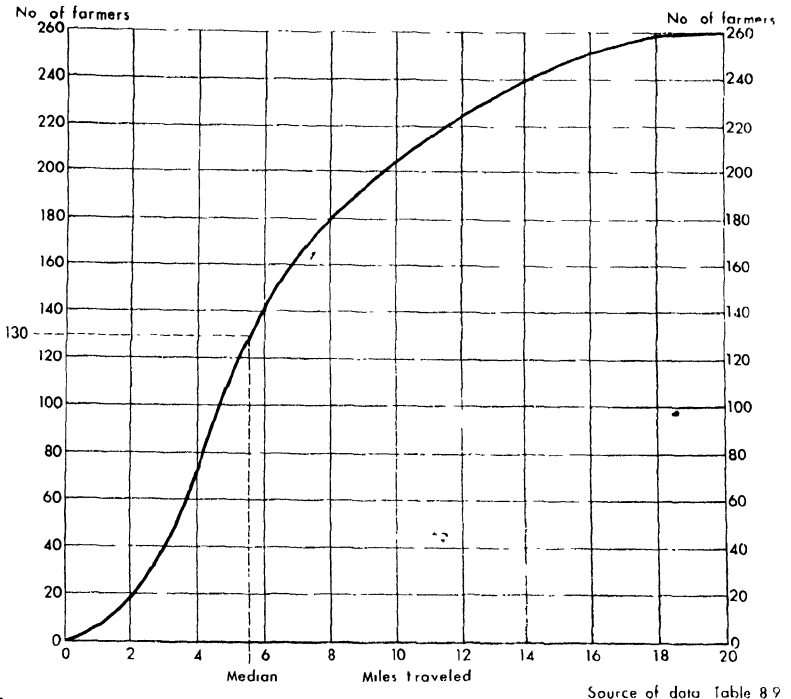
1. Divide the number of frequencies by two. (N/2 = 130)
2. Find the number of frequencies less than the lower limit of the class in which the median falls. (From Table 8-9 this number is found to be 71.)
3. Subtract from N/2 the number of frequencies just found. This will be the number of spaces between the median and the lower limit of the class. (130 − 71 = 59)

Graphical interpolation may be substituted for the above method with quite satisfactory results. Chart 8-1 is an ogive drawn from the data in Table 8-9 (the cumulative frequencies), with the plotted points connected with a curve instead of straight lines. The point at which the curve cuts the horizontal line representing frequency $N/2 = 130$ is the median. The curve gives a more accurate representation of the distribution than connecting the points by a straight line; so if the curve is carefully drawn, and departs considerably from a straight line, the median may be interpolated more accurately by employing the graphic method than the arithmetic method.

Chart 8-1

LOCATING THE MEDIAN GRAPHICALLY

Number of Farmers in the Middle West Traveling Less than Given Distance to Buy Gasoline and Oil



Source of data Table 8 9

Mode

The value of the variable at which the numbers registered are most numerous is called the *mode*. It is the position of greatest density, the predominant or most common value — the value which is the fashion (*la mode*).

The concept of the mode is simple and easy to grasp, but it is not so easy to establish its location in a given case. All the simple methods of locating the mode give merely approximations, and any value found by one of these methods must be used with the realization that it is only approximate.

It should be noted here that the difficulty is in locating the mode of continuous data. When data are definitely discrete, as the distribution of gasoline taxes, the mode is the value of the variable that occurs most frequently. Chart 7-5, page 132, shows two persons to be the modal number of occupants of dwellings in the United States in 1950. It is important to remember that this method of locating the mode is applicable only when the data are discrete; the method cannot be applied to continuous data.

The simplest value for the mode is the mid-point of the class with the greatest frequency. This assumes, of course, that the class taken as the modal class has the same class interval as the other classes. The mid-point of the modal class is referred to as the *crude mode.* It cannot be taken as a very accurate measure of the modal value, for if the data were classified, using class intervals of a different size, the value of the mode might be entirely different. If the same size interval is used, but the class limits located differently, the mode usually will be different, too. The mode, when located by this method, is dependent on the choice of the scale of class intervals; and so it cannot be considered anything more than a crude approximation of the modal value.

Instead of taking the mid-value of the modal class as the mode, it is possible to secure by interpolation within the modal class a closer approximation of the mode. The modal class in Table 8-9 is 4.0 to 5.9 miles, with 70 frequencies. The location of the mode within this class may be made to depend upon the number of frequencies in the two adjoining classes. The first step is to compute the difference between the frequency of the modal class and the frequency of the two adjoining classes. The symbols given below may conveniently be used.

f_1 = frequency of class immediately preceding modal class
f_2 = frequency of class immediately following modal class
f_m = frequency of the modal class
L = lower limit of the modal class
i = class interval

$$d_1 = f_m - f_1 = 70 - 52 = 18$$
$$d_2 = f_m - f_2 = 70 - 39 = 31$$

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2} i \quad \cdots\cdots\cdots \quad (9)$$

$$\text{Mode} = 4.0 + \frac{18}{18 + 31} 2 = 4.0 + .73 = 4.73 \text{ miles}$$

When the frequencies on each side of the modal class are the same, the mode is the mid-value of the modal class. When the frequency is the same for two classes that are tied for greatest frequency, the mode falls on the class limit that divides them.

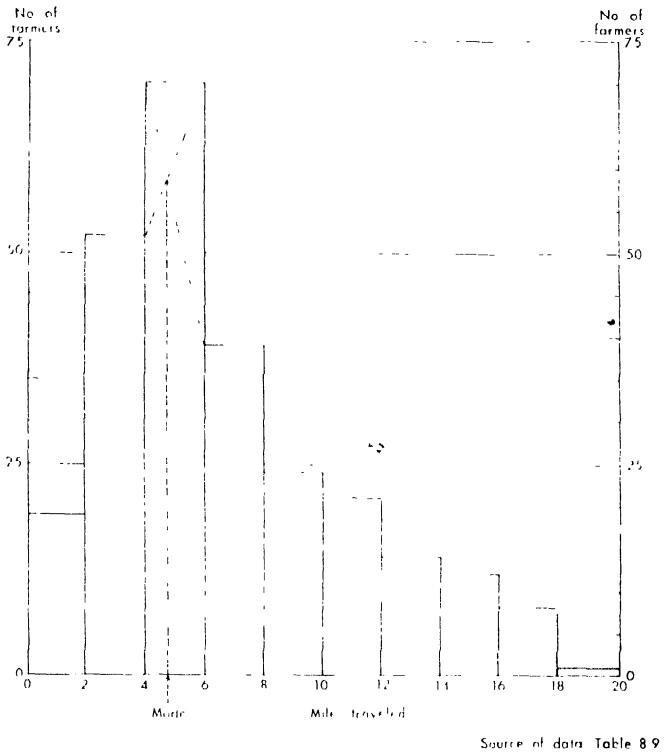
Graphic interpolation of the mode. The mode as just located may also be computed graphically on a histogram. This method is illustrated in Chart 8-2. Lines are drawn diagonally from the upper corners of the rectangle representing the modal class to the upper corners of the adjacent rectangles. These diagonal lines intersect at the modal ordinate; that is, a line through this point and perpendicular to the X axis intersects the X axis

at the modal value. If the chart is drawn accurately, the modal value will agree with the value computed by means of formula (9) on page 181.

Chart 8-2

LOCATING THE MODE GRAPHICALLY

Distance Traveled by 260 Farmers in the Middle West to Buy Gasoline and Oil



Source of data Table 8 9

Empirical mode. In a distribution that is perfectly symmetrical, the mean, median, and mode are the same. As the distribution departs from perfect symmetry, these three values separate and maintain a rather definite relationship to each other; although as the asymmetry becomes more pronounced, this relationship does not hold so closely. When dealing with distributions that are nearly symmetrical, it is possible to find an approximation of the mode by basing it on the relationship between the values of the mean, median, and mode, since the mean and the median may usually be determined. The relationship between the three averages is expressed by the following equation:

$$\textbf{Mode} = \textbf{mean} - 3\,(\textbf{mean} - \textbf{median}) \quad . \quad . \quad . \quad . \quad . \quad \textbf{(10)}$$

This may be described by saying that the median lies one third of the distance from the mean toward the mode.

Applying formula (10) to the data in Table 8-9 for distances farmers traveled to buy gasoline and oil, the following value of the mode is secured:

$$\begin{aligned} \overline{X} &= 6.73 \\ \text{Median} &= 5.69 \\ \text{Empirical mode} &= 6.73 - 3(6.73 - 5.69) = 3.61 \text{ miles} \end{aligned}$$

Interpolating within the modal class gives a value of 4.7 miles for the mode. The fact that the two methods give results that differ so much illustrates that these methods are only approximations, and the results must not be considered to be precise. Locating the mode with greater accuracy than described in the methods above is beyond the scope of this book.

CHIEF CHARACTERISTICS OF THE PRINCIPAL AVERAGES

The arithmetic mean is without doubt the most generally useful of the averages. Its general nature is readily understood, even by people with no training in statistical methods. This is an important characteristic for an average that is to be used in business statistics.

The arithmetic mean is a computed average and is rigidly defined; that is, there is no choice left to the individual computing it. He must follow a rigidly defined procedure; and a number of people using the same data will always arrive at the same answer, assuming, of course, that they make no mistake in calculation. This characteristic may be illustrated by comparison with the mode. If three people were asked to locate the mode of a distribution as accurately as they could, they might each arrive at a different value. This characteristic of the mode is a definite weakness, while the fact that the arithmetic mean is rigidly defined is one of its advantages.

The arithmetic mean lends itself to further algebraic manipulation; to some statisticians this is the most important characteristic an average may possess. If averages of several component parts of a larger group are given, the average for the whole group may readily be expressed in terms of the component parts.

Neither the median nor the mode lends itself to further algebraic treatment. Table 8-10 gives the median family income of each of the four East South Central states in 1950. It also shows the computation of a median family income for the East South Central region, weighting each state median by the number of families in the state. The median income computed from the data for all four states is $1,577. But the median computed from the data for the region, ignoring the state in which the family lived, is $1,555, and is the correct median for the larger group. The median

computed by weighting the median income for each state is not correct, because the median does not lend itself to further algebraic treatment as does the mean. There is no way of computing accurately the median for the region from the medians of the individual states.

Table 8-10

COMPUTATION OF WEIGHTED MEDIAN

FAMILY INCOMES IN THE EAST SOUTH CENTRAL STATES, 1950

State	Median income	Number of families	Median times number of families
Kentucky.....	$1,774	864,730	1,534,031,020
Tennessee.....	1,749	975,485	1,706,123,265
Alabama.....	1,580	865,395	1,367,324,100
Mississippi.....	1,028	619,820	637,174,960
Total.....	...	3,325,430	5,244,653,345

SOURCE: Bureau of the Census, *U. S. Census of Population: 1950.*

$$\text{Weighted median} = \frac{5,244,653,345}{3,325,430} = \$1,577$$

The ease with which an average may be calculated is another characteristic that sometimes must be considered. When the original data are available, there is no significant difference in the ease with which the mean and the median may be calculated. The calculation of the geometric mean requires more work than either the arithmetic mean or the median; but if the conditions warrant the use of the geometric mean, it should be used.

There are times when sufficient information is not available to compute the median. At other times it is possible to compute the median but not the mean. When the total of the values and the number are known, but not the individual values, the mean can be computed, but the median cannot. When there is an open-end class, the mean cannot be computed accurately unless the total or average value of the items in the open-end class is known. The open-end class does not interfere, however, with the location of the median. The location of the mode does not depend at all on the values of the variable at the extremes of the series; so open-end classes make no difference in its location.

The arithmetic mean is based on all the values of the variable, and its size is influenced by each of these values. A few extremely large or a few extremely small values of the variable have an undue amount of influence on the mean. This tendency of the mean is seen in the following distribution of five grades of a student.

83%	89%
89%	83%
81%	81%
20%	77%
77%	20%
5)350	
70%	

The mean is the total, 350, divided by 5, or 70 per cent.

The median is the middle item of the array, 81 per cent.

The mean is 11 less than the median, and an examination of the methods of computation will show why the mean is more affected by the extremely low grade of 20 per cent. The median is influenced only by the *position* of the amounts being averaged, while the mean is affected by the *size* of the amounts since the total of all the amounts is taken. If, instead of being 20 per cent, the lowest grade were 75 per cent, the median would still be 81 per cent and the mean would also be 81 per cent.

83%	89%
89%	83%
81%	81%
75%	77%
77%	75%
5)405	
81%	

The mean is the total, 405, divided by 5, or 81 per cent.

The median is the middle item of the array, 81 per cent.

The question of most importance at this point is, "Which average is more typical of the five grades ranging from 89 per cent down to 20 per cent? The mean is 70 per cent and the median is 81 per cent; they cannot both be considered as equally representative of the performance of a student for the period covered by the tests. When performance of students has been fairly uniform all through a course, it is not so important which average is used to summarize the work, since they will all give approximately the same results. But when an extremely low grade is made in one or two tests, it makes considerable difference in the final grade.

If 70 per cent is taken as the typical grade, the nearest individual grade is 77 per cent, and there are three larger. If the median, 81 per cent, is taken as the typical grade, the nearest individual grade is 83 per cent, and the next removed is four less, 77 per cent. There is one larger than 83 per cent, and then the lowest grade, 20 per cent. In other words, the median grade comes much nearer the individual items except for the one extremely low grade. This practically means that the median ignores this extreme amount, and

is mostly influenced by the remaining amounts concentrated between 77 per cent and 89 per cent. The justification for giving the 20 per cent grade less influence than in the case with the mean is that it is so different from the rest that it is impossible to find any amount that can be representative of it and at the same time be representative of the four grades between 77 per cent and 89 per cent.

The mean attempts to be representative of all the grades, with the result that it is pulled away from the four grades between 77 per cent and 89 per cent toward the 20 per cent grade. It ends by being representative of neither the lowest grade, nor the four higher grades. The median is more representative of the four grades between 77 per cent and 89 per cent, since it is not influenced to any important degree by the one extremely small amount.

A great many distributions of business data, particularly dollar figures, have extremely large amounts which have an undue influence on the mean. In many of these cases the median is the better average to use in summarizing the data. Table 8-11 shows the number of banks in the United States insured by the Federal Deposit Insurance Corporation that had been placed in receivership through 1955. These banks have been classified according to the amount of their deposits.

Table 8-11

BANKS INSURED BY THE FEDERAL DEPOSIT INSURANCE CORPORATION PLACED IN RECEIVERSHIP BETWEEN 1934 AND 1955

CLASSIFIED BY AMOUNT OF DEPOSITS

Size of deposits	Number of banks
$100,000 or less	107
$100,000 and under $250,000	109
$250,000 and under $500,000	59
$500,000 and under $1,000,000	63
$1,000,000 and under $2,000,000	42
$2,000,000 and under $5,000,000	30
$5,000,000 and under $10,000,000	10
$10,000,000 and under $25,000,000	5
$25,000,000 and under $50,000,000	4
Total	429

SOURCE: *Statistical Abstract of the United States, 1956*, p. 447.

The deposits of the 429 banks totaled $571,907,000, and the mean deposit was $1,333,000. The median of the data is $248,000, computed by the methods illustrated on pages 177 to 179.

Since the deposit size of a large number of banks concentrated between $100,000 and $250,000, it seems logical to think of the median deposit size of $248,000 as being more representative of the banks than the mean amount of $1,333,000. The large deposits of a few banks exert an undue influence on the mean. Four banks had deposits between $25,000,000 and $50,000,000. The interpretation of the median is that there were the same number of banks with deposits of less than $248,000 as there were banks with more than $248,000. The fact that the banks with the larger deposits were spread out between deposits of $248,000 and $50,000,000 does not influence the median; it is affected only by the number of banks with deposits larger than the median. The mean is influenced by the size of each individual amount, and so the larger amounts have by far the most influence.

Another distribution of business data that is far from symmetrical is shown in Table 8-12.

Table 8-12

SIZE OF FARMS IN THE UNITED STATES, 1950

Number of acres	Number of farms
Under 10	484,914
10– 29	853,608
30– 49	624,242
50– 69	426,887
70– 99	620,914
100–139	579,111
140–179	523,451
180–219	275,009
220–259	212,316
260–499	478,084
500–999	182,264
1,000 and over	121,362
Total	5,382,162

SOURCE: *Statistical Abstract of the United States, 1956*, p. 624.

A businessman planning a trip to Europe wanted to know the average size farm in the United States, since he expected to compare a typical farm in this country with a typical European farm. A glance at the above frequency distribution will show that it was difficult for him to find any average that could be used as typical of all farms in the United States. The data

are decidedly heterogeneous, since they include so many different kinds of farms in different parts of the country. The mean size farm in 1950 was 215.5 acres, computed by dividing the total land in farms, 1,159,789,000 acres, by the total number of farms 5,382,162. The median size farm was 84.6 acres. This is another case of a small percentage of the items being extremely large, and therefore exerting an undue amount of influence on the mean, pulling it away from the part of the distribution where the items concentrate. The median is a more representative average than the mean in this case.

The following summary briefly outlines the chief characteristics of the arithmetic mean, geometric mean, median, and mode.

Arithmetic Mean

1. The value of the arithmetic mean is based on all the observations, and thus is affected by all the values of the variable. This may result at times in giving certain extreme values too much influence.

2. The arithmetic mean is rigidly defined and is always determinate if the individual values of the variable are available.

3. The arithmetic mean may be calculated if the individual items are known, if the total value and the number of items are known, or if a frequency distribution is available from which it is possible to estimate closely the average value of the items falling in each class.

4. The arithmetic mean lends itself to further algebraic manipulation.

Geometric Mean

1. The geometric mean is based on all the observations, and thus is affected by all the values of the variable. However, it gives less weight to extremely large values than does the arithmetic mean.

2. The geometric mean is zero if any value of the variable is zero, and it may become imaginary if negative values occur. Except for these cases the value is always determinate and is rigidly defined.

3. The geometric mean is the average to use when rates of change or ratios are being averaged, and *it is intended to give equal weight to equal rates of change.*

4. The geometric mean lends itself to further algebraic manipulation.

Median

1. The median is affected by the position of each item in the series but not by the value of each item. This means that extreme deviations from the central part of the distribution affect the median less than the arithmetic mean.

2. The median may be located when the data are incomplete, e.g., if the values of the variable at the extremes are unknown, but their general location is known.

3. The median is, strictly speaking, indeterminate for an even number of cases, although by general agreement it is the mean of the two central values of the variable. When several items in the center of the distribution are the same size, the median may be indeterminate to a slight degree.

4. The median does not lend itself to algebraic treatment in so satisfactory a manner as the arithmetic mean or the geometric mean.

Mode

1. The value of the mode is determined by the items at the point of greatest concentration, and is not affected by the remaining values of the variable.

2. Since the mode is the point of greatest concentration, it is typical of the distribution. When more than one mode exists, it probably means that the distribution should be broken into more than one distribution to secure greater homogeneity.

3. The true mode is difficult to compute, although an approximate value is easily found.

4. The mode does not lend itself to algebraic treatment.

STUDY QUESTIONS

1. The average-size farm in the United States in 1954 was 242 acres. Explain how this figure might be considered to be a ratio.

2. What danger is there in using an arithmetic mean without knowing anything about the size of the individual items included in the total?

3. If a salesman says that he must call on three prospects "on the average" to make one sale, which average do you think he means? Explain.

4. Why is the arithmetic mean the most generally used of the averages? Do you think it should be the most widely used average?

5. The fact that the arithmetic mean can be used logically in further calculations is given as an advantage over the use of the median and the mode. Why is this factor considered to be an important advantage?

6. Do you consider the fact that the geometric mean cannot be used if one of the items is zero to be an important disadvantage of the geometric mean? Give a reason for your answer.

7. Explain why the averaging of percentages, ratios, and rates must be done with unusual care.

PROBLEMS

1. Assume that you ride a commuter train to work and that you live 10 miles from the railway station. You allow 20 minutes to drive to the station since it is normally no problem to average 30 miles per hour. One morning, because of street repairs, your speed is reduced to 15 miles per hour for half of the distance. How fast will you need to drive for the last half of the distance in order to catch the train? (You had allowed yourself the usual 20 minutes, thinking that you could drive at 30 miles per hour all the way, so you must average 30 miles per hour.)

2. The average hourly earnings of 7,630,000 production employees in durable-goods manufacturing concerns in the United States in 1956 was $2.10. The average hourly earnings of 5,444,000 production employees in nondurable-goods manufacturing concerns was $1.81. Compute the average hourly earnings of all production workers in manufacturing in 1956.

3. Refer to the table in Problem 4, page 68. If the solution to Problem 4 has not been retained, it will be first necessary to complete step (a) below.

(a) Set up a tally sheet and classify the employees according to age.
(b) Compute the arithmetic mean age of the 225 employees, using formula (1), page 164, and the individual ages.
(c) Compute the arithmetic mean age of the 225 employees, using formula (2), page 166. How does this mean compare with that secured in (b)?
(d) Compute the arithmetic mean age of the 225 employees, using deviations from an arbitrary origin. (If uniform class intervals were used, deviations in class interval units may be used.)
(e) Compute the median age of the 225 employees. Is the median or the arithmetic mean the better average to use for this distribution? Why?
(f) Compute the mode for the ages of the 225 employees by two methods. Which method do you consider to be better? Does the mode have any distinct advantage over the mean and the median for this distribution?
(g) *Retain this solution for use in Problem 1 of Chapter 9.*

4. Each of five typewriting students copied 10 pages of printed matter at the following speeds. Each page contained 300 words.

Words Per Minute
45
60
73
56
34

(a) What was the average speed of the five typists (in words per minute)?
(b) If the five students had each typed for 10 minutes at the speeds given, what would be the average speed (in words per minute)?

5. A fund of $30,000 was invested in 1947, and for 10 years all dividends and interest payments received were reinvested. At the end of the 10 years the total of the fund was $42,312.64. What was the average rate of return, compounded annually, on the original investment?

6. Table 6-5 (page 116) gives the retail prices of five commodities for November, 1954, and November, 1955. The following tabulation gives the ratio of each of the November, 1955, prices to the November, 1954, prices.

(a) Compute the geometric mean and the arithmetic mean of these ratios.
(b) Explain the significance of each of these averages.

Commodity	Ratio of Nov., 1955, to Nov., 1954
Round steak	.953
Frozen peas	1.109
Grapes	.805
Coffee	.885
Eggs	.875

7. The following table gives a classification of earnings in dress manufacturing shops in Cleveland and Dallas.

(a) Express the frequencies in each distribution as percentages of the total of the distribution.
(b) From the relative frequencies secured in (a), construct two frequency polygons on the same grid.
(c) Summarize what the chart shows about the distribution of earnings in the two cities.
(d) Compute the following values for each distribution:
1. Arithmetic mean.
2. Median.
3. Mode [by formula (9), page 181].

DISTRIBUTION OF STRAIGHT-TIME HOURLY EARNINGS OF PRODUCTION WORKERS IN WOMEN'S AND MISSES' DRESS MANUFACTURING SHOPS IN CLEVELAND AND DALLAS

Hourly earnings (Cents)	Number of workers	
	Cleveland	Dallas
60 and under 80	26	273
80 " " 100	182	405
100 " " 120	241	398
120 " " 140	184	336
140 " " 160	112	158
160 " " 180	55	89
180 " " 200	31	36
200 " " 220	34	9
220 " " 240	17	9
240 " " 260	16	...
260 " " 280	2	2
280 " " 300	4	...
300 " " 320	4	2
Total	908	1,717

SOURCE: *Monthly Labor Review*, Vol. 79, No. 5 (May 1950), p. 539.

9 DISPERSION

Another characteristic of considerable importance in summarizing a mass of figures is the *scatter*, or *dispersion*, among the individual items. Some groups of data tend to cluster closely around a certain value, while others scatter widely. Chart 7-13 (page 138) shows the difference in the dispersion of salaries paid women calculating machine operators in New Orleans and St. Louis. The salaries paid in New Orleans concentrate more at about $35, and do not spread over so wide a range as the salaries paid in St. Louis. The difference in the variation in salaries paid in the two cities is a significant characteristic of the two distributions. In general, differences in the degree of dispersion are important in the analysis of the values of a variable.

IMPORTANCE OF DISPERSION

The use of an average as typical of a distribution should be made carefully. The value of the mean depends essentially on the dispersion of the items around it and on its position in the series. Before a mean may be considered typical, the dispersion of the series around it should be examined. The highest degree of concentration would be to have all the individual items the same size. The scatter in such a case would be zero, and the mean would be exactly the same as the individual values of the variable. The more the individual amounts differ from each other, the less typical of the whole distribution an average will be.

Assume that a golf foursome is made up of individuals with the following annual incomes: Adams, $1,000,000; Baker, $800,000; Cole, $5,500; and Davis, $5,000. What is the typical income of these individuals? The mean is $452,625, but is it *typical* of the incomes of the four individuals? In this particular case the mean is much larger than two, and much smaller than the other two, and so it is difficult to think of it as being representative of any of the four.

Since an average may not always be typical of the distribution from which it is computed, it is important to have some method of testing the reliability of measures of typical size. How well a particular average

represents a distribution depends on the degree of scatter among the items in the distribution. Therefore, the problem of determining the representativeness of an average is a problem of measuring the degree of dispersion among the amounts averaged. Since dispersion may not be the same in both directions, it is often important to measure the *skewness*, or lack of symmetry, of the data. Measures of skewness are discussed on pages 207 to 209.

The degree of dispersion is related to the homogeneity of the data used. If wages of unskilled laborers in a manufacturing concern are analyzed, there will ordinarily be less dispersion than is found for all laborers, skilled and unskilled. The second group, including both classes of laborers, is less homogeneous than if skilled and unskilled laborers are analyzed separately. The problem of homogeneity raises itself whenever an average is computed. It may well be taken as a general rule that an average will be a satisfactory measure of typical size only when it is derived from data that are homogeneous for the purposes of the particular investigation.

Dispersion may be important not as merely supplementary to the mean, but because the scatter in a distribution may itself be significant. In evaluating the performance of a student, it might be significant to measure the *consistency* of his work. If the grades made on different tests showed a wide dispersion, it would mean that his work was sometimes good and sometimes poor. Two students might have the same median grade on the work for the year, yet the individual grades might be decidedly different. All of one student's grades might be close to the median, while the grades of the second might vary from 100 per cent to 20 per cent. In other words, dispersion is often an important characteristic and must be measured for the information it gives about the data.

A manufacturer interested in controlling the quality of his product must try to prevent variations between individual units. A manufacturer of light bulbs obviously should try to produce bulbs that will burn a long time, but it is also desirable that there be as little variation as possible in the length of life between individual bulbs. *Uniformly* high quality in a product is better than a high average in quality with wide variations between units.

When a measure of typical size and a measure of dispersion have been computed for a series, ordinarily the two most important characteristics have been summarized. It is necessary to compute very precise measures of dispersion since many statistical analyses require a concise summary of dispersion. If the dispersion of a series is to be included in a formula, this dispersion must be expressed as a quantitative measure. Inclusion in further computations represents the most important use of certain measures of dispersion.

MEASURES OF ABSOLUTE DISPERSION

The measures of absolute dispersion that will be discussed in the following paragraphs are:

1. Range.
2. Quartile deviation.
3. Average deviation.
4. Standard deviation.

Range

The simplest measure of dispersion is the *range* of the data, that is, the distance between the smallest and the largest amounts. Very frequently this is expressed by merely giving the smallest and the largest amounts. A production manager might say that the average daily wage in a certain department is $23.98, and that the individual daily wages range from $12 to $39. This would give a rough measure of scatter, which could be compared with other departments. A second department might have a mean daily wage of $24.61, with daily wages ranging between $19 and $30. It would be concluded from these figures that the average of the second department is probably more representative of the wage distribution than the average for the first, since there is less scatter in the wages received by employees in the second department. The high and low prices for which securities or commodities sell on an exchange in a given day are reported regularly in financial periodicals, and serve as a reasonably accurate measure of the dispersion in prices for each day.

The obvious weakness of such a measure of dispersion is that it is based on only two items, and tells nothing about the manner in which the other items scatter. The range is a simple and easily understood measure of scatter, but it is less accurate than other measures.

The fact that the range can be computed with very little work has resulted in its being used almost exclusively to measure dispersion in the construction of control charts for variables used in statistical quality control. Maintaining control charts requires the measurement of the dispersion in many distributions, and the saving of time resulting from the use of the range becomes an important factor. The types of distributions used in this work do not distort the range as much as some distributions, so it serves as a reasonably accurate measure of dispersion. The use of the range in connection with statistical quality control is described in detail in Chapter 23.

Quartile Deviation

The median was defined as the point that divided the distribution into two equal parts, half the items falling on one side and half on the other.

By exactly the same process it is possible to divide an array or a frequency distribution into four equal parts. Three points must be located. The first is called the *first quartile* (Q_1), and is the point that has one fourth of the frequencies smaller and three fourths larger. The *third quartile* (Q_3) has one fourth of the frequencies larger and three fourths smaller. The *second quartile* is the median.

Obviously half the frequencies fall between the first and third quartiles, since one fourth fall between the first and second quartiles, and one fourth between the second and third. The distance between the first and third quartiles is the *interquartile range*; the smaller this distance, the greater the degree of concentration of the middle half of the distribution. A measure of dispersion based on the interquartile range does not give any influence to the items above and below the middle half of the distribution, but it is much superior to the range. As stated before, the range is based on only two items, the largest and the smallest amounts of the array.

The measure of dispersion based on the interquartile range is called the *quartile deviation* or *semi-interquartile range*. It is computed by the formula

$$Q = \frac{Q_3 - Q_1}{2} \quad \ldots\ldots\ldots\ldots \quad (11)$$

It is obvious that the chief problem in computing the quartile deviation is locating the first and third quartiles. In the following array of daily wages, taken from Table 9-1, the quartiles are found by locating the points that mark off the lower fourth and the upper fourth of the values of the variable. Q_1 is \$12.00, and Q_3 is \$17.25; hence

$$Q = \frac{\$17.25 - \$12.00}{2} = \frac{5.25}{2} = \$2.63.$$

\$11.10
11.50
12.00 —— Q_1
12.20
12.45
13.50 —— Median
14.00
16.00
17.25 —— Q_3
18.50
20.00

When the quartiles are to be located from a frequency distribution, the procedure resembles the location of the median. The only difference is that instead of using $N/2$ in the computations, $N/4$ is used in computing the

first quartile, and $3N/4$ in computing the third quartile. After $N/4$ and $3N/4$ have been found, the procedure is the same as that used in locating the median.

For the data in Table 8-9, page 178, $\frac{N}{4} = \frac{260}{4} = 65$. There are 19 frequencies less than 2.0 miles (19 farmers traveled less than 2.0 miles to buy gasoline and oil) and 71 less than 4.0 miles. Therefore, Q_1 falls between 2.0 miles and 4.0 miles. Out of the 52 spaces in this class, 46 of them are needed to reach the first quartile ($65 - 19 = 46$), so

$$Q_1 = 2.0 + \frac{46 \times 2}{52} = 2.0 + 1.77 = 3.77$$

In the same table, $\frac{3N}{4} = \frac{3 \times 260}{4} = 195$. Cumulating the frequencies, 180 are found to be less than 8.0 miles and 204 are less than 10.0 miles. Therefore, Q_3 falls between 8.0 miles and 10.0 miles. Out of the 24 spaces in this class, 15 are needed to reach the third quartile ($195 - 180 = 15$). The following computations give the third quartile.

$$Q_3 = 8.0 + \frac{15 \times 2}{24} = 8.0 + 1.25 = 9.25$$

Since the first quartile is 3.77 miles and the third quartile is 9.25 miles, the interquartile range is 5.48 miles (the distance between the two quartiles). Half of the interquartile range (the semi-interquartile range) is the quartile deviation.

$$Q = \frac{9.25 - 3.77}{2} = \frac{5.48}{2} = 2.74$$

Average Deviation

The *average deviation* is one measure of dispersion that is based on all the items in a distribution. The computation of the average deviation may be illustrated by the daily wage distribution given in Table 9-1.

The third column in Table 9-1 shows the deviation of each of the individual items from the median. The average deviation is found by computing the arithmetic mean of these individual deviations, disregarding the signs. It will be noticed that the signs of the deviations were not entered in Table 9-1 since they are not used in the computations. The average deviation from the mean may be computed in the same manner as from the median, and is frequently used as a measure of dispersion.

Table 9-1

COMPUTATION OF THE AVERAGE DEVIATION

DAILY WAGES OF ELEVEN EMPLOYEES

Employee	Daily wages (Dollars)	Deviations from the median ($13.50)
Chester	11.10	2.40
Alexander	11.50	2.00
Smith	12.00	1.50
Burnette	12.20	1.30
Johnson	12.45	1.05
Davis	13.50	0
Henry	14.00	.50
Murphy	16.00	2.50
Cohen	17.25	3.75
Yates	18.50	5.00
Cunningham	20.00	6.50
Total	...	26.50

SOURCE: Hypothetical data.

$$\text{Average deviation} = \frac{26.50}{11} = \$2.41$$

It is common to have data in the form of a frequency distribution rather than in the form of individual amounts, and so the problem arises of computing the average deviation of data in classified form. Table 9-2 on the next page shows the computation of the average deviation of items grouped into a frequency distribution.

The first step in computing the average deviation is to determine the deviation of the mid-value of each class from the median. It is assumed that this shows the amount of deviation for *each value in the class*, and so the total deviation of all the items in a class is found by multiplying the deviation by the frequency of the class. The sum of these products, disregarding the signs, is the sum of the deviations from the median. Dividing this sum by 260 gives the average deviation.

Standard Deviation (σ)

Another method of measuring the dispersion of a series resembles the average deviation, but employs a different method of averaging the deviations. In computing the *standard deviation*, the deviations from the *mean*

Table 9-2

COMPUTATION OF THE AVERAGE DEVIATION (GROUPED DATA)

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

Miles traveled (Class interval)	Miles traveled (Mid-values) *m*	Number of farmers *f*	Deviation of mid-values from median (5.69)	Frequency times deviation
0.0– 1.9	.95	19	4.74	90.06
2.0– 3.9	2.95	52	2.74	142.48
4.0– 5.9	4.95	70	.74	51.80
6.0– 7.9	6.95	39	1.26	49.14
8.0– 9.9	8.95	24	3.26	78.24
10.0–11.9	10.95	21	5.26	110.46
12.0–13.9	12.95	14	7.26	101.64
14.0–15.9	14.95	12	9.26	111.12
16.0–17.9	16.95	8	11.26	90.08
18.0–19.9	18.95	1	13.26	13.26
Total.....	...	260	...	838.28

SOURCE: Table 8-2.

$$\text{Average deviation} = \frac{838.28}{260} = 3.22$$

are first squared, which makes all of them positive. These squared deviations are totaled, divided by the number, and then to compensate for squaring the deviations the square root is taken. The result, after extracting the square root, is the standard deviation. It is the *square root of the mean of the deviations squared*, or the root-mean-square deviation from the mean. If deviations from the mean are represented by x,

$$\sigma^2 = \frac{\Sigma x^2}{N} \quad \text{. . . . . (12)}$$

The computation of the standard deviation is illustrated in Table 9-3. The mean is found to the nearest cent, \$14.41, and then the deviation of each observation from it is written in the third column. It will be noticed that the sum of this column is $-.01$; the total differs from zero because the mean was rounded off. Finally, each deviation is squared, and the squares entered in the last column. A table of squares (see Appendix C) may be used for this, or the deviations may be squared by multiplication on a calculator. The sum of the squares is 93.5241. Therefore

$$\sigma^2 = \frac{93.5241}{11} = \$8.5022$$

$$\sigma = \$2.92$$

Table 9-3

COMPUTATION OF THE STANDARD DEVIATION

DAILY WAGES OF ELEVEN EMPLOYEES

Employee	Daily wages (Dollars) X	Deviations from the mean ($14.41*) x	Square of the deviations x^2
Chester	11.10	− 3.31	10.9561
Alexander	11.50	− 2.91	8.4681
Smith	12.00	− 2.41	5.8081
Burnette	12.20	− 2.21	4.8841
Johnson	12.45	− 1.96	3.8416
Davis	13.50	− .91	.8281
Henry	14.00	− .41	.1681
Murphy	16.00	1.59	2.5281
Cohen	17.25	2.84	8.0656
Yates	18.50	4.09	16.7281
Cunningham	20.00	5.59	31.2481
Total	158.50	− .01	93.5241

SOURCE: Table 9-1.

*It is not necessary to carry the mean to more decimal places than the original data, for slight errors in the mean will not appreciably affect the value of the standard deviation.

$$\bar{X} = \frac{\Sigma X}{N} = \frac{158.50}{11} = \$14.41$$

$$\sigma^2 = \frac{\Sigma x^2}{N} = \frac{93.5241}{11} = 8.5022$$

$$\sigma = \$2.92$$

A quantity analogous to the standard deviation may be defined in more general terms. Let A be any arbitrary value of X, and let d denote the deviations of X from A; i.e., let

$$d = X - A$$

Then we may define the root-mean-square deviation (s) from the arbitrary origin A by the equation

$$s^2 = \frac{\Sigma d^2}{N} \quad \text{(13)}$$

The standard deviation is the root-mean-square deviation from the *mean*, rather than from any other point. There is a very simple relationship

between the standard deviation and the root-mean-square deviation from any other origin. Let

$$X - A = c$$

so that $$d = x + c$$

Then $$d^2 = x^2 + 2xc + c^2$$

$$\Sigma d^2 = \Sigma x^2 + 2c \cdot \Sigma x + Nc^2$$

But the sum of the deviations from the mean is zero, so the second term vanishes, and

$$\Sigma d^2 = \Sigma x^2 + Nc^2$$

$$\frac{\Sigma d^2}{N} = \frac{\Sigma x^2}{N} + c^2$$

But $$\frac{\Sigma x^2}{N} = \sigma^2 \quad \text{[equation (12)]}$$

and $$\frac{\Sigma d^2}{N} = s^2 \quad \text{[equation (13)]}$$

Therefore $$s^2 = \sigma^2 + c^2$$

$$\sigma^2 = s^2 - c^2 \quad \ldots\ldots\ldots\ldots \quad \textbf{(14)}$$

In computing the standard deviation, it is usually a saving of time to take some round number as the arbitrary origin and compute the root-mean-square deviation from it. Sometimes the arbitrary origin is taken near the mean, but zero may be taken. If the arbitrary origin is taken at zero, the deviations are merely the values of the variable (X). This saves the time necessary to subtract the arbitrary origin, but it gives larger deviations to square. When zero is the arbitrary origin

$$s^2 = \frac{\Sigma X^2}{N} \quad \ldots\ldots\ldots\ldots \quad \textbf{(15)}$$

and $$c^2 = \left(\frac{\Sigma X}{N}\right)^2 \quad \ldots\ldots\ldots\ldots \quad \textbf{(16)}$$

The computation of the standard deviation by assuming an arbitrary origin near the mean is illustrated in Table 9-4. The computations are very much like those in Table 9-3 except that the deviations are measured from \$10 instead of from the mean, \$14.41. The value of s^2 is found instead of the value of the standard deviation, and it therefore becomes necessary to correct this to find the standard deviation.

Basically the standard deviation is similar to the average deviation, being merely a different method of averaging the deviations. Squaring the individual deviations gives the larger deviations more influence than is given them by the method of averaging used in computing the average deviation. For this reason the standard deviation is always larger than the average deviation from the mean, except, of course, when all the deviations being averaged are the same size, in which case the average deviation and the standard deviation are identical.

Table 9-4

COMPUTATION OF THE STANDARD DEVIATION USING AN ARBITRARY ORIGIN

DAILY WAGES OF ELEVEN EMPLOYEES

Employee	Daily wages (Dollars) X	Deviations from the arbitrary origin ($10) d	Square of the deviations d^2
Chester	11.10	1.10	1.2100
Alexander	11.50	1.50	2.2500
Smith	12.00	2.00	4.0000
Burnette	12.20	2.20	4.8400
Johnson	12.45	2.45	6.0025
Davis	13.50	3.50	12.2500
Henry	14.00	4.00	16.0000
Murphy	16.00	6.00	36.0000
Cohen	17.25	7.25	52.5625
Yates	18.50	8.50	72.2500
Cunningham	20.00	10.00	100.0000
Total	158.50	...	307.3650

SOURCE: Table 9-1.

$$s^2 = \frac{\Sigma d^2}{N} = \frac{307.3650}{11} = 27.9423$$

$$\bar{X} = \frac{158.50}{11} = 14.41$$

$$A = 10.00$$

$$c = \bar{X} - A = 14.41 - 10.00 = 4.41$$

$$c^2 = 19.4481$$

$$\sigma^2 = s^2 - c^2 = 27.9423 - 19.4481 = 8.4942$$

$$\sigma = \$2.91$$

Importance of the standard deviation. The importance of the standard deviation is similar to that of the average deviation. Since they are both averages of the deviations from an average, they measure the average amount of scatter in a distribution. But statisticians cling rather tenaciously to the standard deviation in preference to the average deviation. The standard deviation is based on all the values of the variable, and it is logical mathematically. Ignoring the signs in computing the average deviation is a mathematical weakness of that measure.

Since the chief reason for computing a measure of dispersion is to use it in further computations, the standard deviation holds an important place in the discussion of dispersion. The quartile deviation and the average deviation are probably as good or better measures of dispersion than the standard deviation, when they are not to be used further in another formula. But for this latter use the standard deviation is accepted as superior.

Probably the most important use of a measure of dispersion in further calculations occurs in the measurement of the reliability of computations from samples. Measures of the precision of estimates made from samples can be computed for various statistical measures. Such computations make use of a measure of dispersion, and in every case the standard deviation is used in preference to any other.

When the distribution from which the standard deviation is computed is a normal distribution (described in detail on pages 209 to 212), 68.27 per cent of the values of the variable will be found between plus and minus one standard deviation from the arithmetic mean. Plus and minus two standard deviations from the mean will include 95.45 per cent of the items, and plus and minus three standard deviations will include 99.73 per cent of the items. When the distribution is not perfectly symmetrical, the percentages of items will vary from the above; but for a distribution only moderately asymmetrical they hold approximately.

To aid in the interpretation of the quartile deviation it was pointed out that half the values of the variable fall between the first and third quartiles, the interquartile range. It follows, therefore, that one fourth the values fall within the semi-interquartile range, which is the quartile deviation. In the same manner it will help in interpreting the significance of the standard deviation of a symmetrical distribution to remember that approximately 68 per cent of the values of the variable fall within a range of two times the standard deviation, 95 per cent within a range of four times the standard deviation, and practically all within a range of six times the standard deviation. These relationships are particularly important in the analysis of the reliability of samples, as will be discussed in the following chapter.

Computation of the standard deviation from a frequency distribution. When data are available only in classified form, it is necessary to have some method of computing the standard deviation from the classification instead of from the individual items. In fact, the computation from a frequency distribution is so much shorter when the number of items is large that it would be a saving of time to make a frequency distribution to use in computing the mean and standard deviation. The basic principle employed in this method is the same as the one used to approximate the mean and the average deviation from the frequency distribution. The mid-point of each class is assumed to be the average value of all the items in the class, and the standard deviation is based on the deviation of these mid-points from the mean. This assumption is not entirely correct, but the grouping error is usually relatively small, and the approximation is widely used without attempting to compute the value more accurately.

The mid-value of one of the classes is chosen as the arbitrary origin (A) from which to measure the deviations. The class interval is used as the unit throughout the arithmetic, and the value of each frequency is assumed to be that of the mid-value of the class.

By equation (13)

$$s^2 = \frac{\Sigma d^2}{N}$$

when d represents the deviations of individual observations from the arbitrary origin. When the deviations are the differences between the mid-values and the arbitrary origin, it becomes necessary to multiply the square of each deviation by the frequency of the respective class. The equation then becomes

$$s^2 = \frac{\Sigma f(d')^2}{N} \text{ in class interval units} \quad \ldots\ldots \quad (17)$$

The equation for the standard deviation is

$$\sigma^2 = s^2 - c^2 \quad [\text{equation (14)}]$$

But if s^2 is in class interval units, c^2 must be kept in the same units. When the standard deviation has been computed, it is transformed from class interval units to original units, by multiplying by the class interval.

The computation of the standard deviation from a frequency distribution is illustrated in Table 9-5 on the next page. It will be observed that this table merely extends Table 8-4 (page 168) by the addition of one column, $f(d')^2$, which gives the value to substitute in equation (17).

Table 9-5

COMPUTATION OF THE STANDARD DEVIATION (GROUPED DATA)

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

Miles traveled (Class interval)	Number of farmers f	d'	fd'	$f(d')^2$
0.0– 1.9	19	– 4	– 76	304
2.0– 3.9	52	– 3	– 156	468
4.0– 5.9	70	– 2	– 140	280
6.0– 7.9	39	– 1	– 39	39
8.0– 9.9	24	0	0	0
10.0–11.9	21	1	21	21
12.0–13.9	14	2	28	56
14.0–15.9	12	3	36	108
16.0–17.9	8	4	32	128
18.0–19.9	1	5	5	25
Total...	260	...	– 289	1,429

SOURCE: Table 8-2.

$$s^2 = \frac{\Sigma f(d')^2}{N} = \frac{1{,}429}{260} = 5.4962$$

$$c^2 = \left(\frac{\Sigma fd'}{N}\right)^2 = \left(\frac{-289}{260}\right)^2 = 1.2355$$

$$\sigma^2 = s^2 - c^2 = 5.4962 - 1.2355 = 4.2607$$

σ = 2.06 intervals. (All the above operations were carried through in class intervals.)

Since the class interval is 2 miles

$$\sigma = 2.06 \times 2 \text{ miles} = 4.12 \text{ miles}$$

CHIEF CHARACTERISTICS OF MEASURES OF DISPERSION

The following summary briefly outlines the chief characteristics of the measures of absolute dispersion described in this chapter.

Range

1. The range is the simplest measure of dispersion with respect to its calculation and its significance, since it is merely the distance between the largest and smallest items in a distribution.

2. Since the range is based on the two extreme values, it tends to be erratic. A few very large or very small values are not unusual in distributions of business data. When extremely large or small items occur, the range measures their dispersion only and ignores the remainder of the values of the variable.

3. The fact that the range is not influenced by any of the values of the variable except the two extremes is its chief weakness. There is always a danger that it will give a completely incorrect description of the dispersion in the distribution.

4. An extensive use of the range is found in statistical quality control where the distributions used do not distort the range badly and the saving in computation time is an important factor.

Quartile Deviation

1. The quartile deviation is fundamentally the same type of measure as the range since it is based on the range over which the middle half of the distribution scatters. Like the range, it is not affected by the dispersion of all the individual values of the variable.

2. Basing the quartile deviation on the range of the middle half of the distribution is an improvement over the range of the entire distribution because extreme deviations in the largest and smallest values cannot distort the quartile deviation to the extent that they can distort the range.

3. The quartile deviation requires somewhat more work to compute than the range but generally less work than the average deviation or the standard deviation.

Average Deviation

1. The average deviation is the average of the absolute size of the deviations of the individual items from their average value (either arithmetic mean or median).

2. By reflecting the dispersion of every item in the distribution, the average deviation is superior to the range and the quartile deviation as a measure of dispersion.

3. Since the average deviation is the average variation of the individual items from their average value, it has a precise meaning. Its significance is logical and is not difficult to understand.

4. The calculation of the average deviation from individual values of a variable or from a frequency distribution is not difficult.

5. The major flaw of the average deviation arises from the fact that it averages the absolute deviations, that is, it ignores the signs of the deviations. This fact makes its use in further calculations less satisfactory than the standard deviation.

Standard Deviation

1. The standard deviation is the square root of the arithmetic mean of the squared deviations of the individual items from the arithmetic mean. Squaring the deviations makes all of them positive so there is no need to ignore the signs as in the computation of the average deviation.

2. The standard deviation is affected by the deviation of every item.

3. The need to square each deviation increases the amount of work required in comparison with the calculation of the average deviation, but the standard deviation can be computed from the individual values of the variable or from a frequency distribution.

4. The standard deviation resembles the average deviation in that it has a precise meaning, but it has a further advantage since there is no mathematical flaw in its calculation.

5. The fact that the standard deviation is logical mathematically means that it can be used satisfactorily in further calculations. This is its outstanding superiority over the other measures of dispersion.

RELATIVE DISPERSION

The preceding pages have discussed the absolute measures of dispersion, which furnish methods for expressing definitely the amount of scatter in a distribution. But there are times when the problem is to compare the scatter in one distribution with that of another; and, if the items in one are decidedly different in size from the other, it is difficult to compare the degree of scatter in the two. For example, assume that an executive wants to compare the variation in the salaries of clerical workers in the organization with the salaries paid the sales force. The clerical payroll is weekly, and the salesmen are paid monthly. The mean clerical salary is $48.50 and the mean salary of the salesmen is $516.76, the former computed from weekly rates and the latter from the monthly rates. The standard deviation of the clerical salaries is $14.50, and of the salesmen's salaries, $206.70.

It is impossible to make an accurate comparison between these two measures of dispersion because of the differences in the original data. Whether $206.70 is a more important amount of scatter for the salaries paid to the salesmen than $14.50 is for the clerical salaries must be answered. Since the salaries of salesmen are larger, partly because they are expressed on a monthly rather than a weekly basis, it is to be expected that they will show more dispersion than the smaller weekly wages. Some method is needed to express the amount of dispersion in the two series on a *relative* basis.

The most common method of comparing amounts of different sizes is to reduce them to a comparable percentage basis. In this case each standard deviation is expressed as a percentage of the mean of the data from which it was computed. The percentage computed in this way is called the *coefficient of variation*,[1] and is represented by V. The formula for computing V is simply

$$V = \frac{\sigma}{\bar{X}} \cdot 100. \quad \ldots \ldots \ldots \ldots \quad (18)$$

Dividing the standard deviation by the mean and multiplying by 100 expresses the standard deviation as a percentage of the mean.

For salesmen's salaries the coefficient of variation is computed as follows:

$$V = \frac{\$206.70}{\$516.76} \cdot 100 = 40 \text{ per cent}$$

For clerical salaries the coefficient of variation is computed as follows:

$$V = \frac{\$14.50}{\$48.50} \cdot 100 = 30 \text{ per cent}$$

The coefficient of variation for salesmen's salaries may now be compared logically with the coefficient for clerical salaries. These two measures show that there was somewhat more scatter among the salaries of salesmen than among the clerical salaries. The two percentages can be compared, whereas the comparison of the two standard deviations means nothing, with the individual items in one distribution so much different from the items in the other.

Other measures of relative dispersion might be based on the average deviation and the quartile deviation, but the coefficient of variation as described above is the only one generally used. For measuring the dispersion of statistical data used in business, the average deviation and quartile deviation are highly satisfactory and more easily understood than the standard deviation. If they become more popular as measures of dispersion, it would be advisable to compute measures of relative dispersion based on them.

MEASURES OF SKEWNESS

Sometimes it is desirable to compute a measure that shows the direction of the dispersion about the center of the distribution. Measures of dispersion indicate only the amount of the scatter, and give no information about the direction in which the scatter occurs. Measures of *skewness* indicate the lack of symmetry in a distribution, and show the direction in which the skewness extends.

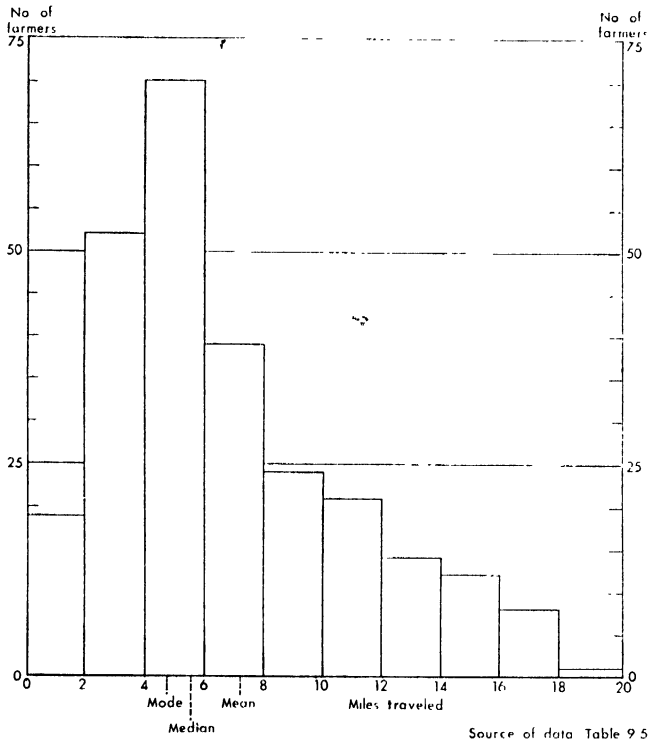
[1] This name was given the measure by Karl Pearson.

The distribution of distances farmers traveled to buy gasoline and oil (Table 9-5, page 204) shows a moderate amount of skewness to the larger values, which is reflected in the difference between the arithmetic mean, median, and mode. Chart 9-1 shows the location of the three averages, with the mean the largest, the mode the smallest, and the median falling between them.

Chart 9-1

DISTANCES TRAVELED BY 260 FARMERS IN THE MIDDLE WEST TO BUY GASOLINE AND OIL

Histogram Showing Location of Mean, Median, and Mode



Source of data Table 9 5

A measure of skewness, based on the difference between the mean and the mode, was suggested by Karl Pearson. The greater the amount of skewness, the more the mean and the mode differ because of the inherent influence of extreme items. By finding the amount of difference between these two averages, it is possible to tell something about the skewness of the distribution. Dividing the difference by the standard deviation gives a coefficient of skewness. The formula is as follows:

$$\text{Skewness} = \frac{\text{arithmetic mean} - \text{mode}}{\text{standard deviation}} \quad \ldots \quad (19)$$

When the distribution is skewed to the larger values, the mean is larger than the mode; therefore the measure of skewness will be positive. When the skewness is to the smaller values, the mean will be smaller than the mode; and the measure of skewness will be negative.

The calculation of the measure of skewness for the distribution of distances farmers traveled to buy gasoline and oil is made by equation (19) as follows:

$$\text{Skewness} = \frac{6.73 - 4.73}{4.12} = \frac{2.00}{4.12} = .49$$

If the empirical mode computed by equation (10) on page 182 is used instead of the mode located within the modal class, the following value of skewness is secured:

$$\text{Skewness} = \frac{6.73 - 3.46}{4.12} = \frac{3.27}{4.12} = .79$$

Or, the same measure of skewness may be computed directly from the values of the mean and the median by the following equation:

$$\text{Skewness} = \frac{3(\text{arithmetic mean} - \text{median})}{\text{standard deviation}} \quad \ldots\ldots \text{(19a)}$$

$$\text{Skewness} = \frac{3(6.73 - 5.64)}{4.12} = \frac{3(1.09)}{4.12} = .79$$

Since neither method of locating the mode is completely satisfactory, both of these measures of skewness should be considered as merely approximations of the degree of skewness.

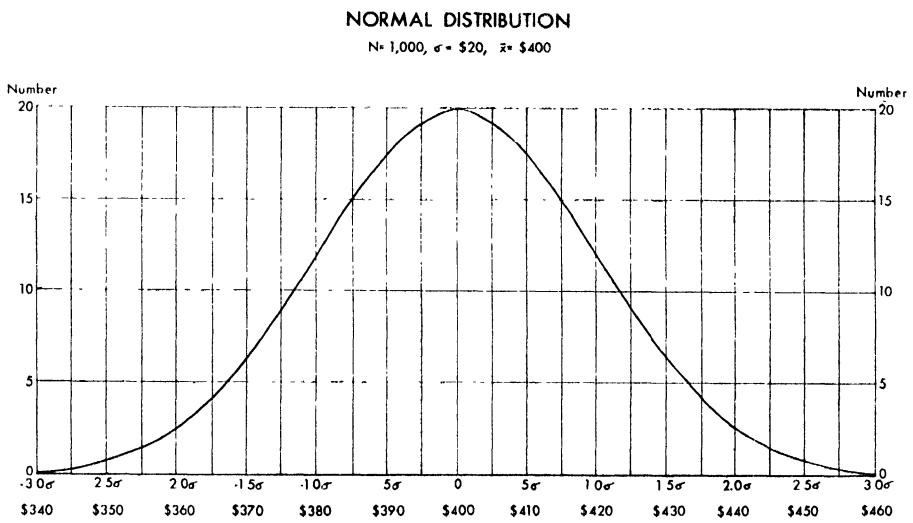
THE NORMAL DISTRIBUTION

The discussion in this chapter and the two preceding chapters has been in terms of actual distributions of quantitative data, such as wages, miles traveled, ages, and other values of variables that could be expressed quantitatively. The discussion in this section will be concerned with a theoretical distribution, known as the *normal distribution*. This distribution was first defined over 200 years ago, and in the early part of the nineteenth century it was considered to be the "law" to which distributions of observations of natural phenomena were believed to conform. But more careful study of natural phenomena has modified this belief as other distribution types have been discovered. Many students have had experience with classes that were "graded on the curve" — an expression generally referring to the assumption that grades follow the normal distribution.

There are varied uses of the normal distribution in describing business data, but the discussion in this text will be limited to its use in sampling. Distributions of business data are probably less inclined to follow the normal distribution than data on natural phenomena, so it is considered that the most important use of the normal distribution in business statistics is in connection with sampling.

When plotted in the same manner as the frequency polygon described on pages 136 to 138, the chart is called the *normal curve*. The origin of the X scale is located in the middle of the distribution. Since the distribution is symmetrical, this mid-point is the mean of the distribution. All the X values are measured as deviations from the mean in standard deviation units and since $x = X - \bar{X}$, the values on the X scale are represented by x/σ. Since the mean is at the middle of the distribution, the values of x/σ will be plus and minus. A normal distribution with a mean of \$400 is shown in Chart 9-2; the values on the X scale are shown both in the original X units and in x/σ units.

Chart 9-2



The height of the *maximum ordinate* of any given normal distribution depends upon the total number in the distribution (represented in this and previous chapters by N) and the value of the standard deviation (σ). The maximum ordinate is represented by Y_0 since it is the value of Y when

$x = 0$. The value of the maximum ordinate of a normal distribution is computed by the following equation:

$$Y_o = \frac{N}{\sigma\sqrt{2\pi}} \text{ or } Y_o = \frac{N}{\sigma} .39894 \quad . \quad . \quad . \quad (20)$$

For a normal distribution with total frequencies of 1,000 ($N = 1{,}000$) and a standard deviation of \$20 ($\sigma = \$20$), the value of the maximum ordinate (Y_o) is computed as follows:

$$Y_o = \frac{1{,}000}{20} .39894 = (50)(.39894)$$

$$Y_o = 19.947$$

Other ordinates of the normal curve may be computed from the values of x by the following equation:[2]

$$Y = Y_o e^{\frac{-x^2}{2\sigma^2}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

Since $Y_o = \frac{N}{\sigma} .39894$, the equation of the ordinates may be written

$$Y = \frac{N}{\sigma} .39894\, e^{\frac{-x^2}{2\sigma^2}}$$

When $x = \$40$, Y is computed as follows:

$$Y = \frac{1{,}000}{20}(.39894)(2.71828)^{-\frac{40^2}{2(20)^2}}$$

$$= \frac{1{,}000}{20}(.39894)(2.71828)^{-2}$$

$$= \frac{1{,}000}{20}(.39894)\frac{1}{(2.71828)^2}$$

$$= \frac{1{,}000}{20}(.39894)(.13534)$$

$$= 2.6995$$

The above value .05399, (.39894)(.13534), can be secured from the Table of Ordinates and Areas under the Normal Curve (Appendix G) under the heading "Ordinate at $\frac{x}{\sigma}$." For the ordinate computed above, $x = \$40$, and since $\sigma = \$20$, $\frac{x}{\sigma} = \frac{40}{20} = 2$. The value from the column "Ordinate at $\frac{x}{\sigma}$" is

[2]In equation (21), the symbol e represents the base of the Naperian or natural logarithms. The value of e to five decimal places is 2.71828.

.05399 when $\frac{x}{\sigma} = 2.00$. This value is multiplied by the value of $\frac{N}{\sigma}$ to secure the ordinate of the normal distribution when $N = 1{,}000$, $\sigma = \$20$, and $\frac{x}{\sigma} = 2.00$.

$$Y = \frac{N}{\sigma}(.05399) = \frac{1{,}000}{20}(.05399) = 2.6995$$

Note that this is the same value computed above by formula (21).

The Table of Ordinates and Areas under the Normal Curve can be used to compute the maximum ordinate (Y_o) in the same manner as any other ordinate. The value in the column "Ordinate at $\frac{x}{\sigma}$" when $\frac{x}{\sigma} = 0$, is .39894, so

$$Y_o = \frac{1{,}000}{20}(.39894) = 19.947$$

By computing a number of ordinates of the normal distribution when $N = 1{,}000$ and $\sigma = \$20$, the distribution may be plotted to give the curve in Chart 9-2. The ordinates used to plot this curve were computed by the use of the values in Appendix G; these values are generally used rather than the formulas.

It is frequently important to know not only the ordinate of the normal curve at a given value on the X axis, but also the proportion of the total frequencies falling between two points on the X axis. For example, 34.13 per cent of the area under the normal curve falls between the maximum ordinate ($x = 0$) and $+1\sigma$. Between $+1\sigma$ and -1σ is 68.26 per cent of the area under the curve. The proportion of the total area lying between any two ordinates may be determined from Appendix G from the column "Area Between Maximum Ordinate and Ordinate at x/σ." Since the curve is symmetrical, the same areas apply to the plus and minus values of x/σ. Important uses of the areas under the normal curve will be discussed in Chapter 10.

STUDY QUESTIONS

1. Why is it important that something be known about the shape of the distribution of the values of a variable if one is to use an average as a measure of typical size?

2. Why is it important that some knowledge of the dispersion be available when using an average as typical of the values of the variable?

3. Why is it important to know something about the skewness of a distribution?

4. Although the range is considered the simplest measure of dispersion, it is not a very reliable measure. What is the chief weakness of the range?

5. Why are the quartiles and the quartile deviation generally better measures of dispersion than the range?

6. The average deviation and the standard deviation are the two measures of dispersion that are affected by all the values of the variable. Are the average deviations and the standard deviations thus better measures of dispersion than other measures? Why?

7. Why is the standard deviation considered a better measure of dispersion than the average deviation?

8. Explain the advantages of using a measure of relative dispersion.

9. Why do equations (20) and (21) for the normal distribution use the standard deviation instead of some other measure of dispersion?

10. Equation (21) is used to compute the ordinate (Y) of the normal distribution for a given value of x. As the value of x becomes larger, the value of Y becomes smaller. How large must the value of x become to give a value of zero for Y? What is the significance of this?

PROBLEMS

1. Refer to the table in Problem 4, page 68, Chapter 3. If the solution to Problem 4 has not been retained, it will first be necessary to complete step (a) below.

(a) Set up a tally sheet and classify the employees according to age.

(b) Compute the first and third quartiles and the quartile deviation for the distribution. What do these measures show about the dispersion in the ages?

(c) Compute the average deviation from the median for the distribution. (If the solution to Problem 3 on page 190 of Chapter 8 has not been retained, it will first be necessary to compute the median.)

(d) Compute the standard deviation (σ) and the coefficient of variation (V) for the distribution. What is the significance of the difference between the standard deviation and the coefficient of variation for the same distribution?

(e) Compute the measure of skewness from formula (19) on page 208 for the distribution. (If the solution to Problem 3 on page 190 of Chapter 8 has not been retained, it will first be necessary to compute the mean and the mode.)

2. Two students made the same average (arithmetic mean) on the three hour examinations given in a course, but the individual grades on the examinations were as follows:

	STUDENT A	STUDENT B
First examination.....	80	55
Second examination...	74	95
Third examination....	83	87
Arithmetic mean.....	79	79

(a) Summarize briefly the difference in the performance of the two students.

(b) Compute the average deviation from the mean for the two sets of grades. What information do these two measures give with regard to the performance of the two students?

(c) Compute the standard deviation of the two sets of grades. Does the standard deviation give any better description of the performance of the two students than the average deviation?

3. Refer to the table on page 191 which gives the distribution of earnings in dress manufacturing shops in Cleveland and Dallas. Compute the following values for each of the distributions.

(a) First quartile (Q_1).

(b) Third quartile (Q_3).

(c) Quartile deviation (Q).

(d) Average deviation from the median.

(e) Standard deviation (σ).

(f) Coefficient of variation (V).

Part

III STATISTICAL INDUCTION

chapter 10 ♦ Sampling

10 SAMPLING

In collecting information from a large aggregate of individuals, business firms, transactions, or other units, the statistician has the choice of collecting the information from every one of the individual units, referred to as the *population* or *universe*, or of studying a selected portion of the aggregate, referred to as a *sample*. It was pointed out in Chapter 3 that the accountant normally records every transaction, and accounting reports are, therefore, based on a complete compilation of all transactions. Internal statistics are also usually based on a complete enumeration. In Chapter 5 it was pointed out that the traditional method of collecting census data has been to secure the information needed from every individual, farm, store, factory, or mine; but recently more of this type of external data is being collected through the use of samples.

IMPORTANCE OF SAMPLING

The practice of making an estimate of the characteristics of a universe by examining a sample is a very old procedure, although a scientific approach to the problem is of recent origin. The cook who tastes a spoonful of soup in order to form a conclusion regarding the whole kettleful is using sampling procedures. The cotton merchant who buys cotton from a sample cut from only a few places in the bale, or the miller who buys a carload of wheat on the basis of a small amount of the grain extracted from the car, is putting his trust in sampling methods. The earlier uses of sampling assumed that it made little difference how the sample was chosen; and when sampling was limited to material that was uniform in its composition, any sample chosen would give approximately the same answer. A given sample of sea water taken in the middle of the Atlantic would probably give about the same concentration of salt as any other sample taken at the same place. A given group of ten people, however, would probably not show the same percentage of unemployment as every other group of ten people that could be selected. The selecting of a sample from a universe that shows great variation among the constituent elements offers many problems, but it is possible to make satisfactory estimates of the whole from a relatively small sample if the proper methods of selection are used.

Although accountants rely almost entirely on complete records, and official statisticians tend to prefer complete enumerations to sample surveys, the results obtained from the analysis of samples are gaining acceptance at a rapid rate. As better sampling procedures have been developed, the accuracy of the estimates made from samples has improved.

It must be understood that the result obtained from analyzing a sample may never be assumed to give exactly the same result that would be secured if the entire universe were included in the study. However, if the sample is selected in such a manner that it is possible to compute the maximum difference that may be expected to occur between the sample result and the result that would have been secured if the universe had been analyzed, the sample data may be just as useful as a complete enumeration. For many purposes an approximate answer is all that is needed. When such is the case, the estimates from a sample will serve the same purpose as a complete enumeration of the universe and will ordinarily entail less expense.

A manufacturer of building materials who wants to select a magazine in which to advertise his products might consider the percentage of home-owning subscribers to the magazines an important factor in his decision as to which magazine to use. Whether 71, 72, or 73 per cent of the subscribers to a given magazine owned their homes would ordinarily make very little difference in his decision, although it would make a difference if only 30 per cent owned their homes. The degree of approximation that can be tolerated in the data varies with the problem being considered, but it is always important to know to what extent it is an approximation.

It is wasteful to use data that are more precise than are needed for the purpose at hand, since the increase in precision increases the cost of the data without making them any more useful. It is doubtful if a very large proportion of external statistical data used by businessmen is worth the cost of making a complete enumeration, and sample surveys are being substituted to an ever-increasing extent. The use of samples has brought the cost of much data down to a point where the businessman can afford it, whereas he would be forced to make decisions without the information if he had to rely on universe data.

The problem of taking a sample that can be depended upon to possess the characteristics of the universe, and the related problem of computing the maximum amount of sampling error that will be present in a given situation will be taken up in the following pages of this chapter.

SAMPLING ERROR

The extent to which sample results will deviate from the results that would be secured from a study of the universe is called the *precision* of an

estimate from a sample, and its measurement is an important part of a sampling study. In general, only methods of selecting a sample that will permit measuring the precision of estimates from it are considered satisfactory.

There are many kinds of errors that may occur in compiling statistical data, but *sampling error* refers only to those variations that arise from the use of a sample instead of a complete enumeration. Both the sample survey and the complete enumeration might have serious error in them, but the same source of error would be present in both results.

For example, college students fill out numerous forms at the time of registration, and one of the questions asked might deal with hobbies in which the student is interested. Assume that every student registering fills out this section of the form, but the registrar's office tabulates only a sample of the schedules. This sample would be expected to give a somewhat different distribution of interests than if all of the schedules were tabulated. The difference between the results secured from the sample and from the universe (all students' registration forms) would be sampling error. This does not mean that there would not be other kinds of errors in the tabulation. The clerk making the tabulation might make mistakes in compiling either the sample results or the complete study. Because people do not fill out forms with complete accuracy, there will be errors in the information written on the records. These errors will affect the sample results and the universe tabulation in the same way.

SELECTION OF A SAMPLE

When a small portion of the universe is to be studied in order to estimate the characteristics of the universe, it is obvious that a great deal of importance attaches to the selection of the sample items. The first reaction of most people is that typical items should be selected if the sample is to be used to represent the universe, but much experience has demonstrated that it is extremely difficult if not impossible for a person to make a selection of typical items. In fact, specifying "typical" items implies that one knows the answer; otherwise he will not know if a particular individual is typical.

Sometimes attempts have been made to avoid this difficulty by selecting individual items that are known to be typical with respect to certain characteristics, on the assumption that if they are typical in these respects they will be typical with respect to the items being studied. For example, in making a study of buying habits of farmers in a given state, a sample of farms might be selected in such a manner that the average size of farms in the sample would be the same as the average for the state; and the distribution of the sample farms among the different types would parallel

the distribution throughout the state. Since the sample of farms would be typical with respect to these characteristics, it would be assumed that the answers given to questions on buying habits would also be typical of all farmers in the state, and the sample would be considered a perfect cross section of the universe.

While it cannot be stated positively that a sample selected as described above *will not* possess the characteristics of the universe, there is no way of demonstrating that it *will* be representative. The quality of the results can be evaluated only as a judgment of someone who has expert knowledge of the situation; it is not possible to compute objectively the precision of the estimates of the universe. Many business data have been collected by the use of samples of this kind, but the use of the data is limited by the fact that no measure of the precision of the results can be made. Samples selected in this manner are called *purposive* or *judgment* samples.

When the probability of including each individual item in the sample is known, the precision of a sample result can be computed. If a person makes a decision that determines the selection of an individual item, there is no method of knowing the probability of including that item; and therefore the sampling error of judgment samples cannot be computed. If the probability of selection is known, the sampling error can be computed and the sample is a *probability* or *random* sample.

It is worth while emphasizing again the importance of being able to compute a measure of the sampling error. Sample results for which nothing is known of the sampling error, in general, are of little value; but when it can be established that a complete enumeration of the universe would give results that would not fall outside known limits, the data can be used with this limitation in mind. An equally important feature of this situation is the fact that if the sampling error is larger than can be tolerated, it is possible to reduce it as much as is desired, at a cost in time and money. This subject will be further discussed later in the chapter, but for any random sample the precision can normally be increased by increasing the size of the sample. There are other methods of increasing the precision that sometimes can be used without increasing the cost; but if the funds are available, increasing the size of the sample can always be employed to reduce the sampling error.

A SIMPLE RANDOM SAMPLE

In order to demonstrate how a random sample may be used to estimate the characteristics of the universe and how the sampling error may be computed, a small hypothetical universe will be used as a case study. Assume that the earnings in Table 9-1 on page 197 comprise the universe,

and the problem is to estimate the average earnings of the eleven employees from a sample of two. The size of the universe will be represented by N, and the size of the sample by n. In this example $N = 11$ and $n = 2$.

To draw a simple random sample of two from the universe of 11, each amount might be written on a card, and the cards shuffled thoroughly before the two are drawn. There are 55 different samples of 2 that can be drawn from a universe of 11, and when the selection is purely random, any one of these 55 samples is equally likely to be drawn. The 55 samples are listed in Table 10-1 on the following page.

It should, of course, be realized that a universe of this size would probably never be sampled; but a small universe will be much easier to use in demonstrating the principles of sampling than one the size that normally would be sampled.

The smallest estimate of the mean earnings of the universe is the sample consisting of the earnings of Chester and Alexander, which average \$11.30. The largest estimate would be obtained if Yates and Cunningham were drawn, with an average of \$19.25. The arithmetic mean of the 55 sample means is \$14.41, which is the arithmetic mean of the universe (see page 199). Since each of these means has the same probability of being selected, a study of the 55 possible sample means provides some evaluation of the reliability of the estimate furnished by one sample mean.

The basis of this evaluation is a measure of dispersion among the sample means. In other words, how much do the 55 sample means vary from the mean of the universe? In Chapter 9 the various measures of dispersion were described, and it was pointed out that the best measure of dispersion to use in further calculations is the standard deviation. This suggests that the standard deviation of the means of the 55 samples be computed as a measure of the dispersion of these sample means. Treating each of the sample means as a value of the variable (X), the standard deviation can be computed by formulas (14), (15), and (16).

The sum of the squares of the 55 means is 11,629.633750 and the standard deviation is computed as follows:

$$s^2 = \frac{\Sigma X^2}{N} = \frac{11{,}629.633750}{55} = 211.45$$

$$c^2 = \left(\frac{\Sigma X}{N}\right)^2 = \left(\frac{792.5}{55}\right)^2 = 207.65$$

$$\sigma_{\bar{x}}^2 = s^2 - c^2 = 211.45 - 207.65 = 3.80$$

$$\sigma_{\bar{x}} = 1.95$$

Table 10-1

ALL POSSIBLE SAMPLES OF TWO THAT CAN BE DRAWN FROM THE UNIVERSE CONSISTING OF ELEVEN EMPLOYEES

Employees included in sample	Earnings (dollars)	Mean earnings
Chester and Alexander	(11.10 + 11.50)	$11.30
" " Smith	(11.10 + 12.00)	11.55
" " Burnette	(11.10 + 12.20)	11.65
" " Johnson	(11.10 + 12.45)	11.775
" " Davis	(11.10 + 13.50)	12.30
" " Henry	(11.10 + 14.00)	12.55
" " Murphy	(11.10 + 16.00)	13.55
" " Cohen	(11.10 + 17.25)	14.175
" " Yates	(11.10 + 18.50)	14.80
" " Cunningham	(11.10 + 20.00)	15.55
Alexander and Smith	(11.50 + 12.00)	11.75
" " Burnette	(11.50 + 12.20)	11.85
" " Johnson	(11.50 + 12.45)	11.975
" " Davis	(11.50 + 13.50)	12.50
" " Henry	(11.50 + 14.00)	12.75
" " Murphy	(11.50 + 16.00)	13.75
" " Cohen	(11.50 + 17.25)	14.375
" " Yates	(11.50 + 18.50)	15.00
" " Cunningham	(11.50 + 20.00)	15.75
Smith and Burnette	(12.00 + 12.20)	12.10
" " Johnson	(12.00 + 12.45)	12.225
" " Davis	(12.00 + 13.50)	12.75
" " Henry	(12.00 + 14.00)	13.00
" " Murphy	(12.00 + 16.00)	14.00
" " Cohen	(12.00 + 17.25)	14.625
" " Yates	(12.00 + 18.50)	15.25
" " Cunningham	(12.00 + 20.00)	16.00
Burnette and Johnson	(12.20 + 12.45)	12.325
" " Davis	(12.20 + 13.50)	12.85
" " Henry	(12.20 + 14.00)	13.10
Burnette and Murphy	(12.20 + 16.00)	14.10
" " Cohen	(12.20 + 17.25)	14.725
" " Yates	(12.20 + 18.50)	15.35
" " Cunningham	(12.20 + 20.00)	16.10
Johnson and Davis	(12.45 + 13.50)	12.975
" " Henry	(12.45 + 14.00)	13.225
" " Murphy	(12.45 + 16.00)	14.225
" " Cohen	(12.45 + 17.25)	14.85
" " Yates	(12.45 + 18.50)	15.475
" " Cunningham	(12.45 + 20.00)	16.225
Davis and Henry	(13.50 + 14.00)	13.75
" " Murphy	(13.50 + 16.00)	14.75
" " Cohen	(13.50 + 17.25)	15.375
" " Yates	(13.50 + 18.50)	16.00
" " Cunningham	(13.50 + 20.00)	16.75
Henry and Murphy	(14.00 + 16.00)	15.00
" " Cohen	(14.00 + 17.25)	15.625
" " Yates	(14.00 + 18.50)	16.25
" " Cunningham	(14.00 + 20.00)	17.00
Murphy and Cohen	(16.00 + 17.25)	16.625
" " Yates	(16.00 + 18.50)	17.25
" " Cunningham	(16.00 + 20.00)	18.00
Cohen and Yates	(17.25 + 18.50)	17.875
" " Cunningham	(17.25 + 20.00)	18.625
Yates and Cunningham	(18.50 + 20.00)	19.25
Total		$792.5

SOURCE: Table 9-1.

The standard deviation of the means of all possible samples ($\sigma_{\bar{x}}$) is distinguished from the standard deviation of the universe ($\hat{\sigma}$) by the use of a subscript to the Greek letter sigma. The term is called the *standard error of the mean*, and is the measure of precision of a sample. When a standard deviation is calculated from a group of computed values, such as means or percentages, it is generally called the "standard error." The term "standard deviation" is used when the computation is based on individual measurements.

CONFIDENCE INTERVALS

The standard deviation of the distribution of arithmetic means (standard error of the mean) summarizes the variation in means of the samples that can be drawn from the universe. It was pointed out in Chapter 9 that for a normal distribution approximately 68 per cent of the items fall within the range of two standard deviations, plus or minus 1 σ from the mean. Under certain conditions the distribution of means of all possible samples that can be drawn from a given universe tend to be normally distributed, which indicates that 68 per cent of these means will fall within a range of plus or minus 1 σ from the mean of all the sample means (which is the mean of the universe). In other words, the probability is 68 out of 100 that the mean of a sample drawn at random will fall within plus or minus 1 σ from the mean of the universe which the sample was drawn to estimate.

Referring back to the example of the 55 sample means, the standard deviation of the sample means was computed to be \$1.95. According to the statement in the preceding paragraph, approximately 68 per cent of the 55 sample means should fall within the range \$14.41 $\pm$ \$1.95 (that is, between \$12.46 and \$16.36). In other words, 37 of the sample means (68 per cent of 55 is 37) should fall within $\pm$ 1 σ of the mean of the universe if they were normally distributed. A count of the means given on page 220 shows that actually 36 of the means fall within this range. This range is called the *confidence interval* and about 68 per cent of the time a sample mean will fall within this interval. This permits us to assume that if we have a sample mean and can determine its confidence interval, we can accept the probability of 68 out of 100 that the universe mean will fall within the confidence interval.

In Chapter 9 it was stated that 95.45 per cent of the items in a normal distribution will fall within a range of $\pm$ 2 σ from the mean. For the 55 sample means this should be 52 (95.45 per cent of 55 is 52); a count of the means shows that actually 53 of them fall within this range. Increasing the width of the confidence interval raises the probability that a sample will not fall outside it.

Within a range of $\pm 3\sigma$ from the mean, practically all the items of a normal distribution fall; and all the 55 means are within this confidence interval.

The percentages given above are correct when the distribution is normal, and the distribution of the means of random samples is approximately normal when the sample is reasonably large. *This is true even though the universe from which the samples are taken is not normally distributed.* Samples with $n = 2$ will not be distributed normally, but samples of this size would not be used to estimate a universe value. The very small sample was used in this example simply to make the job of demonstrating the behavior of samples easier.

When the size of the sample taken from a given universe is increased, the variation in the means becomes less. The reason for this is that the larger the number of items in the sample, the more opportunity is offered the extremely large and small items drawn to offset each other. This results in the sample means being more nearly the same than when a smaller sample was taken and there was less chance for this offsetting to take place.

It would take a considerable amount of space to give in detail all the samples of $n = 3$ to $n = 10$, and it is doubtful if it would be worth while. All that is shown below is a table giving the standard deviation of each of the sample distributions that would result from taking all possible samples of different sizes. These standard deviations give proof that the sample means show less and less variation as the size of the sample is increased.

Table 10-2

STANDARD DEVIATIONS OF THE DISTRIBUTIONS OF MEANS OF ALL POSSIBLE SAMPLES OF DIFFERENT SIZES THAT CAN BE DRAWN FROM THE UNIVERSE CONSISTING OF ELEVEN EMPLOYEES

Size of sample	Standard deviation of the distribution of means
2	1.95
3	1.49
4	1.22
5	1.01
6	.83
7	.69
8	.56
9	.43
10	.29

SOURCE: Table 9-1.

COMPUTATION OF $\sigma_{\bar{x}}$

The previous discussion aimed at demonstrating that it is possible to measure the sampling error, but it is obvious that the method described could not be used in a practical case. Taking all possible samples from even a small universe would be absurd, since it would be so much simpler to make a complete enumeration of the universe. The method was used simply to demonstrate that the dispersion in the sample means can be used to measure the reliability of estimates from samples.

The value of the standard deviation of all possible sample means (the σ of the mean) can be computed by the formula:

$\sigma_{\bar{x}}$ represents the standard deviation of the distribution of means with samples of size "n."

$\hat{\sigma}$ is the standard deviation of the universe.

N is the size of the universe.

n is the size of the sample.

$$\sigma_{\bar{x}} = \hat{\sigma}\sqrt{\frac{N-n}{n(N-1)}} \quad \ldots\ldots \quad (22)$$

or

$$\sigma_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \ldots\ldots \quad (23)$$

The data on the earnings of the 11 employees were analyzed in Chapter 9, where the standard deviation was found to be \$2.91. The size of the universe was 11, and the size of the sample was 2; so $N = 11$ and $n = 2$. The following calculations give a value of $\sigma_{\bar{x}}$ equal to \$1.95, which is the same value computed on page 219 for the standard deviation of the distribution of means.

$$\sigma_{\bar{x}} = 2.91\sqrt{\frac{11-2}{2(11-1)}} = 2.91\sqrt{\frac{9}{20}} = 2.91 \times .6708 = 1.95$$

When we know the standard deviation of the universe from which a sample is drawn, it is a simple matter to compute the standard deviation of the sample means ($\sigma_{\bar{x}}$) by formula (22) or (23). At this point, however, the question inevitably arises as to how one can measure the standard deviation of the universe with which he is dealing. If the problem is to estimate the arithmetic mean of the universe, it is obvious that the mean is unknown, and if this is true it does not seem likely that the standard deviation of that universe will be known.

The answer to this question is that the standard deviation of the universe can be estimated from the same sample that was used to estimate the mean of the universe. It seems reasonable to assume that if the sample will give a satisfactory estimate of the universe mean it should give an equally good estimate of the universe standard deviation. For samples that are reasonably large, say over 100, this is generally true, except for a very abnormal universe with extreme or unusual items.

This means that a sample can be used to estimate a universe characteristic (such as the arithmetic mean) and also to compute a measure of the precision of this estimate. This is the great advantage of a random sample; other types of samples do not supply the necessary information from which a measure of the precision of the result can be computed.

BIAS IN THE SAMPLE

The standard deviation of a sample is not an unbiased estimate of the standard deviation of the universe, and when the sample is small the bias is large enough to need correction. An unbiased method of estimation is one in which the mean of all values computed from all possible samples is the same as the value computed from the universe. The amount of the bias can always be computed, and whenever it is large enough to be important, an adjustment can be made. The value of the standard deviation computed from a sample must be multiplied by $\sqrt{\frac{n}{n-1}}$ to secure an unbiased estimate of the standard deviation of the universe. If the estimated standard deviation of the universe is represented by σ' and the standard deviation of the sample by σ, the formula for estimating the universe standard deviation from a sample is:

$$\sigma' = \sqrt{\sigma^2 \frac{n}{n-1}} \quad \text{(24)}$$

or, written in a different form

$$\sigma' = \sigma\sqrt{\frac{n}{n-1}} \quad \text{(25)}$$

The term $\sqrt{\frac{n}{n-1}}$ is so close to unity for a large sample that it is generally ignored. For example, when $n = 500$, $\frac{n}{n-1} = \frac{500}{499} = 1.002$. Multiplying σ^2 by 1.002 makes such a little change that it is not worth while. On the other hand, for a sample in which $n = 30$, $\frac{n}{n-1} = \frac{30}{29} = 1.034$. Multiplying σ^2 by 1.034 would increase this value by 3.4 per cent, and a correction of this size would probably be worth making.

Although the correction is not important for large samples, it is important in the analysis of small samples; and it is wise always to remember that the sample standard deviation is always a biased estimate of the universe standard deviation, even though the bias may be so small that it can be ignored. In other words, much of the time σ is approximately equal to σ'. Very rarely is $\hat{\sigma}$ (the true value of the universe standard deviation) known.

If it is known that the standard deviation of the sample will be used as an estimate of the universe standard deviation, the correction can be made at the time the standard deviation is computed. If this is done, the formula is:

$$\sigma' = \sqrt{\frac{\Sigma x^2}{n-1}} \quad \ldots\ldots\ldots\ldots \quad (26)^*$$

Since $\hat{\sigma}$ (the standard deviation of the universe) is seldom known, the following formula is most commonly used to compute the standard error of the mean ($\sigma_{\bar{x}}$).

$$\sigma_{\bar{x}} = \frac{\sigma'}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \ldots\ldots\ldots \quad (27)$$

This formula differs from (23) in that it uses σ' (the universe standard deviation estimated from the sample) instead of $\hat{\sigma}$ (the true universe standard deviation).

IMPORTANCE OF THE SIZE OF THE UNIVERSE

When persons unfamiliar with sampling first consider the question of the influence of the size of the universe on the precision of sampling results, the intuitive answer seems to be that a sample of a given size, say 1,000, will give a much more precise result when drawn from a universe of 100,000 than when drawn from 800,000. In the first case the sample is one out of every 100 items in the universe, while in the second universe the sample is one out of every 800 items. It appears obvious that the first sample will be more precise than the second. However, if the dispersion in the two universes, as measured by their standard deviations, is the same, the precision of the sample result from a sample of 1,000 will be practically the same for the larger universe as for the smaller.

Formula (27) for the standard error of the mean can be used to demonstrate that the above statement is true. Assume that 8,295 was the average number of miles driven last year by a sample of 1,000 car owners from a universe of 100,000, and that the standard deviation of the universe was

*This formula is derived as follows:

$$\sigma' = \sqrt{\sigma^2 \frac{n}{n-1}} = \sqrt{\frac{\Sigma x^2}{n}\frac{n}{n-1}} = \sqrt{\frac{\Sigma x^2}{n-1}}$$

$x = X - \bar{X}$ when X represents the values of the individual items and $\bar{X}$ represents the arithmetic mean of the X values.

estimated from the sample study to be 3,100 miles. The standard error of the estimate of the average number of miles driven by the 100,000 car owners in the universe would be

$$\sigma_{\bar{x}} = \frac{\sigma'}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3{,}100}{\sqrt{1{,}000}}\sqrt{\frac{100{,}000-1{,}000}{100{,}000-1}} = 98.03\sqrt{.990} = 98.03 \times .995 = 97.5$$

If the sample of 1,000 had been taken from a universe of 800,000 car owners, the standard error of the sample mean would be

$$\sigma_{\bar{x}} = 98.03\sqrt{\frac{800{,}000-1{,}000}{800{,}000-1}} = 98.03\sqrt{.998751} = 98.03 \times .999 = 97.9$$

For all practical purposes the sampling error is the same for the two samples since the second term in the equation is approximately unity in each case. Whenever the size of the sample is small in relation to the size of the universe, this term will be close to unity and can be ignored in computing the standard error. When the sampling fraction, $\frac{n}{N}$, does not exceed 5 per cent, it is customary to ignore the second term in the above formula. Sometimes it is ignored when the sampling fraction is as high as 10 per cent, since it results in overstating slightly the standard error. This means that the precision of the estimate from a sample is somewhat greater than shown by the computed standard error.

The following example will be used to illustrate the estimation of a universe mean from a sample, and the determination of the precision of this estimate. Assume that you are working in the personnel office of a company employing 30,000 people and you are asked to make an estimate of the average age of all the employees of the company. You have a complete set of personnel records, from which you could take the age of every employee to use in computing the average age. However, the information is wanted immediately, and you are told that an approximation will be satisfactory. You select a sample of 282 employees at random from the files and make the following frequency distribution of the ages of the 282 employees. (A discussion of the problems of taking a random sample is given later in this chapter; it is sufficient for this illustration to concede that the sample was drawn at random.) The arithmetic mean of this sample of 282 employees is 39.6 years and the standard deviation is 9.6 years. These values were computed in Table 10-3 by the methods given in Chapters 8 and 9.

The mean of 39.6 years is the best available estimate of the mean of the 30,000 employees and you report this figure in less than an hour (the time required to take the sample of 282 personnel folders, tally the frequency distribution, and compute the arithmetic mean).

Table 10-3

COMPUTATION OF MEAN AND STANDARD DEVIATION

AGES OF 282 EMPLOYEES

Ages (Years)	Number f	d'	fd'	$f(d')^2$
20 and under 25	7	− 4	− 28	112
25 and under 30	34	− 3	− 102	306
30 and under 35	65	− 2	− 130	260
35 and under 40	56	− 1	− 56	56
40 and under 45	47	0	0	0
45 and under 50	25	1	25	25
50 and under 55	22	2	44	88
55 and under 60	18	3	54	162
60 and under 65	8	4	32	128
Total.....	282	...	− 161	1,137

SOURCE: Hypothetical data.

$$\bar{X} = 42.5 + \frac{-161}{282}5 = 42.5 - 2.9 = 39.6 \text{ years}$$

$$s^2 = \frac{1,137}{282} = 4.0319$$

$$c^2 = \left(\frac{-161}{282}\right)^2 = .3260$$

$$\sigma^2 = s^2 - c^2 = 4.0319 - .3260 = 3.7059$$

$\sigma = 1.925$ (class interval units)

Class interval (i) = 5 years

$\sigma = 1.925i = 1.925 \times 5 = 9.6$ years

It is also possible to compute the precision of this estimate by using formula (27). The value of $\hat{\sigma}$ (the universe standard deviation) is not known, but σ' can be estimated from the standard deviation of the sample as follows:

$$\sigma' = \sigma\sqrt{\frac{n}{n-1}} = 9.6\sqrt{\frac{282}{281}} = 9.6\ \sqrt{1.00\overline{36}} = 9.6 \times 1.002 = 9.6 \text{ (rounded to one decimal place)}$$

The result of the correction for the bias in the standard deviation of a sample is so slight that it does not change the result when carried to one decimal place and σ' is not enough different from σ to justify its computation. In practice, this correction would be ignored for a sample of this size since it has such a slight effect on the estimate, although for small samples it may be important.

Using formula (27) and the values already computed, the process of solving for the standard error of the mean is shown below:

$$\sigma_{\bar{x}} = \frac{9.6}{\sqrt{282}}\sqrt{\frac{30{,}000 - 282}{30{,}000 - 1}} = \frac{9.6}{16.79}\sqrt{\frac{29{,}718}{29{,}999}} = .57 \times .995 = .57$$

In this calculation the value of $\sqrt{\frac{N-n}{N-1}}$ is so near unity that it also would in practice be ignored. Its use does no harm, but when the sampling fraction is as small as it is in this example, nothing is gained in accuracy by using it.

The estimate of the mean age of all employees of the company from the sample of 282 did not have a large amount of error due to the use of the sample. The standard error of .57 years gives a measure of the precision of the estimate from the sample. The probability is very high (it is practically certain) that the mean of the universe would fall within the confidence interval 39.6 ± 3 (.57), or within the range 37.9 to 41.3 years. The probability is 95 out of 100 that the universe mean falls within the confidence interval 39.6 ± 2 (.57), or within the range 38.5 to 40.7. It would be quite appropriate to inform the persons using this estimate of the mean age of all employees that if the time were taken to average the 30,000 ages there is very little chance that the result would fall outside the interval 37.9 to 41.3.

This example is typical of the use of sample data to estimate values in the universe. A simple random sample such as the one described above will give an unbiased estimate of the average age of all employees in the universe, and also will provide the data needed to compute the precision of the estimate. If a more precise estimate is needed, it can be secured by increasing the size of the sample. If the above results had been secured from a sample of 1,128, the sampling error would have been considerably less. Ignoring the correction for the size of the universe, the following calculation using formula (27) gives the sampling error:

$$\sigma_{\bar{x}} = \frac{9.6}{\sqrt{1{,}128}} = \frac{9.6}{33.6} = .286$$

The confidence interval is cut in half when the size of the sample is quadrupled. Increasing the precision of a sample result thus becomes more and more costly, and there is no good reason for using data that are more precise than is necessary.

STANDARD ERROR OF A PERCENTAGE

The examples of estimates from samples have all been estimates of the arithmetic mean of the universe, but probably equally important to the business statistician is the estimate of the percentage of units in the universe

that possess a given characteristic. The percentage of consumers in a given market that prefer a product, the percentage of subscribers to a magazine that own their homes, and the percentage of voters that will support a given candidate are all examples of data that belong in this category.

The universe of eleven employees used to illustrate the estimate of the universe mean can be used to illustrate the estimating of the universe percentage from a sample. Of the eleven employees in the universe, five owned their homes and six lived in rented homes. If we wanted to estimate the percentage of homeowners in the universe from a sample of two employees, we could draw 55 combinations of two from the universe, just as in the example where we were estimating the average earnings of this group.

Table 10-4

ALL POSSIBLE SAMPLES OF TWO THAT CAN BE DRAWN FROM THE UNIVERSE CONSISTING OF ELEVEN EMPLOYEES

Employees included in sample	Per cent owning homes	Employees included in sample	Per cent owning homes
Chester and Alexander	100	Burnette and Murphy	50
" " Smith	100	" " Cohen	50
" " Burnette	100	" " Yates	50
" " Johnson	100	" " Cunningham	50
" " Davis	50		
" " Henry	50	Johnson and Davis	50
" " Murphy	50	" " Henry	50
" " Cohen	50	" " Murphy	50
" " Yates	50	" " Cohen	50
" " Cunningham	50	" " Yates	50
		" " Cunningham	50
Alexander and Smith	100		
" " Burnette	100	Davis and Henry	0
" " Johnson	100	" " Murphy	0
" " Davis	50	" " Cohen	0
" " Henry	50	" " Yates	0
" " Murphy	50	" " Cunningham	0
" " Cohen	50		
" " Yates	50	Henry and Murphy	0
" " Cunningham	50	" " Cohen	0
		" " Yates	0
Smith and Burnette	100	" " Cunningham	0
" " Johnson	100		
" " Davis	50	Murphy and Cohen	0
" " Henry	50	" " Yates	0
" " Murphy	50	" " Cunningham	0
" " Cohen	50		
" " Yates	50	Cohen and Yates	0
" " Cunningham	50	" " Cunningham	0
Burnette and Johnson	100	Yates and Cunningham	0
" " Davis	50		
" " Henry	50	Total.....	2,500
		Mean.....	45 5/11

SOURCE: Hypothetical data.

Assume that Chester, Alexander, Smith, Burnette, and Johnson own the houses they live in, and the remaining six are tenants. A sample of two individuals will show either zero, 50, or 100 per cent home ownership. All of the possible 55 sample percentages are shown in Table 10-4. For 15 of the samples the percentage of homeowners is zero; for 30 samples the percentage of homeowners is 50; for 10 samples the percentage of homeowners is 100. The mean of the 55 percentages is 45 5/11 per cent, which is the correct value for the universe (5 out of 11).

The dispersion in the sample percentages is computed in Table 10-5 in exactly the same manner as the dispersion in the sample means was computed for Table 10-3. The deviation of each percentage from the universe percentage (45 5/11) is computed, the deviations squared, totaled, divided by N (55), and the square root taken. The computation uses formula (12) with x representing the deviation of each percentage from the average percentage for the universe.

Table 10-5

COMPUTATION OF STANDARD DEVIATION: ALL POSSIBLE SAMPLE PERCENTAGES

Per cent owning homes	No. of samples f	Dev. from mean (45 5/11) x	x^2	fx^2
0	15	−45 5/11	2,066.11	30,991.7
50	30	4 6/11	20.66	619.8
100	10	54 6/11	2,975.21	29,752.1
Total	55	...	...	61,363.6

SOURCE: Table 10-4.

$$\sigma_p = \sqrt{\frac{61,363.6}{55}} = \sqrt{1,115.70} = 33.4$$

The standard deviation of the percentages resulting from taking all possible samples from the universe measures the scatter in the sample values, and gives an indication of the precision of an estimate made from one sample. The value of the standard deviation computed from the sample percentages is 33.4 per cent, and since approximately 68 out of 100 of the items in a normal distribution fall within $\pm 1\sigma$ from the mean, 37 of the sample percentages should fall within the confidence interval 45 5/11 $\pm$ 33.4. A count of sample percentages shows that 30 fall within the confidence interval. In the same manner the 2σ and 3σ confidence intervals can be computed, just as with the distribution of means. In this case all the percentages fall within 45 5/11 $+2\sigma$.

It was shown on page 223 that the standard deviation of the distribution of means can be computed without the laborious job of taking all possible samples from the universe, and the same can be done for the distribution of percentages. The formula for the standard deviation of all possible sample percentages from a universe with a percentage P is

$$\sigma_p = \sqrt{\frac{PQ}{n}\frac{N-n}{N-1}} \qquad (28)$$
$$Q = 100 - P$$

When this formula is used, the following value is found for the standard error of the estimate of the percentage of homeowners in the universe of 11 ·

$$\sigma_p = \sqrt{\frac{\left(45\frac{5}{11}\right)\left(54\frac{6}{11}\right)}{2}\,\frac{11-2}{11-1}} = \sqrt{\frac{2{,}479.34}{2}\left(.9\right)} = \sqrt{1{,}115.70}$$

$$\sigma_p = 33.4$$

The value of σ_p is the same as that computed on page 230, the standard deviation of all possible sample percentages. If the value of P for the universe is not known, the value derived from the sample may be used as an estimate, just as the value of σ computed from the sample was used as an estimate of the universe standard deviation in computing $\sigma_{\bar{x}}$.

DETERMINING THE SIZE OF SAMPLE NEEDED

The preceding discussion has dealt entirely with the problem of determining the precision of the results from a sample, but in actual practice an equally important procedure is determining how large a sample is needed before a study is made. When a preliminary analysis of a problem at hand indicates that information is needed, almost the first question that will be asked is how much will this information cost. If there has been any previous experience with collecting data, it will be possible to estimate approximately the cost per interview; but it will also be necessary to know how many interviews will be needed. If it appears that the cost of making an interview and tabulating the results will be approximately $5 per schedule, and if the precision desired in the study will require a sample of 2,000, the total cost of the study will be approximately $10,000. If the management is not willing to spend this much for the information, it is desirable to find this out before any money is spent. This means that it is extremely important to estimate at least approximately how large a sample is needed before any work is done on the study.

A large dairy was considering the advisability of selling milk in paper cartons through grocery stores; but before going further with their consideration of the idea, the manager wanted to know how well consumers liked this type of container. It was suggested that a small consumer survey might give an approximate answer to the question of how well consumers would like this form of container, and it was decided to explore the cost of making such a survey.

Before any answer could be given to this question, it was necessary to determine (1) how much it would cost per interview to secure the information, and (2) how many interviews would be needed. In making an estimate of the size of the sample needed, it was first necessary to secure agreement on the amount of sampling error that could be tolerated in the results. This decision must always be made, and can be made only by the persons that will make use of the data. The statistician planning the survey generally will discuss the required accuracy with a user of the information in order to find out the degree of precision needed in the data.

The manager of the dairy might decide that, if the survey showed correctly within a range of 6 either way the percentage of people that like paper milk cartons, the data would be satisfactory. This would represent a confidence interval of ± 6, although this statement would not indicate the probability level; if one wanted to be practically certain that the universe percentage would fall within this range, it would be well to use the 3σ level. If this level of probability were used, it follows that $3\sigma = 6$ and $\sigma = 2$.

Ignoring the correction for the size of the universe, the formula for the standard error of a percentage is

$$\sigma_p = \sqrt{\frac{PQ}{n}} \quad \text{. . . . . (29)}$$

If the sampling fraction is small, the correction factor can be ignored without having any appreciable effect on the result. If we have an approximation of the value of P, it will be possible to substitute values for P, Q, and σ_p, leaving only the value of n unknown. If the intention is to solve the equation for n, it is more convenient to use this equation:

$$n = \frac{PQ}{\sigma_p^2} \quad \text{. . . . . (30)}$$

If it were believed that the percentage of persons liking paper milk cartons would be approximately 70, this value could be substituted for P, and a calculation made for the size of n needed.

$P = 70$
$Q = 100 - 70 = 30$
$\sigma_p = 2$

$$n = \frac{(70)\,(30)}{2^2} = \frac{2{,}100}{4} = 525$$

If the value of P were thought to be approximately 60, the size of the sample needed would be somewhat larger, calculated as follows:

$$P = 60 \qquad Q = 100 - 60 = 40 \qquad \sigma_p = 2 \qquad n = \frac{(60)\ (40)}{2^2} = \frac{2{,}400}{4} = 600$$

If it is impossible to make an approximation of the size of the percentage, it is customary to assign P a value of 50, since this value requires the largest sample for a given degree of reliability. The sample size required for the confidence interval when $P = 50$ is computed as follows:

$$P = 50 \qquad Q = 100 - 50 = 50 \qquad \sigma_p = 2 \qquad n = \frac{(50)\ (50)}{2^2} = \frac{2{,}500}{4} = 625$$

The sample of 625, required when $P = 50$, is the largest sample that would be needed for a confidence interval of ± 6. For any other value of P, the value of PQ would be less than 2,500, which means that the sample size needed would be slightly less. The general practice is to use a sample large enough to give the required precision if $P = 50$, even though there is reason to believe the value of P will differ considerably from 50. If this results in taking a somewhat larger sample than is needed, it will normally not increase the cost a great deal and will have the advantage of never providing a smaller degree of precision than was intended.

If the problem is to estimate the size of the sample needed to determine the mean value of some characteristic, such as the average distance traveled by farmers to buy shopping goods, formula (27) is used, usually after solving for n.

$$n = \frac{\sigma^2}{\sigma_{\bar{x}}^2} \qquad \text{(31)}$$

Formula (31) ignores the correction for the size of the universe since this factor usually has a very slight effect on the result.

The problem of estimating the size of sample needed is the same as discussed above for a percentage. First, a decision must be made as to the sampling error that can be tolerated; and second, an estimate must be made of the standard deviation in the universe. The latter estimate is not so easy to make for this situation as it is for the percentage. The maximum sample is needed when $P = 50$, but there is no way of knowing what the maximum standard deviation will be. One approach is to try to estimate the amount of dispersion from a previous study; but if this cannot be done, it may be necessary to make a preliminary study and estimate the size of the

standard deviation from these data. A small sample would not give a very accurate estimate of the standard deviation, but it would be better than a guess and would at least give an approximation of the size of the sample that would be needed.

A study was made by a market research organization to learn something about the buying habits of farmers in the Middle West with the object of finding out the distance farmers traveled to buy various kinds of goods, and one problem in planning the survey was the determination of the size of sample needed. A number of studies similar to this had been made by market research organizations, and on the basis of the information in these surveys it seemed reasonable to assume that the average distance traveled would not be more than 40 miles and the standard deviation not more than 20 miles. Although these amounts were known to be not much more than guesses, it was decided from this information to make a preliminary estimate of the size of the sample needed.

The executives who were to use the data wanted to have a rather small amount of sampling error, and finally it was agreed that a confidence interval of 1.2 miles should be set up as an objective. Assuming that this was at the 3σ level, the required standard error of the mean was 0.4 miles.

$$3\sigma_{\bar{x}} = 1.2 \text{ miles} \qquad \sigma_{\bar{x}} = 0.4 \text{ miles}$$

The size of the sample that would be required was then computed as follows:

$$n = \frac{\sigma'^2}{\sigma_{\bar{x}}^2} = \frac{20^2}{0.4^2} = \frac{400}{.16} = 2{,}500$$

When the study was made, the standard deviation of the universe was found to be 14.3 miles instead of 20 miles, so the sampling error was less than the specifications required. The sample that was finally secured was 2,467, and the precision of the sample result was as follows:

$$\sigma_{\bar{x}} = \frac{14.3}{\sqrt{2{,}467}} = .29 \text{ miles}$$

This was somewhat smaller than the 0.4 miles specified for the study. The more accurately the standard deviation of the universe can be predicted, the more accurately the size of a sample needed can be determined. If a preliminary study had been made and the size of the universe estimated from a small sample, it usually would be more accurate than the above example. However, if the decision to make a study depends upon the size of sample needed, there may not be any funds available to take a preliminary sample. In such a case it becomes necessary to make a rough approximation, even though this be merely a guess.

SAMPLE DESIGN

During the discussion in this chapter it has been assumed that samples were selected in such a manner that each item in the universe had an equal chance of being included in the sample, but no specific methods of insuring this were described. It was considered important to establish first the methods of using random samples, but the discussion now will be concerned with the problem of selecting a sample in a manner that will fulfill the conditions stated above.

On page 219 it was suggested that a random sample from the universe of 11 employees might be drawn by writing each item on a card, and shuffling the cards thoroughly before drawing each name. Mechanical methods of mixing the items in the universe thoroughly are cumbersome and often fail to secure complete randomness. For example, if the universe consists of the names of 30,000 employees in the personnel files, there is no practicable method of mixing them thoroughly enough to permit a random sample to be drawn, even if the director of personnel would permit his files to be treated in such a manner.

Undoubtedly the best method of taking a random sample from a list is the use of *random numbers*, which consist of long series of digits from 0 to 9. Each digit occurs approximately the same number of times, and the order is one that statisticians have accepted as random. A portion of a table of random numbers is illustrated in Fig. 10-A. If the universe consists of 9 or

13284	16834	74151	92027	24670	36665	00770
21224	00370	30420	03883	94648	89428	41583
99052	47887	81085	64933	66279	80432	65793
00199	50993	98603	38452	87890	94624	69721
60578	06483	28733	37867	07936	98710	98539
91240	18312	17441	01929	18163	69201	31211
97458	14229	12063	59611	32249	90466	33216
35249	38646	34475	72417	60514	69257	12489
38980	46600	11759	11900	46743	27860	77940
10750	52745	38749	87365	58959	53731	89295
36247	27850	73958	20673	37800	63835	71051
70994	66986	99744	72438	01174	42159	11392
99638	94702	11463	18148	81386	80431	90628
72055	15774	43857	99805	10419	76939	25993
24038	65541	85788	55835	38835	59399	13790
74976	14631	35908	28221	39470	91548	12854
35553	71628	70189	26436	63407	91178	90348
35676	12797	51434	82976	42010	26344	92920
74815	67523	72985	23183	02446	63594	98924
45246	88048	65173	50989	91060	89894	36036

Bureau of Transport Economics and Statistics, Interstate Commerce Commission

Fig. 10-A **Random Numbers**

less items, one column of numbers is used; if the universe consists of 10 to 99 items, two columns are used, etc. The numbers are read consecutively from the table; if a number is repeated, it is not used a second time; if a number larger than the largest number in the universe is selected, it is ignored. A sample of items selected in this manner will be random; that is, each item in the universe will have the same probability of being selected.

Instead of taking a random sample from the list, a *systematic sample* may be selected. If the sample is to consist of one per cent of the universe, one item would be selected at random from the first 100 items on the list, and then every 100th item throughout the remainder of the list. A similar procedure would be followed for any sampling fraction that was desired.

A systematic sample is approximately the same as a simple random sample when the items in the universe are in random order. An example would be a file of names in alphabetical order, if the item being studied bore no relationship to the names of the individuals. There are some dangers in using a systematic sample, but it is a widely used method of selecting a sample and in general is considered simple and foolproof.[1]

CLUSTER SAMPLING

In sampling business and human universes, securing a complete list of all the items in the universe is often either impossible or very expensive. Many universes exist in the form of lists, such as telephone directories, city directories, members of trade associations, employees, students in a school, charge customers, and subscribers to a magazine or newspaper. However, for many universes no satisfactory list exists, and compiling one would be too expensive. Sometimes when it is impossible to secure a list, or it is too expensive to compile one for the study, *cluster sampling* may be used.

In making a consumer survey in a city, the sampling units might be families; and it might be intended that the survey consist of a random sample of all families in the city. Assume that a recent listing of families was not available (such as an up-to-date city directory). It would be possible to make a listing of all the families in the city, and from this list draw a sample of families to be interviewed; but obviously this would be very costly. A cheaper method would be to divide a map of the city into a number of small areas; these areas might be large, consisting of several blocks, or they might be small, consisting of single blocks or even portions of blocks. Each area would be given a number and from this list of areas a random sample or a systematic sample might be drawn.

[1]William Edwards Deming, *Some Theory of Sampling* (New York: John Wiley & Sons, Inc., 1950), p. 83.

Since the family represented the unit to be interviewed, the next step would be to interview all of the families in each area selected, or choose either a random or a systematic sample of families from each area. Since every part of the city would have a chance of being selected, the sample of families would be random. This type of cluster sample is called an *area sample*, and it can be used in the absence of a complete list of the individual units. If the universe can be broken into geographic units that include all of the universe, a random sample of these geographic units can be selected. By selecting a small enough sample of areal units, it is always possible to list all of the individual units in these areas and either interview all of these individual units or select a sample of units from the listing. In this manner it is not necessary to list all of the units in the universe, but only those that fall in the selected areas.

The sampling error for a sample selected as described above will generally not be measured correctly by the formulas previously given. If the individual units in the sampling areas tend to resemble each other, the sampling error will be greater than would be secured from a simple random sample from all of the individual units in the universe. There is a tendency for the individuals in the small areas to be alike; and the greater this tendency, the more it increases the sampling error of the results from the sample. If the individual units in the areas show no greater resemblance than for the whole universe, a cluster sample will give as precise an estimate as a simple random sample, but it is usually found that people living in the same neighborhood tend to resemble each other, at least to some extent; so the precision of a cluster sample tends to be lower than for a simple random sample.

The discussion earlier in the chapter demonstrated that the sampling error can be reduced by increasing the size of the sample, so the fact that a cluster sample loses precision may not be serious. By increasing the size of the sample, this decline in precision can be offset. In many cases it costs much less to take a considerably larger cluster sample than it would cost to take the smaller simple random sample with the same precision. The cluster sample would save in the listing of the units in the universe, and it would also be possible to make the interviews at a lower cost per interview, since they were grouped into clusters. A simple random sample would be scattered over all portions of the universe, with the result that interview costs would generally be high. The cluster sample, on the other hand, would concentrate the interviews in a limited number of areas, with a resulting saving in cost that often would more than compensate for the larger number of interviews that would be needed. In fact, the saving in the cost of interviews is so great that area samples are used even when a complete list of the individual units in the universe is available.

The problem of the statistician in designing a sample is to secure the maximum precision per dollar of cost. Even though a given sample design might double the number of interviews that were needed for the sample, the fact that the larger number of interviews could be made at a smaller cost would make this design preferred. The most efficient design in any given case is the one that secures the desired precision at the lowest cost.

The problems of designing the most efficient sample may be extremely complex and no attempt will be made to do more in this text than to point out that they exist. Great advances have been made in recent years in increasing the efficiency of sample designs and this field of statistical work represents a highly specialized activity.

TESTING HYPOTHESES

The discussion of sampling up to this point has covered the problem of estimating certain characteristics of the universe from which the sample was drawn. It was pointed out that estimates of values for the universe are subject to error but that limits may be determined within which, with a given degree of confidence, the universe values may be expected to fall.

It is important to recognize that the estimates from samples are only approximate and this fact must be taken into consideration in making decisions based on such estimates. If a computation is made of the amount of overtime worked by all employees in two departments in a certain month, the data for the two departments can be considered as completely accurate. If employees in Department A worked an average of 2 hours and 15 minutes overtime, and those in Department B worked an average of 2 hours and 13 minutes, the conclusion that Department A worked slightly more overtime per employee than Department B is definitely established. The difference is very small; but if we assume that the records were checked sufficiently to be certain they were accurate, there is no question but that there was a difference between the two departments.

If the two averages were based on samples, however, both of them would be approximate and it would not be immediately clear whether the difference between the two averages was a real difference or was due to the approximate nature of the results. In such a situation the analysis of sample data introduces a problem not present in the use of results based on complete enumerations. In using data based on complete enumerations, there is, of course, always the possibility that different kinds of mistakes have been made in compiling the data; but if careful work has been done, there is usually no reason why such mistakes cannot be eliminated. This is not the case with respect to sampling error, since it is inevitable that a variation will exist between the results of a complete enumeration and a sample

study, even though the same care has been taken to insure accuracy in the collection and analysis of the data.

Since a properly designed sample study provides the confidence interval of the estimate, there is enough information available to permit drawing conclusions even though the information is only approximate. The following example may be used to illustrate how this can be done.

The advertising sales department of a farm paper believed that the farmers who subscribed to the paper had a higher average income than all farmers in the area; but, in using this information to sell advertising, it was necessary to have some facts to support the statement. The circulation department of the farm paper took a random sample of 3,600 names from the circulation list in one state and contracted with the Bureau of the Census to make a special tabulation of the census schedules for these farmers. (The Bureau of the Census will make such special tabulations when it can be done without revealing any confidential information from the individual schedules.)

The tabulation of the 3,600 schedules made possible the comparison of the subscribers to the farm paper with all other farmers in the state, but allowance had to be made for the fact that the results of the sample were only approximate The average value of farm products sold per farm was $4,290 for the sample of 3,600 subscribers, and $4,162 for all farms in the state. This seems to indicate that the gross income of the subscribers to the farm paper was higher than for all farms in the state; but since the average for the farm paper subscribers was based on a sample, an allowance must be made for that fact.

The problem presented by this situation is an illustration of the testing of hypotheses by setting the hypotheses against observed data. If the observed facts are clearly inconsistent with the hypothesis, it must be rejected. If the facts are not inconsistent with the hypothesis, it may be given tentative acceptance, subject, of course, to rejection when additional information is secured.

We may set up and test the hypothesis that the true mean value of farm products sold by subscribers to the farm paper is the same as the mean value of farm products sold by all farmers, that is, $4,162 instead of the $4,290 secured from the sample of 3,600. If the test indicates that the above statement is not a reasonable one, we must reject the hypothesis.

The standard deviation of the universe was $2,280 and will be represented by $\hat{\sigma}$ as on page 223. It has already been demonstrated that the means of all possible samples of size n from a universe will form approximately a normal distribution, except when n is very small. The distribution of means will have a mean value that is the same as the mean of the universe. Since in this case the mean of the universe was known to be $4,162,

the mean of the distribution of all sample means would be \$4,162. The standard deviation of the distribution of sample means may be computed as follows:

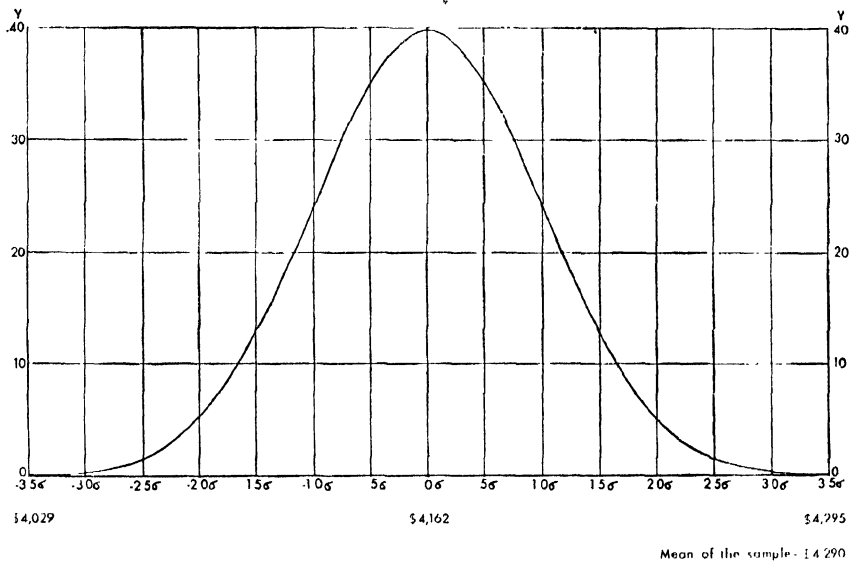
$$\sigma_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{\$2,280}{\sqrt{3,600}} = \$38$$

The difference between the universe mean, which may be designated by M, and the sample mean $(\bar{X})$ is \$128 $(\bar{X} - M = \$4,290 - \$4,162 = \$128)$. The question to be answered is whether or not a difference of this much can be explained as a variation due to sampling, or whether there is a significant difference between the universe mean and the mean of the sample.

Chart 10-1 shows the distribution of means of samples of 3,600 drawn from a universe with a mean of \$4,162 and a standard deviation of \$2,280. The mean of this distribution of means is also \$4,162; but the standard deviation of the distribution of means ($\sigma_{\bar{x}}$) is \$38, as computed above. The X scale on the chart is shown in standard deviation units running from $-3.5\,\sigma$ to $+3.5\,\sigma$, or in dollars, from \$4,029 to \$4,295. The mean of the

Chart 10-1

NORMAL DISTRIBUTION



sample of 3,600 subscribers to the farm paper is \$4,290 and falls at the extreme right of the scale. The position of the mean on the scale indicates that it is not impossible for a sample of 3,600 drawn from the universe described above to have a mean of \$4,290, but it seems very unlikely.

The Table of Areas under the Normal Curve (see Appendix G) shows that 99.73 per cent of the area falls within the limits of $\pm 3\,\sigma$. Since the

sample means are distributed approximately normally, it follows that approximately 99.73 per cent of these sample means will fall between the $3\sigma_{\bar{x}}$ limits. The mean of the 3,600 subscribers was 3.37σ from the universe mean; the Table of Areas under the Normal Curve shows that 99.92 per cent of the sample means would fall within these limits. The probability of such a deviation resulting from sampling variation is so small that it seems very unlikely, so we are justified in saying that the average value of gross farm income is significantly higher among subscribers to the farm paper than among all farmers. It is rather generally accepted that a deviation of 2.58σ is large enough to establish that a difference is significant, rather than the difference being due to sampling variation.

The testing of hypotheses by the use of sample data offers many complex problems that are properly included in advanced courses in statistics. The preceding discussion is intended merely to indicate the nature of the subject; a specific application in business is described in more detail in Chapter 23.

STUDY QUESTIONS

1. Explain why business has found the use of samples so profitable in collecting external statistical data. Give particular attention to the question of cost of the information in relation to its accuracy.

2. What is the important difference between the methods of taking a judgment sample and a probability (random) sample? Explain why the probability sample is considered better than the judgment sample.

3. Once an individual is chosen at random as part of a probability sample, it is considered wrong to substitute another person if the one drawn is not at home or for any other reason cannot be interviewed. Explain why.

4. An opinion study is planned among students on the campus to secure their opinion of a compulsory fee of $15 per semester which will be used for the support of the student union. Assume that at the present time there is no fee for support of the union, and that there is a general feeling among the students that money for the support of the union should be provided from some source. With approximately 20,000 students in the school it is unlikely that anyone would pay for a complete enumeration of all the students.

(a) Describe how a strictly random sample of students could be selected for this study.

(b) How large a sample would you recommend be taken for this study? Explain how you decided upon this size.

5. A dairy has been selling milk in glass bottles ever since it started in business, but there is a growing feeling among some of the executives that it would be profitable to offer milk in paper cartons as well as in glass bottles. Some of the staff believe that their customers would like to have a choice of containers, but other staff members insist that their customers prefer milk in glass bottles.

Finally the dairy executives decide to conduct an opinion survey to learn their customers' views on the use of a paper carton. The dairy sells milk in stores and also delivers to homes. It is desired to find out how their store customers and their home-delivery customers would like a paper container.

Describe how a random sample of all their customers could be taken.

6. Describe how you would select a random sample of all the tire dealers in your state.

7. Assume that a random sample of 1,000 passengers is taken at stated intervals during the day at a large city airport and that the passengers are asked whether they are changing planes at this airport or whether they are beginning or ending their flight there.

(a) How accurate would you expect the answers to the 1,000 interviews to be in comparison with the answers you would get if you asked all the passengers the same question?

(b) How would you take a random sample of all passengers leaving and arriving at the airport?

8. Assume that you have a sample of retail stores in a state that reports retail sales each month. The sample is random and from it an estimate of total retail sales for the state is made each month. The confidence interval is not too large and the information is found to be very useful to businessmen.

You are asked to use your sample of stores to make estimates of retail sales in each county since you have stores in the sample from every county in the state. It is not expected that you should secure any larger sample; you are asked merely to tabulate the sample results by counties and use the sample for each county to estimate retail sales in the county. It is argued that if the sample gives a good enough estimate of retail sales for the state, it should give equally good estimates for the individual counties, since the sum of the county estimates would be the same as the estimate for the state.

What answer would you give to this request?

9. Assume that a manufacturer has designed a completely new package for his product and after a year of using this new package he is considering making a consumer survey to find out if his customers like it. It is obvious that a sample should be used, but the question immediately arises as to the cost of such a study You estimate that interviews will cost $2.35 each to make and to tabulate the results.

Explain how you would go about estimating the cost of conducting the survey.

PROBLEMS

1. A market research study was undertaken to determine how much was spent in a particular year by automobile owners in the United States for gasoline and oil. A sample of 1,000 automobile owners was taken at random. The average expenditure of this sample was $203. The standard deviation of the 1,000 amounts reported was $86.

(a) What would you estimate to be the average expenditure for gasoline and oil of all automobile owners in the United States?

(b) Compute the confidence interval of this estimate and explain what it means.

(c) How would you select a random sample of all automobile owners in the United States in order to make the survey described?

2. Assume that you want a quick estimate of the average age of the 9,000 employees in the field offices of a company. You select 540 cards at random from the personnel file and compute the average age of these 540 employees to be 36.4 years. The standard deviation computed from the 540 ages on the personnel cards is 10.8 years.

(a) Make an estimate of the average age of the 9,000 employees of the field offices, and compute the confidence interval of this estimate. Explain what these two measures mean.

(b) Describe how you would take a random sample of the 9,000 cards in the personnel file.

3. A state employment commission has reports on employment of 200,000 firms in the state during the first quarter of the year. A random sample of 100 of these reports is taken, and the number of employees for each firm is secured from the report. The total number of employees reported by these 100 firms is 2,513. (The 100 individual numbers could be given, but to save space they have been omitted.) You also need to know the sum of the squares of these 100 numbers; this is found to be 76,875. (In other words, $\Sigma X = 2{,}513$ and $\Sigma X^2 = 76{,}875$.)

(a) What would you estimate to be the number of employees per firm in the state?

(b) Compute the confidence interval of this estimate and explain what it means.

4. A manufacturing company employing 10,000 persons has a progressive employee relations program. It was decided to make a survey to determine the state of employee morale and to secure information that would be used in future labor relations efforts.

A random sample of 600 employees was interviewed at home. One question asked was how much the company contributed to the retirement plan and how much the employees contributed. Thirty per cent of those interviewed did not know.

(a) If all of the 10,000 employees had been interviewed, what would you estimate to be the percentage that would reply that they did not know the answer to the retirement plan question?

(b) Do you think the sample is accurate enough for the use of management in deciding whether or not an educational program should be initiated to inform the employees about the company's retirement plan? (The company paid all of the contribution to the retirement plan.) Before answering this question, it would be wise to compute the confidence interval of the estimate secured from the sample.

5. Assume that you are making a survey to ask a number of questions about consumer brand preferences. All of your tabulations will be in the form of percentages (such as the percentage of people who prefer your brand, and the percentage of people who do not use your product). You are instructed to secure a confidence interval of not less than $\pm$ 1.8 for each percentage. (It is satisfactory if the confidence interval is less than $\pm$ 1.8 for some percentages, but it must not be more.)

Compute the size of the sample that you would need to be certain that the confidence interval did not exceed $\pm$ 1.8.

6. Assume that you are planning a study of family incomes in a city and you have reached the point where you must make an estimate of the cost of the study. This requires an approximation of the number of interviews to be taken. Since the size of the sample needed will vary with the accuracy required in the estimate of family incomes, a conference is held to decide how wide a confidence interval the users of the data can tolerate. At this conference it is decided that the estimates must not vary more than $240 from the average income that would have been found if a complete study were made of all families in the city.

Other income studies suggest that the average income per family might be somewhere in the neighborhood of $3,600 and that the standard deviation might be approximately $1,200. This information is not much better than a guess, but it is the best information that you have on which to base your computation of the size of sample needed.

When you tabulate your data, you find that the mean income of your sample is $3,858 and the standard deviation of the sample is $1,310.

(a) Compute the size of the sample that you would recommend, taking into consideration the information above.

(b) Compute the confidence interval you have secured in your study. If your confidence interval is smaller than the $240 specified, your study is complete.

(c) If the confidence interval is larger than $240, what should you do?

7. The study described in Problem 6 was made in a city of 800,000. Later a survey was wanted in a city of 4 million, and it became necessary to compute the size of the sample needed in the second survey. There was good reason to believe that the dispersion in incomes for the larger city would be essentially the same as in the first city studied.

What size sample would you recommend for the city of 4 million, assuming that the same degree of reliability was wanted as in the first survey?

Part

IV ANALYSIS OF BUSINESS CHANGE

chapter 11 ♦ Time Series

chapter 12 ♦ Secular Trend—The Straight Line

chapter 13 ♦ Secular Trend—Nonlinear

chapter 14 ♦ Seasonal Variation

chapter 15 ♦ Cyclical Fluctuations in Time Series

chapter 16 ♦ Construction of Index Numbers

11 TIME SERIES

Data classified on the basis of intervals of time constitute vital information in the control of business activity, since this is the most effective method of showing the changes that are taking place in a business, an industry, or in total economic activity. Closely related to the problem of measuring changes in business activity is the making of forecasts of future activity. The management of a business enterprise requires a continual making of decisions regarding the future, and the basis for such forecasts is the record of past performance.

It is possible to express information concerning past performance of business in general terms, but this is decidedly inferior to a more definite quantitative expression. For example, it may be said that business is "somewhat better than usual this fall and it is expected that further improvement will be registered." However, this statement does not approach in definiteness a tabulation of sales and production figures, compiled on a monthly basis over a period of years in the past and projected into the future.

Time series showing information important to a business may come from its internal records or may be collected outside the business. Table 11-1 shows the total registrations of trucks in the United States for the prewar period 1928 to 1941 and for the postwar years. The war years, when production of most models had ceased and sales of all motor vehicles were rationed, have been omitted from the table. In addition to the registration of all trucks, the table gives the figures for one manufacturer, International Harvester Company. These two series are typical of the way in which external industry data may be compared with the same type of information for one company. In this case the data for the company are not internal data since they did not come from the records of the business, but from external sources. This is an example of a situation in which a higher degree of comparability is secured by using external data for the individual business instead of the concern's internal records. The data on registration of motor vehicles are used to represent sales, since it is assumed that a vehicle is registered very soon after the sale is made. However, there probably would be some variation in the sales figures compiled by the individual company and the external data compiled from the registration of vehicles. Since registration data are used to represent sales

Table 11-1

NEW TRUCK REGISTRATIONS — ALL MAKES AND INTERNATIONAL HARVESTER TRUCKS: 1928 TO 1956

(THOUSANDS)

Year	All new trucks registered	International Harvester Trucks	
		Number of new trucks registered	Percentage of all trucks
1928	341.4	26.2	7.7
1929	526.6	31.4	6.0
1930	410.3	23.7	5.8
1931	313.9	21.1	6.7
1932	180.4	15.8	8.7
1933	245.9	26.7	10.8
1934	403.9	31.6	7.8
1935	510.7	53.5	10.5
1936	611.6	72.0	11.8
1937	618.2	76.2	12.3
1938	365.3	55.8	15.3
1939	486.7	66.0	13.6
1940	576.3	77.9	13.5
1941	640.7	92.5	14.4
1942	*	*	*
1943	*	*	*
1944	*	*	*
1945	*	*	*
1946	625.2	78.2	12.5
1947	879.1	113.4	12.9
1948	1,035.2	125.3	12.1
1949	962.0	91.4	9.5
1950	1,142.3	98.2	8.6
1951	1,003.9	95.4	9.5
1952	812.1	92.6	11.4
1953	930.3	95.8	10.3
1954	829.1	84.2	10.2
1955	957.0	100.4	10.5
1956	894.4	108.0	12.1

*All motor vehicles rationed.

SOURCE: Standard and Poor's Industry Surveys.

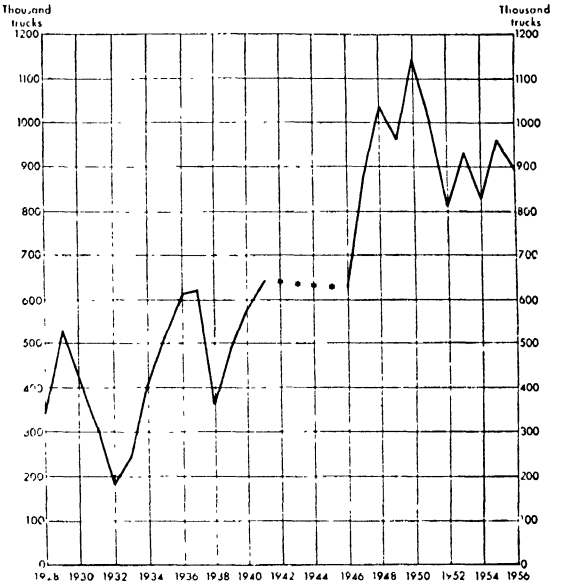
of all trucks, there is an advantage in using registration data for the individual company, even though the concern might have its own sales records.

The last column in the table shows the ratio of International Harvester Company trucks to total trucks registered in each of the years. The simple analysis depicted by this column of percentages is a summary of the position of this one company in the industry. For example, during the four years immediately before the war International Harvester Company had its greatest share of the truck market in the period covered by the data. It can also be seen from this column that the share of the market supplied by International Harvester Company has been declining since the early postwar years.

Chart 11-1 shows total new truck registrations, and Chart 11-2 shows registrations of International Harvester Company trucks. It can be seen from these two charts that the fluctuations in these two series followed somewhat different paths. The total registrations of new International Harvester trucks increased in 1947 and again in 1948, but the latter was the peak year. Total truck registrations (except those for 1949), however, continued to rise until 1950.

Chart 11-1

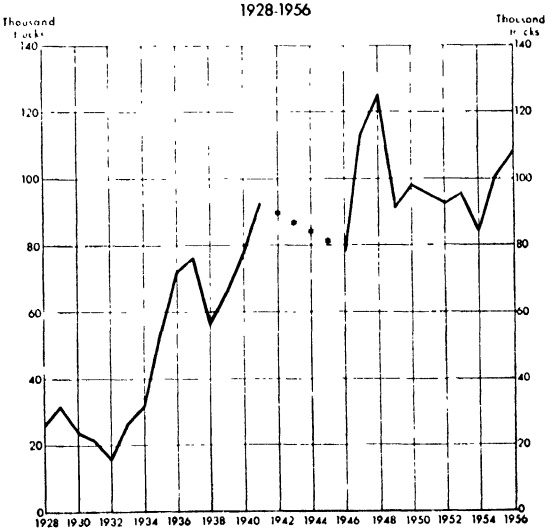
NEW TRUCK REGISTRATIONS IN THE UNITED STATES
1928-1956



*All motor vehicles rationed
Source of data Table 11-1

Chart 11-2

REGISTRATIONS OF NEW INTERNATIONAL HARVESTER TRUCKS IN THE UNITED STATES
1928-1956



*All motor vehicles rationed
Source of data: Table 11-1

Chart 11-3

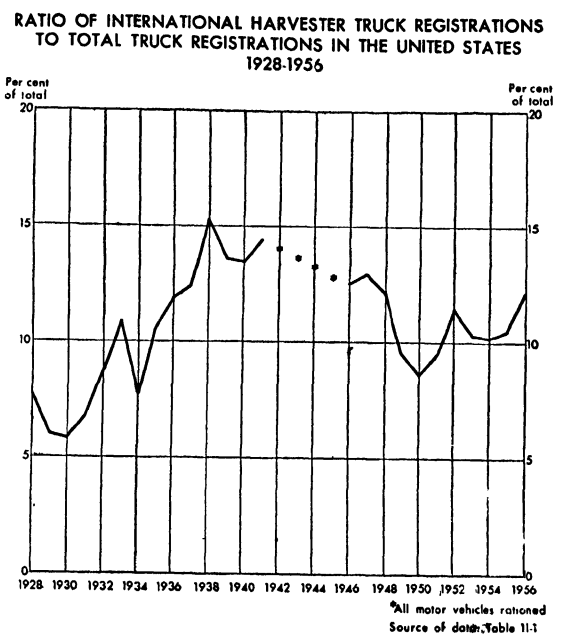


Chart 11-3 shows graphically the ratio of International Harvester truck registrations to total registrations for the period covered by Table 11-1.

Examples of data classified on the basis of time are found in the following discussion in the form of tables and charts. Since economic change occurs constantly, businessmen use large numbers of time series in keeping informed of changes that are taking place in their own concerns and in the economic environment outside their businesses.

TYPES OF FLUCTUATION[1]

The changes that occur in statistical series classified by periods of time are usually grouped into the following four categories, each of which represents a rather well-defined type of economic change.

1. Secular trend.
2. Seasonal variation.
3. Cyclical fluctuations.
4. Random or erratic fluctuations.

It is desirable to understand the nature of these different forces that interact on data expressed in terms of time units, since the whole subject of time series will be developed around this classification. Methods of measuring each type of fluctuation will be described, and the forecasting of time series will consist of forecasts of the separate types of fluctuation.

The forces affecting time series might be classified in a number of ways, but the classification presented above is one that is commonly used to distinguish between long-term movements, periodic movements throughout a year, recurring changes that do not necessarily occur in a fixed period, and random movements. The first of these movements, referred to as *secular trend*, reflects the effect of forces making for gradual growth or

decline. These forces operate over a long period of time, and are not subject to sudden reversals in direction. The second type of fluctuation is known as *seasonal variation*, because it represents a regularly recurring variation with the seasons of the year. Other types of periodic fluctuations, such as variation throughout the day, represent the same kind of change and are analyzed by the same methods as are used for fluctuations over the period of a year. The third type of fluctuation is distinguished from seasonal variation by the fact that it does not have a fixed period, although it is a recurring type of change. This phenomenon is referred to as the *business cycle* and the changes are called *cyclical fluctuations.* Intermingled with the three well-defined types of variation described above are innumerable small variations that are essentially random in nature, resulting from a large number of factors, most of which are relatively unimportant. These fluctuations are called *random* or *erratic fluctuations.*

Secular Trend

One of the outstanding characteristics of our modern economic system has been the growth of business and industry. This growth, or secular trend, is based primarily on a growing population, and the rate of growth in the total economy reflects directly the growth of the population. Unless a decrease in the standard of living is accepted, a growing population offers a market for an increasing amount of all kinds of commodities. Throughout the history of the United States the growth of population has been accompanied by an increasing per capita production of goods and services, with the result that the growth of total industry has been faster than the growth of population.

The growth of total industry means that industries and individual business concerns have grown, although the rate of growth has shown wide variations among industries. Synthetic fibers and other materials, household equipment, automobiles, and radio and television are typical examples of industries growing at very rapid rates. Some new industries replace existing industries, while others add their output to the total without actually replacing other products. Natural gas and petroleum have displaced a large amount of coal as a source of energy; motor vehicles have replaced horse-drawn vehicles; and rayon and nylon have replaced a major portion of the silk consumed and are making strong inroads on many uses of cotton.

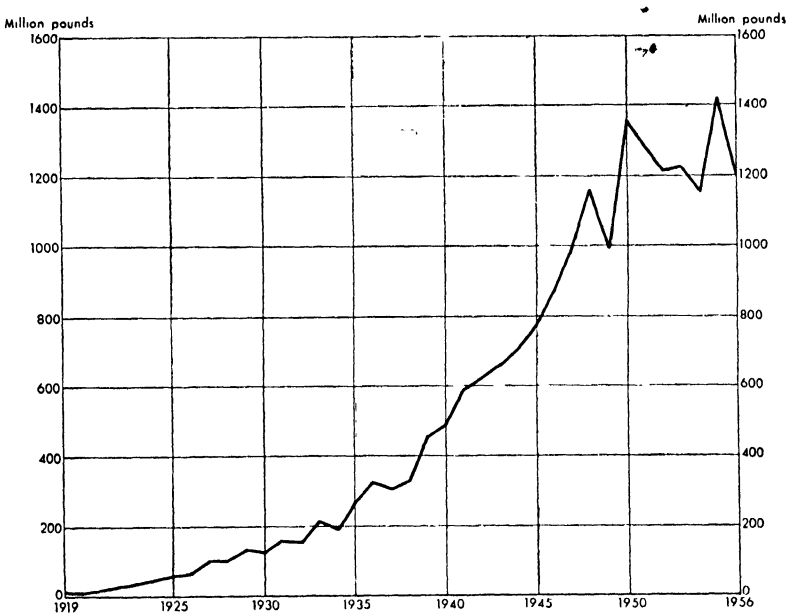
The changing character of modern industry makes it inevitable that some business units and industries supplying goods and services should grow and others decline. As long as inventions produce new products or new technical methods of producing products, it is to be expected that industries and businesses will continue to grow. And as long as consuming

habits change, it may be expected that certain industries will decline. Because this characteristic of time series is so prevalent, any analysis of time series must take this type of fluctuation into consideration.

Secular trend is to be found in prices as well as in physical volume series. Prices of commodities tend to decline as cheaper methods of production are developed. On the other hand, if the production of goods depends upon a supply of raw materials that is gradually being exhausted, the price may show a secular trend upward. In addition, the long-time movements of the general price level take on the characteristics of secular trend, although these movements have many of the characteristics of long cyclical fluctuations.

Chart 11-4

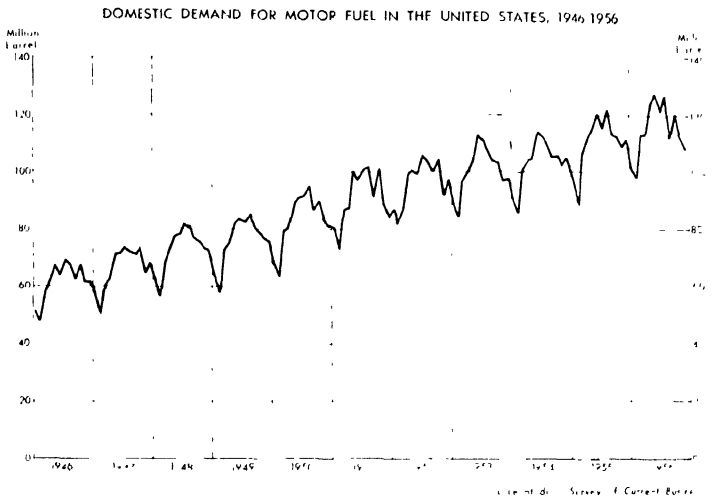
CONSUMPTION OF RAYON AND ACETATE FIBER
IN THE UNITED STATES, 1919-1956



Source of data Textile Organon, March, 1957

Chart 11-4 shows the consumption of rayon and acetate fiber in the United States since 1919, and illustrates the rapid expansion of these new products; this series is typical of the industries in which new products have been capturing an increasing share of a market. Chart 11-5 shows the increase in consumption of motor fuel since 1946; and although the series is classified on a monthly basis, the rising trend is clearly visible. In eleven years the demand for motor fuel nearly doubled.

Chart 11-5



Seasonal Variation

The changes in business and economic activities that result from the changing seasons are true periodicities, and are reflected in changes in industrial production, agricultural production, consumer buying, and prices. Crops are produced in a regular annual cycle, except in the extreme southern portions of the country where more than one crop can be grown each year. Prices and stocks of farm products are influenced by the fact that production is geared to a yearly cycle, while consumption is distributed throughout the year. Consumer needs vary widely from one season to the next. Antifreeze for automobile radiators is needed in the winter; coal, oil, and gas are needed for fuel in the winter; but greater amounts of electricity are used in the summer for air conditioning. The seasonal pattern of the consumption of electricity formerly showed a peak in the winter when the days were short and lights were burned more hours, but more recently the consumption of electricity for air conditioning in the summer has pushed the seasonal peak to the hot months of the year.

Social customs dictate the time of year many articles are purchased, and distinctive seasonal patterns, independent of changes in the weather, are found. Christmas gifts swell the December retail trade figures; June commencements and weddings increase the business in certain commodities that are suitable for gifts. The buying of new clothes for Easter may be partly a result of the change in the weather, but it is probably much more related

to the custom of acquiring new spring clothes for that particular Sunday, the date of which may vary as much as a month from year to year. The practice of introducing new models each year gives a seasonal pattern to the buying of certain types of durable goods — people wait for the new models, sales decline, and then a sharp increase in sales is experienced when the new models are put on the market. The automobile industry is characterized by this type of seasonal fluctuation, combined with a seasonal pattern resulting from the fact that people like to buy durable consumer goods at the beginning of the season of greatest use.

Periodic fluctuations in consumer purchasing, of course, cause sales of retailers to be distributed unevenly throughout the year, which in turn affects the regularity of sales of wholesalers and manufacturers. Since many manufacturers produce only when the sale of goods is assured, this bunching of orders causes the rate of production to fluctuate and to follow a regular pattern. This seasonal pattern is then reflected in the purchase of materials and supplies and in the employment of labor. Thus a seasonal pattern in the buying habits of the consuming public causes a number of seasonal variations in different phases of business activity.

There are other periodicities in business besides those fluctuating with the seasons; in fact, seasonal variation may be thought of as merely one special type of periodic fluctuation. The number of telephone calls varies rather regularly with the time of day. This variation is so pronounced that special rates are given during the periods of slack business to build up a greater volume. City transit systems experience a decided periodic fluctuation between the different hours of the day. Intercity passenger transportation has peaks over week ends and on holidays. Many types of retail stores show a wide variation in sales on different days of the week. All of these periodic fluctuations can be analyzed in the same manner as seasonal variation by simply using the proper period: day, week, month, or any other period of time except the year.

The data in Chart 11-5, giving domestic demand for motor fuel in the United States, show a pronounced seasonal pattern. Demand rises regularly from a low point in the winter to a peak in midsummer, then declines steadily until midwinter. This seasonal pattern undoubtedly reflects the effect of the weather on driving habits of the public. Inclement weather brings about less driving in the winter than in the summer, and the practice of taking the annual vacation in the summer increases the amount of motor fuel consumed. As cars and highways were improved, the seasonal peak in the summer was reduced and cars were used more in the winter months. There tends to be less seasonal variation in the demand for motor fuel in the states farther south where driving all year is easier and more pleasant.

Cyclical Fluctuations

In many respects the cyclical fluctuations of time series represent the most important type of variation, chiefly because these fluctuations are less predictable than seasonal variation and secular trend and therefore can have a more damaging effect on the fortunes of a business enterprise. The cyclical variations in individual series are related to the cycles in the total economy. Business cycles have been described by Burns and Mitchell as successive waves of expansion and contraction that occur at about the same time in many economic activities.[1] Expansion is followed by a recession, which gradually merges into a revival and expansion, with the different phases following each other in regular fashion. These phases are recurrent — but not periodic, as in the case of seasonal variation. The length of a complete cycle varies from a year to approximately twelve years, and the amplitude of the swings varies from one cycle to another.

The cycle in total economic activity may be thought of as the average of the cycles in individual segments of the economy; however, it is very common to consider the cyclical fluctuations in total business activity as a phenomenon distinct from the cycles in the individual series that measure the various parts of the total economy. Regardless of one's theory about the relationship of the individual components of economic activity to the over-all level of activity, it is important to measure the cyclical patterns that prevail in the individual series. After this has been done, it is in order to compare the pattern in an individual series with the cycle in total activity and with other individual series. Although there is a considerable degree of relationship between the cyclical fluctuations in individual series, the measurement of the difference in behavior of the individual series is significant. Some series lead at the turning points, such as the peaks and valleys of the business cycle, while others lag. Some series show wide fluctuations through a complete cycle, while others show a very mild variation. The study of specific industries and individual business concerns to determine the cyclical behavior of their production, sales, and prices, for example, gives management much information that is of importance in making decisions.

Erratic Fluctuations

The day-to-day operations of a business are influenced by a multitude of factors. In nearly every business organization, sales vary from one day to another because of purely chance factors. In a retail store the weather

[1]Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946), p. 3.

on any particular day makes a difference in the amount of business done. In a concern where sales are in large units, some days will show no sales and others will show very large sales, for no other reason than that the orders happened to be closed on one day rather than another.

A great many of the day-to-day irregularities in business activity are of this nature. When data on the amount of sales or production are classified by days, the figures for different days would be expected to show the effect of these irregularities. If the days are added together to make a weekly classification, there would be some tendency for these irregularities to smooth themselves out. If the period were a month, there would be still more of a tendency for this process to operate; and if it were a year, the smoothing would be still greater.

As businessmen try to keep closer control over business, they have made use of figures classified on smaller and smaller intervals of time. At one time financial statements were made out once a year, but there is a steady tendency toward more frequent statements. Yearly data are not now considered adequate. Quarterly data are likewise not considered the best possible division, and more emphasis is being placed on reporting data monthly, weekly, or even daily. *The shorter the interval of time on which the data in the form of totals are classified, the more tendency there is for erratic fluctuations to influence the series.*

Some types of data are more affected by these erratic forces than others. In general, the statistical series for an individual business are likely to be more irregular than for an industry as a whole. When a large number of individual business units combine their data into a total, there is a tendency for individual irregularities to cancel out. Also, within the same organization certain types of information may show a greater tendency to be affected by these irregular forces than others.

Occasionally an irregular fluctuation may be important enough to be given special study, as when retail sales figures in a particular community show the effect of a convention held there, or production figures of a company show the effect of a strike. Unless the irregularity is large enough to be especially significant, it is ordinarily pointless to try to find the specific causes, since the causes may not be repeated with a uniformity that aids forecasting. The irregular, "saw-toothed" effect seen in many time series classified on a monthly or weekly basis usually results from the fact that the time unit used was so small that the irregular forces affecting the data were not given opportunity to cancel out. When it is desirable to use the small time unit, such as a week or a month, these irregularities in the data must simply be ignored, or some simple method of smoothing should be used.

IMPORTANCE OF TIME SERIES ANALYSIS

It has already been emphasized that the basic purpose of all business statistics is to give the businessman information on the workings of his business and the environment in which it operates. The basic purpose of time series analysis is to give the businessman a convenient method of measuring changes in his business and relating these changes to the changes in the business environment. The time series measuring changes in one's own business are supplied by the internal records of the company, while information on changes in the whole industry and in business in general will come from various external sources.

The special methods of time series analysis will be given detailed treatment in the following chapters. Of the four types of classification described on pages 19 to 25, time series generally require the most extensive analysis before being of maximum usefulness to the business executive. For many statistical data used in business, mere classification is all the analysis needed to make them ready for the executive's use; but this is seldom true for time series. For this reason, more of this book is devoted to the analysis of time series than to any other one topic.

A time series requires more analysis because of the number of different influences that regularly affect it. Since it may be a composite of any one or all of the four types of fluctuations described above, a considerable amount of analysis may be necessary to determine what caused the various changes.

The basic problem is the isolation of the different types of fluctuations, particularly trend, seasonal, and cyclical. It is not enough merely to keep records or collect data on the various aspects of business activity. Data supplied by such records show the composite change in activity; but since several forces are at work, the data show only the net results of these forces. For example, the volume of production of an industry might show a net increase that was the result of a sharp cyclical decline combined with a regular seasonal rise that more than offset the cyclical drop.

The seasonal problems of a business are very different from those growing out of the cyclical swings of business. And both of these differ from the problems of long-term trends. The measurement of each of these fluctuations becomes a basic step in the use of data classified according to time.

When management uses any type of historical data, some problem of forecasting is generally involved. The record of the past is the main basis for making an estimate of the future. Since business decisions affecting the future inevitably require a forecast, the statistical record of past events represents an element in management decisions.

Just as the measurement of fluctuations in time series requires a separation of the different types of fluctuations, any forecast should be based on a forecast of the different types of variations. Separate forecasts should be made of the trend, seasonal, and cyclical variations although it is doubtful if a forecast of erratic fluctuations is feasible. Any budgeting in a seasonal business must take account of the seasonal element. A concern with a heavy fixed investment cannot avoid forecasts of the long-term trend, since the business expects to operate for a long time. Very few concerns can afford to ignore the cyclical fluctuations and their inevitable effect on the individual industry and business.

ADJUSTMENT FOR CALENDAR VARIATION

There are occasions when the varying length of the months introduces an irregularity in monthly time series. For example, daily production in a factory might be almost the same in February and March; but because there were only 20 working days in February and 23 in March, there would be an increase in the March figures. Seven months have 31 days, four have 30 days, and February has either 28 or 29 days. The number of Sundays in a month may be either four or five, except that February can have only four Sundays when it is 28 days long.

These variations in monthly data due solely to the number of days cause illogical and erratic fluctuations in many time series. A method of removing this irregularity is to divide the data for a month by the number of days, or by the number of working days in the month. A word of caution should be given, however, for making this adjustment is not so simple as it might appear to be. Whether the correction should be made on the basis of total days in the month or the number of working days depends upon the data. In the example above, it is assumed that production is on a five-day week basis, but in some industries production is continuous. If such is the case, the adjustment should be made by dividing production by the number of days in the month. In production series there is a more clearly defined relationship between the number of working days and the amount of output than in such series as retail sales. In evaluating the results of retail business for a month, it is clearly important to make allowance for the number of days a store is open; but it is somewhat uncertain that the adjustment can be made accurately as in the case of a manufacturing plant. In retailing, certain days of the week are better business days than others; so it is sometimes desirable to make an allowance for the number of Saturdays in a month as well as the total number of days a store is open.

The Board of Governors of the Federal Reserve System publishes extensive statistical data on production in manufacturing industries and has

made detailed studies of the work week used in different industries.[2] The continuous-process industries generally use a seven-day week, while in other industries the five-day week now predominates. Using this information it is possible to compute the necessary adjustment for the number of working days in each month. Table 11-2 shows lumber production for the twelve months of 1956 in millions of board feet, the number of working days in each month, and the average production per working day for each month. The lumber industry uses the five-day week, which means that the smallest number of working days possible in a month is 20 and the largest number is 23. Since 1956 was a leap year, and since there were four Saturdays and four Sundays in February, there were 21 working days. Although September has 30 days, in 1956 there were five Saturdays and five Sundays, so there were only 20 working days.

Table 11-2

LUMBER PRODUCTION IN THE UNITED STATES, 1956

TOTAL MONTHLY PRODUCTION AND PRODUCTION PER WORKING DAY

Month	Production[1] (Millions of board feet)	Number of working days[2]	Production per working day (Millions of board feet)
January	2,940	22	134
February	2,922	21	139
March	3,092	22	141
April	3,145	21	150
May	3,431	23	149
June	3,312	21	158
July	3,067	22	139
August	3,538	23	154
September	3,147	20	157
October	3,403	23	148
November	2,975	22	135
December	2,554	21	122

SOURCE: [1]*Survey of Current Business*, Vol. 37, No. 3 (March, 1957), p. 31.

[2]Computed on the assumption that a five-day week is the general practice in the lumber industry.

[2]*Federal Reserve Bulletin*, Vol. 39, No. 12 (December, 1953), pp. 1,280-1,281.

The importance of the adjustment for the number of working days in the month is demonstrated in Table 11-2 by the production of lumber in January and February, 1956. February production was less than in January, but there was one less working day in February. When production is adjusted for the number of working days, February production was 3.7 per cent higher than January.

In adjustment for calendar variation, the basic difficulty is with the calendar, and this difficulty would be eliminated by a change which would make all months the same length. One proposal for calendar reform is to divide the year into thirteen months of four weeks each, but this is such a break with tradition that it has very little chance of being accepted. A less drastic revision would be to retain the twelve months but have two thirty-day months and one thirty-one-day month in each quarter. This would make all the quarters the same length and would give only a small variation between the lengths of the various months. Each system would total only 364 days, so it is generally proposed that New Year's Day be an extra day between the last day of December and the first day of January and not included in any month. On leap years an extra day could be added between the last day of June and the first day of July. Agitation for calendar reform has been going on for a long time, but the prospects are none too bright that it will be accomplished in the near future.

For many years the adjustment for calendar variation took into account the number of holidays in the month, but the present practice generally is not to make this adjustment. The reasoning used is that the holidays occur in the same month every year[3] and their effect can be included in the measurement of seasonal variation, since seasonal variation is defined as the regularly recurring fluctuations within a year.

SEMILOGARITHMIC CHARTS

The presentation of time series in graphic form was discussed in Chapter 7 and examples of bar charts and line charts were given. The principles of construction are essentially the same for all types of bar charts and line charts discussed in Chapter 7. In general the classes into which the data are grouped are shown on the X axis and the numbers or amounts in the various classes are shown on the vertical scale. All of the examples in Chapter 7 use an arithmetic scale for the Y axis, which signifies that a given space on the scale always means a certain number or amount wherever

[3]The most important exception to this statement is the date of Easter, which may come in March or April. Whenever the date of Easter has an important effect on a series, as it does in department store sales, it becomes important to allow for a different seasonal pattern depending on the date of Easter.

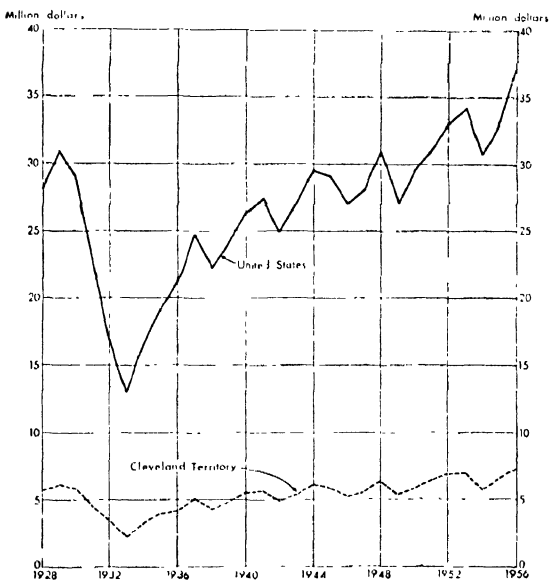
it appears. Thus, a given space that represents $1,000 is the same whether it shows the difference between $2,000 and $3,000, or between $98,000 and $99,000.

When a logarithmic scale is used for the *Y* axis, it is laid out according to the logarithms of the numbers on the scale. This is a specialized type of chart used for time series, and its purpose is to show the rate of change from one period to another. The ruling on the chart is such that the figures are automatically reduced to a percentage basis when plotted, and *the same vertical distance anywhere on the chart shows the same percentage of change.* A decrease of a given percentage represents the same distance on the scale, whether the decrease be from $1,000,000 to $500,000, or from $10,000 to $5,000. On the arithmetic scale the first decrease, $500,000, would take up 100 times as much space as the second decrease, $5,000, and yet the percentage decrease is the same. So if the interest is in percentage changes in the data, the *semilogarithmic chart* should be used. This type of chart is useful in comparing percentage changes in two series, or in two parts of the same series. If two series are shown on the same chart, the slope of each line shows the *percentage* change in that series. By comparing the slopes of the two lines it is possible to compare the percentage changes in the two series.

Chart 11-6 shows the total sales of a company, and the sales of one salesman in the company's Cleveland territory. The chart is drawn on the

Chart 11-6

SALES OF COMPANY X, 1928-1956
United States Total and Cleveland Territory



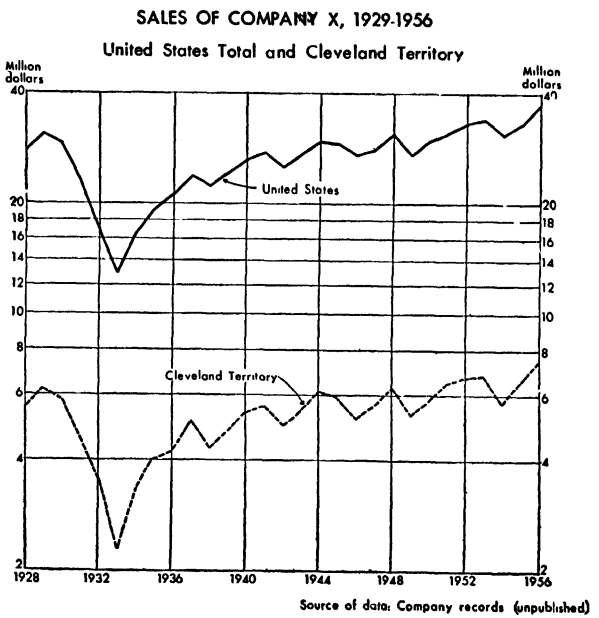
Source of data: Company records (unpublished)

arithmetic scale, and so the distance of each line from the zero line of the chart shows the amount of sales. This chart shows how the salesman's volume compares with the total business of the company, but this is not the information that would most likely be wanted.

In this particular case the chart was being used in a discussion of the question of how well this salesman had maintained business in his territory during depressions. The comparison being made was between his sales and the total sales of the company. The salesman maintained that his sales had fallen off less than the total sales of the company, and referred to Chart 11-6 for support of his statement. In making such a comparison, however, the important factor is not the *amount* of change in either the company's sales or the salesman's sales but the *percentage* changes in the two.

The chart drawn on the arithmetic scale compared the amounts of change, but the data should have been presented on a semilogarithmic chart, as in Chart 11-7, so that the fluctuations in the two lines could be

Chart 11-7

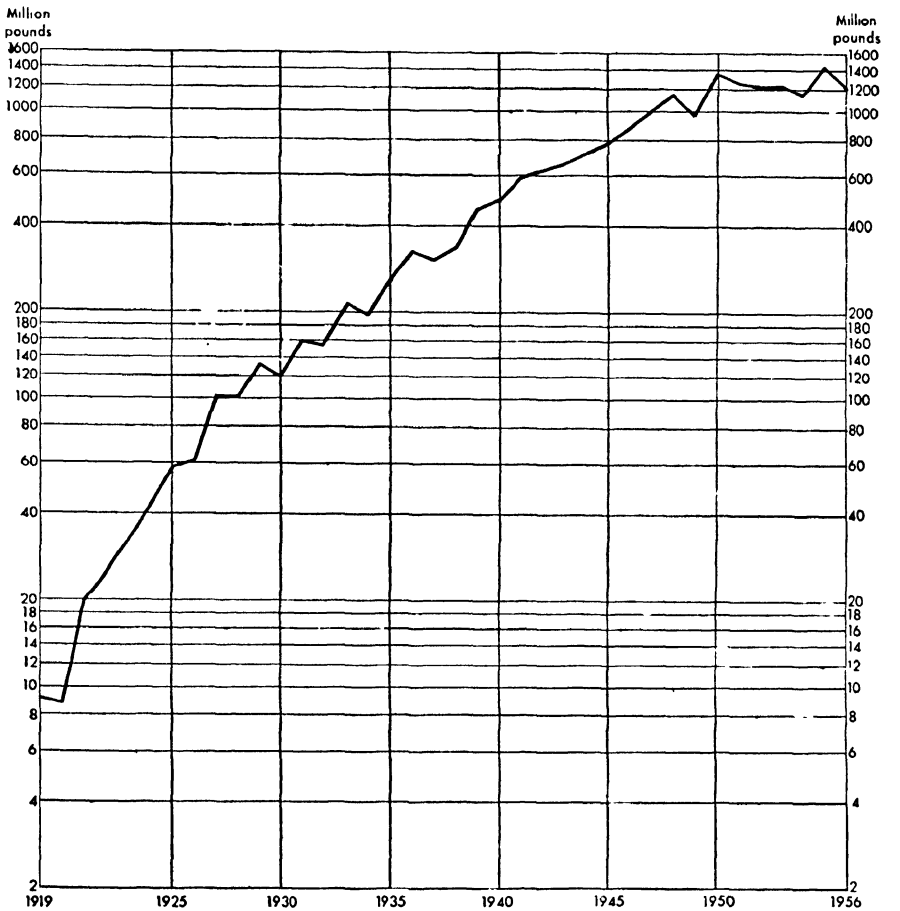


compared. The decline in the salesman's total from 1929 to 1933 is seen to be practically the same percentage of decline as the company suffered in its total sales. The question as to whether this particular man kept his business up as well as the business as a whole seems to be answered by this chart.

Charts 11-4 and 11-8 illustrate the growth of rayon and acetate fiber consumption in the United States. Chart 11-4 shows the amount of increase from year to year, while Chart 11-8 shows changes in the rate of growth. On a semilogarithmic chart a straight line shows a constant rate of growth.

Chart 11-8

CONSUMPTION OF RAYON AND ACETATE FIBER IN THE UNITED STATES, 1919-1956



Source of data *Textile Organon*, March, 1957

A constant rate has been maintained for four or five years at a time, but over the whole period the rate of growth has gradually slowed down. Comparison of the slopes at different points on the line will show the comparative rates of growth.

Chart 11-9

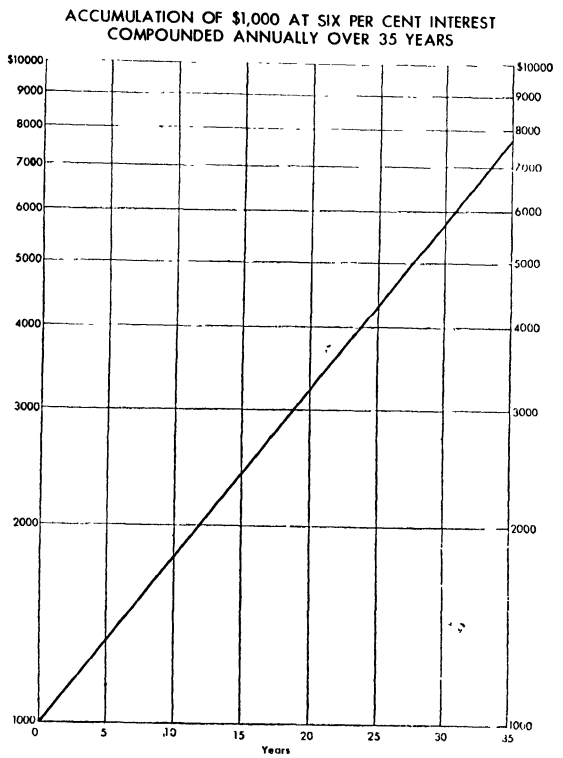


Chart 11-9 shows the growth of $1,000 left to accumulate at 6 per cent interest, compounded annually. The money increases at a constant rate — 6 per cent per annum. The line is, therefore, a straight line on semilogarithmic paper, and increases to $7,686.09 in 35 years.

The fact that a straight line on semilogarithmic paper shows a constant percentage of change can be used to advantage in computing depreciation charges when the method used involves finding a constant rate of depreciation to be applied to the diminishing value of the asset. The value of the asset, therefore, decreases by the same percentage each year. When plotted on a semilogarithmic chart, the net value of the asset is a straight line.

Table 11-3 shows the depreciation on an asset costing $6,000 and having an estimated scrap value of $400 at the end of eight years, the uniform rate on diminishing value being 28.7166 per cent. Chart 11-10 shows the data on a semilogarithmic chart.

Instead of computing the data in Table 11-3 and then plotting the values of the asset, it is possible to plot the data on the chart first and read the figures off the chart to get the carrying value of the asset at different years. Since the line is a straight line, it is necessary to locate only two points on it, and draw the line connecting these two points. The first and last values are known; in the above example the cost was $6,000 and the asset was to be depreciated to $400 in eight years. These two figures give two points on the straight line, which make possible the drawing of the line. The carrying value of the asset for any year between the first and the eighth may be read off the chart with sufficient accuracy, if the chart is carefully drawn. The credits to the reserve for depreciation may then be computed.

Table 11-3

DEPRECIATION TABLE — UNIFORM RATE ON DIMINISHING VALUE

RATE: 28.7166%

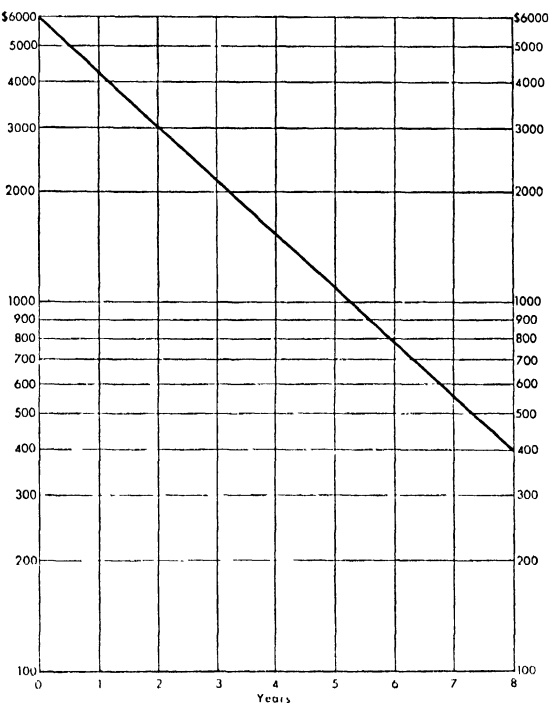
Year	Carrying value	Credit to reserve
Original Cost	$6,000.00	...
1	4,277.00	$1,723.00
2	3,048.79	1,228.21
3	2,173.28	875.51
4	1,549.19	624.09
5	1,104.32	444.87
6	787.20	317.12
7	561.14	226.06
8	400.00	161.14
Total.....	...	$5,600.00

One of the objections to the above method of determining depreciation charges is the difficulty of computation. Using the semilogarithmic chart as previously described provides an easy way of making the computations.

In using the semilogarithmic chart, it must be remembered that its primary purpose is not to compare the size of different series. Chart 11-7 gives the impression that the Cleveland territory sales were about one third the sales of the company as a whole, since the line showing the Cleveland territory sales is about one third as far from the bottom of the chart as the

Chart 11-10

DEPRECIATION AT A UNIFORM RATE ON DIMINISHING VALUE



line showing the total sales. Chart 11-6 on the arithmetic scale shows the relative size of the two series, and it can be seen that the total sales are about five times as large as sales in the Cleveland territory. The purpose of Chart 11-7 is to show *percentage changes in the two series; it is not an accurate picture of the absolute size of the series.* It is, of course, possible to read the figures on the chart and see the absolute size of the data, but the chart does not give an accurate impression of the magnitude of the values at a glance.

Setting Up the Scale

Ruled paper on which to construct semilogarithmic charts can ordinarily be obtained wherever graph paper is sold, or it can be constructed for the particular purpose at hand.[4] The size of the paper is designated by the number of cycles. Chart 11-8 shows three-cycle paper. The first line in a cycle is always ten times the value of the first line in the preceding cycle. The ruling in all cycles is exactly the same, and so the value of any line in any cycle is always exactly ten times the value of the corresponding line in the cycle immediately below. If this relationship is kept in mind, there need never be any confusion in setting up the scale.

Printed semilogarithmic paper has horizontal guide lines close together to facilitate plotting the data, but when data are plotted on a specially drawn grid, usually only the major guide lines are drawn. This improves the appearance of the chart, although it makes the reading of values from it rather difficult and may result in slight inaccuracies. It is also easier to plot the points if a piece of printed graph paper is used as a plotting scale.

The layout of the semilogarithmic chart is fundamentally no different from that of the other charts discussed previously. Since it does not make any comparison of size, the slope of the line is the only significant part. For this reason the chart starts at whatever point is convenient for showing the data, rather than at zero as in the arithmetic chart. It is obviously impossible to show zero on the semilogarithmic chart. Each line is one tenth the corresponding line in the cycle above, so the chart could be extended downward any number of cycles desired and still would not reach zero. This also means that there is never any part of the scale omitted as in the case of the arithmetic chart. Since the first line on the grid is not the base line, it should not be emphasized.

Using Two Scales

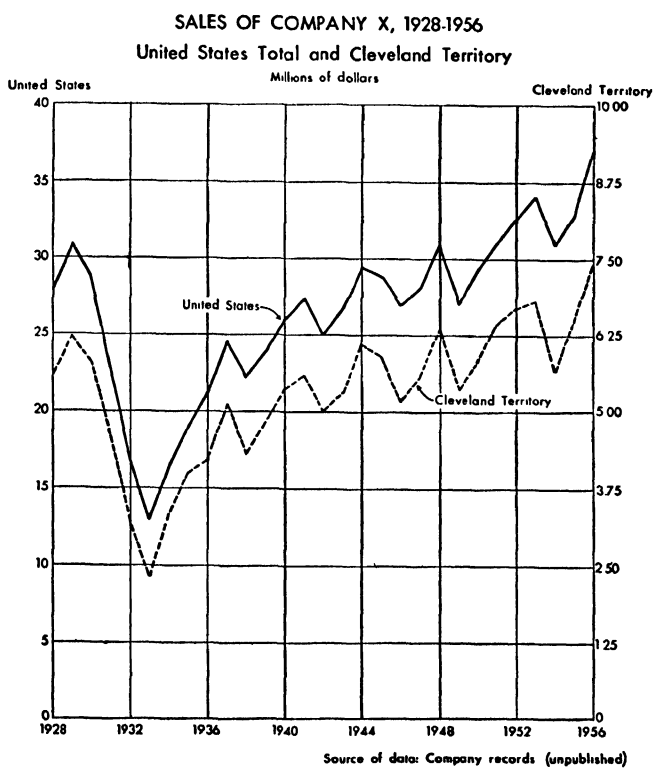
When two series are shown on arithmetic charts it is preferable that they should both use the same scale. Since the basic purpose of this chart

[4] A method that is perhaps even easier is to plot the logarithms of the actual data on an arithmetic chart.

is to show the size of the figures plotted, comparison can be made directly between two lines or parts of the same line. When a different scale is used for each line plotted, this comparison is impossible. Chart 11-11 shows the total company sales and the sales in the Cleveland territory, previously discussed. This chart is similar to Chart 11-6, except that two scales are used on the *Y* axis of Chart 11-11. It is quite difficult to tell from this chart how much larger the total sales were than the sales in the Cleveland territory. It does make easier, however, the comparison of the fluctuations in the two series. This suggests the principle that two lines can be compared on the arithmetic chart in the same manner as on the semilogarithmic chart *if two scales are used on the arithmetic chart in proportion to the two series to be plotted.* If the two scales are drawn in proportion to the two series, equal percentage fluctuations in the two lines will appear the same on the chart.

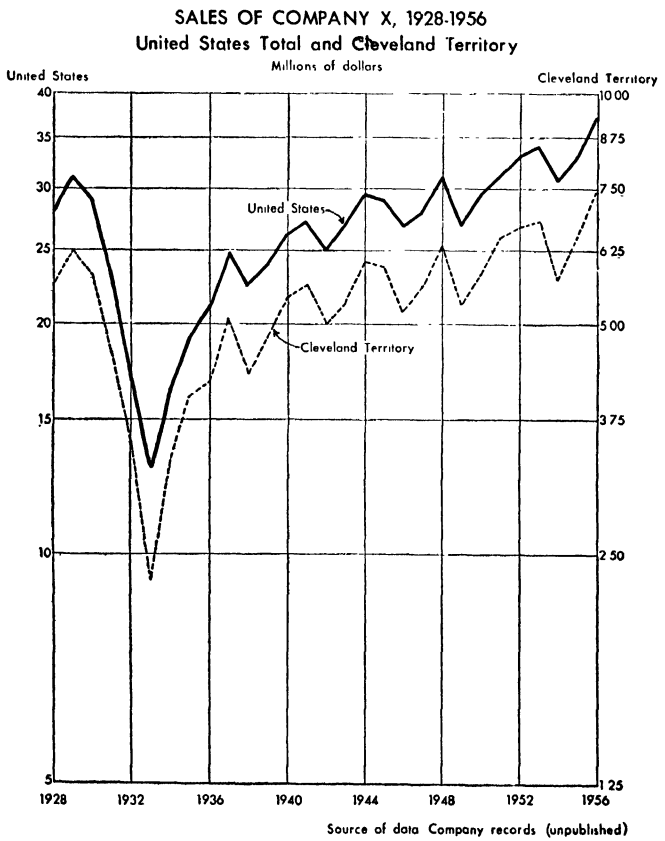
When using two scales on the arithmetic chart, it is necessary, first, to determine the ratio between the two series, and then to construct two scales that stand in this ratio to each other. Since there is so much danger that this will not be done accurately, it is better to use a semilogarithmic chart for this type of comparison.

Chart 11-11



Two scales can be used also on the semilogarithmic chart, as shown by Chart 11-12. The additional advantage of the semilogarithmic chart is that there is no problem of getting the scales in the same proportion. This is merely another way of saying that the two curves will have exactly the same shape, regardless of the scale that is used or where the lines come on the chart. This removes the danger of misrepresentation of the facts. In Chart 11-12 the fluctuations in the lines are exactly the same as when they were drawn on a chart with only one scale, as Chart 11-7. The only effect of using the second scale is to change the position of one line on the chart. In Chart 11-12 the line for the Cleveland territory is higher on the chart, but since the distance from the bottom of the semilogarithmic chart is not significant, this makes no difference. And there is the advantage that the two lines are closer together, which makes it easier to compare the fluctuations in them. So the use of two scales on the semilogarithmic chart involves no disadvantages, and frequently improves the chart.

Chart 11-12



Source of data Company records (unpublished)

DEFLATION OF TIME SERIES

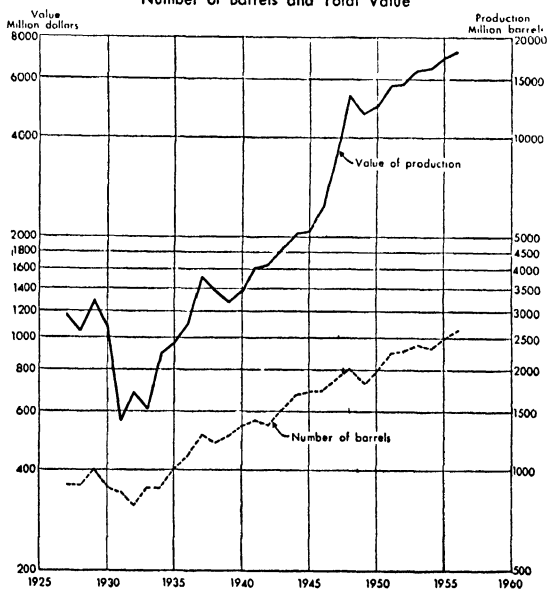
A great many time series are dollar values, such as sales, value of construction contracts awarded, value of building permits issued, national income, wages, and bank deposits. Series of this type are important for the facts they give, and they have wide use in business. However, all statistical series that measure aggregate values are the composite result of changes in physical volume and changes in the level of prices. Such a series as sales of a company, or of all the business concerns in an industry, are affected by two factors: (1) *the prices of the items sold,* and (2) *the number of units sold.* If the dollar sales figures of an industry increased during a period when prices were declining, it means that the volume of goods sold increased at a faster rate than the dollar sales figures.

For individual commodities the information on number of units produced or sold is usually supplemented by information on the total value of sales. Chart 11-13 shows both kinds of information for petroleum production in the United States for the years 1927 to 1956. Since this chart is drawn on a semilogarithmic grid, the changes in the two curves represent the relative changes in the value of petroleum produced and the amount produced. The greater decline in value between 1930 and 1931 is explained by the fact that the price fell sharply. The steeper rise in the value curve in recent years was caused by a rise in the price of petroleum. If the price had remained constant, the curve representing value would have risen at the same rate as the production curve.

When it is possible to express the changes in volume in units of measure, as can be done with petroleum production in the United States, both the value series and the volume series represent logical, well-defined time series. However, in

Chart 11-13

PETROLEUM PRODUCTION IN THE UNITED STATES, 1927-1956
Number of Barrels and Total Value



Source of data: *Statistical Abstract,* 1957

many situations no adequate physical unit is available. Sales of a department store, for example, are expressed in dollars; but it is difficult to conceive of a method of expressing sales in terms of any one physical unit. The total volume of goods and services produced by the economic system of a country is an important item of information, but it is impossible to design a physical unit of measure that can be applied to all kinds of goods and services.

In situations where no unit of physical volume can be applied, it is possible to derive a measure of the aggregate change in volume by adjusting a value series to remove the effects of the changes in the prices of the individual items, provided that a measure of the change in the average price of the items is available. For example, the total consumer expenditures in the United States increased between 1946 and 1956 for two reasons: the volume of goods and services purchased increased, and the prices paid for the goods and services rose. If we know the amount of the aggregate consumer expenditures in dollars and the average increase in retail prices, it is possible to compute a statistical series that reflects only the changes in the physical volume of goods and services purchased. This can be done even though there is no single unit of measure that can be applied to all of the items purchased by consumers.

The method used to adjust dollar figures for the changes in prices is called *deflation* of the dollar series. The violent rise in prices that occurred during World War I caused value series to increase much more than the related series representing physical volume. Since the series after adjustment for the level of prices were on a lower level than before adjustment, the term "deflation" was used to describe the process. This term continues to be used instead of a more generally descriptive term such as "adjustment for price changes," even though prices may be falling and the adjustment *increases* the adjusted series. In such cases the term "deflation" is hardly applicable, but it has come to mean simply adjustment of value series for changes in prices.

Value series representing income, such as wages or national income, are commonly called *real wages* or *real income* when adjusted for changes in the level of prices. The term "real wages" means the same as "deflated wages" and is generally preferred when used in connection with items of income. Another method of expressing the same idea is the term *constant dollars.* The United States Department of Commerce refers to deflated gross national product as "gross national product in constant dollars." This emphasizes the fact that the purchasing power of money in the different periods is the same, or the values are expressed in dollars of constant purchasing power.

The mechanics of deflating a value series are (1) to divide the series by a percentage relative that measures the change in the level of prices between the period under consideration and a base period, and (2) to multiply this quotient by 100. For example, the gross national product in 1956 was $412.4 billion and prices in 1956 were 124.86 per cent of prices in 1947. The deflated gross national product was $330.3 billion.

$$\text{Deflated gross national product} = \frac{412.4}{124.86} \times 100 = 330.3$$

The division by the relative measuring the relation of prices to 1947 expresses gross national product in 1947 dollars, or represents the total value of the gross national product if prices had remained unchanged from the 1947 level. (Multiplying by 100 is necessary since the measure of prices is expressed as a percentage. The price relative is $\frac{124.86}{100}$ and when dividing by this relative it is inverted and multiplied by the value being deflated.)

The deflated gross national product might be expressed as "330.3 billions of 1947 dollars" instead of using the term "constant dollars." If another period such as 1939 were used as 100 per cent in expressing the change in the price level, the results would have been in "1939 dollars."

It is essential that the relative used to measure the changes in the level of prices accurately reflect the average change in prices. Such a relative is called a *price index*; the construction of such price indexes is described in detail in Chapter 16. It is sufficient to state at this point that if the measure of the change in prices is not accurate, the resulting deflated series will not be an accurate representation of the changes in physical volume. It is extremely important to insure that the price index used to measure the changes in the prices be comparable with the value series being deflated. For example, if dollar building figures are being deflated, the price index must measure the changes in the cost of building the kind of structures included in the aggregate values. If export figures are being deflated, the price index must measure the prices of export commodities. Frequently the precise index needed is not available, but it should be remembered that when another is substituted the deflated series will be inaccurate to the degree that the price index used deviated from the correct change in prices.

In spite of the difficulty of securing an accurate measure of the prices included in all aggregate value series, the method of deflating time series is extremely useful. Many series can be compiled only as aggregate values, but when deflated with even an approximate measure of price changes, the series will serve as an approximate measure of physical volume. The concept of "real wages" and "real income" is extremely useful since the fluctuations in total dollar series fail to give a complete picture of the situation in periods of major shifts in the price level.

Annual hourly earnings in manufacturing industry in the United States in January, 1956, were \$2.05, compared with \$.70 in 1939. The increase of 198 per cent did not mean that wage earners had that great an increase in purchasing power. The index of consumer prices was 59.4 in 1939 and 118.2 in January, 1956. Each of these indexes was a percentage of the 1947–1949 average index, so the \$.70 wage rate in 1939 was \$1.18 in 1947–1949 dollars and the \$2.05 rate in 1957 was \$1.73 in the same purchasing power.

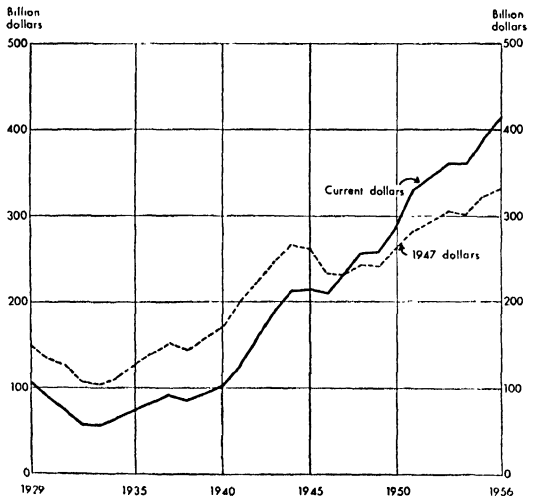
$$\text{Real wages, 1939} = \frac{\$.70}{59.4} = \$1.18 \qquad \text{Real wages, Jan., 1956} = \frac{\$2.05}{118.2} = \$1.73$$

The real wages in January, 1956, were 47 per cent more than in 1939. This percentage increase measures the change in the amount of goods and services an hour of work in a factory would buy in 1957 compared with 1939.

Chart 11-14 and Table 11-4 show the gross national product in the United States for the years 1929 to 1956, expressed both in 1947 dollars and in the current prices prevailing each year. Since 1947 there has been a large rise in the total value of goods and services produced in the United States, but the deflated series has risen much less. This means that a substantial portion of the increase in the value of the total output of goods and services has resulted from rising prices. The series expressed in constant (1947) dollars represents the rise in the physical volume of goods and services produced.

Chart 11-14

GROSS NATIONAL PRODUCT, 1929-1956
Current Dollars and 1947 Dollars



Source of data Table 11-4

Table 11-4

GROSS NATIONAL PRODUCT: 1929 TO 1956

CURRENT DOLLARS AND 1947 DOLLARS

Year	Current dollars (Billions)	1947 dollars (Billions)	Year	Current dollars (Billions)	1947 dollars (Billions)
1929.....	104.4	149.3	1944.....	211.4	268.2
1930.....	91.1	135.2	1945.....	213.6	263.1
1931.....	76.3	126.6	1946.....	209.2	233.8
1932.....	58.5	107.6	1947.....	232.2	232.2
1933.....	56.0	103.7	1948.....	257.3	243.9
1934.....	65.0	113.4	1949.....	257.3	241.9
1935.....	72.5	127.8	1950.....	285.1	264.7
1936.....	82.7	142.5	1951.....	328.2	282.9
1937.....	90.8	153.5	1952.....	345.4	293.3
1938.....	85.2	145.9	1953.....	363.2	305.3
1939.....	91.1	157.5	1954.....	360.7	300.8
1940.....	100.6	171.6	1955.....	390.9	322.4
1941.....	125.8	198.2	1956.....	412.4	330.3
1942.....	159.1	223.6			
1943.....	192.5	248.9			

SOURCE: *Survey of Current Business*, July, 1957.

STUDY QUESTIONS

1. Why is the classification of statistical data with respect to time important to the business executive?

2. Give an example of a situation in which the measurement of the seasonal variation in business would make possible a better decision by a businessman.

3. Give three examples of periodic fluctuations in business other than variations related to the seasons. (These examples should be in addition to those given in the text.)

4. Some businesses offer price reductions to stimulate business during periods of the day, week, or year when volume is abnormally low. Give three examples of this type of pricing.

5. Give an example of an industry in which the measurement of secular trend would be important to the businessman.

6. How can the semilogarithmic chart be used to compare the growth of two industries?

7. Why does the business executive need information on the cyclical fluctuations of business?

8. A number of years ago a businessman was shown the usefulness of a semilogarithmic chart for the analysis of economic data, and he was so much impressed that he decided to use only graphs drawn on this type of grid. Do you consider this decision wise? Why or why not?

9. Why is it logical to plot two curves representing series in different units (such as production and prices) on the same semilogarithmic chart? Why is it not so satisfactory to plot two such curves on an arithmetic chart?

PROBLEMS

1. The following table gives the production and cold storage holdings of American cheese in the United States during 1956.

(a) Plot these series on a semilogarithmic chart.
(b) What seasonal fluctuations do you observe in these two series?
(c) Give an explanation of why these seasonal fluctuations appear in the production and cold storage stocks of cheese. Assume that the industry is on a five-day week.

FACTORY PRODUCTION AND COLD STORAGE STOCKS OF AMERICAN CHEESE, 1956

(IN THOUSANDS OF POUNDS)

Month	Production	Stocks, end of month
January	96,775	496,746
February	97,025	464,397
March	119,070	460,421
April	129,185	456,279
May	154,465	484,154
June	158,420	524,505
July	134,090	551,334
August	116,970	554,518
September	102,445	533,107
October	98,010	489,385
November	90,700	454,729
December	96,240	441,082

SOURCE: *Survey of Current Business*, March, 1957, page S-27.

2. The following table gives the production of shoes in the United States during the year 1956, during which time the industry operated on a five-day week.

(a) Adjust the production for calendar variation and plot the monthly production and the production per working day on the same semilogarithmic chart.
(b) Explain the difference between the two series.

PRODUCTION OF SHOES, SANDALS, AND PLAY SHOES (EXCEPT ATHLETIC), 1956

(THOUSAND PAIRS)

Month	Production	Month	Production
January	49,356	July	36,856
February	51,863	August	46,469
March	50,077	September	37,189
April	43,727	October	42,183
May	42,314	November	38,407
June	38,751	December	37,107

SOURCE: *Survey of Current Business*, March, 1957, page S-31.

3. The table below gives the number of revenue passenger miles on Class I steam railways and on scheduled airlines in the United States for the years 1939 to 1956.

(a) Plot these two series on the same semilogarithmic chart.

(b) Summarize what this chart shows about the trends of the two modes of travel.

REVENUE PASSENGER MILES OF CLASS I STEAM RAILWAYS AND SCHEDULED AIRLINES: 1939 TO 1956

(IN BILLIONS OF PASSENGER MILES)

Year	Class I steam railways	Scheduled airlines	Year	Class I steam railways	Scheduled airlines
1939	22.6	.7	1949	35.1	6.6
1940	23.8	1.0	1950	31.8	7.8
1941	29.4	1.4	1951	34.6	10.2
1942	53.7	1.4	1952	34.0	12.1
1943	87.9	1.6	1953	29.3	16.2
1944	95.6	2.2	1954	29.3	16.2
1945	91.8	3.3	1955	28.5	19.2
1946	64.7	5.9	1956	28.2	21.6
1947	45.9	6.0			
1948	41.1	5.8			

SOURCE: *Business Statistics*, 1957 Edition.

4. The following table gives annual disposable personal income and personal consumption expenditures for the years 1929 to 1956.

(a) Plot these two series on the same arithmetic chart.
(b) Summarize briefly what this chart shows concerning the cyclical fluctuations in income and expenditures over this period.

DISPOSABLE PERSONAL INCOME AND PERSONAL CONSUMPTION EXPENDITURES: 1929 TO 1956

(IN BILLIONS OF DOLLARS)

Year	Disposable personal income	Personal consumption expenditures
1929	83.1	79.0
1930	74.4	71.0
1931	63.8	61.3
1932	48.7	49.3
1933	45.7	46.4
1934	52.0	51.9
1935	58.3	56.3
1936	66.2	62.6
1937	71.0	67.3
1938	65.7	64.6
1939	70.4	67.6
1940	76.1	71.9
1941	93.0	81.9
1942	117.5	89.7
1943	133.5	100.5
1944	146.8	109.8
1945	150.4	121.7
1946	159.2	146.6
1947	169.0	165.0
1948	187.6	177.6
1949	188.2	180.6
1950	206.1	194.0
1951	226.1	208.3
1952	237.4	218.3
1953	250.2	230.5
1954	254.5	236.6
1955	270.2	254.4
1956	287.2	267.2

SOURCE: *National Income*, 1954 Edition. *Survey of Current Business*, July, 1957.

5. A machine with an expected life of five years is to be depreciated at a uniform rate on diminishing value. The original cost of the machine was $10,000 and it is estimated that it will be worth $200 at the end of five years.

(a) Compute the credit to the reserve for depreciation each year.

(b) What difficulty would be encountered if the machine were estimated to have no value at the end of five years?

6. The following table shows personal consumption expenditures in the United States for the years 1947 to 1956. During this period prices were rising, with the result that some of the increase in consumer expenditures reflected a price increase rather than an increase in the volume of goods bought by consumers. The table also gives the implicit price deflator, which is a measure of the change in the prices of goods and services purchased by consumers.

(a) Compute personal consumption expenditures in 1947 dollars for the years 1947 to 1956.

(b) Plot the two series, consumer expenditures in current dollars and in 1947 dollars, on the same arithmetic chart.

(c) Write a description of what the chart shows concerning the changes in the amount of money consumers spent and the volume of goods and services they secured over the ten-year period.

PERSONAL CONSUMPTION EXPENDITURES AND IMPLICIT PRICE DEFLATOR: 1947 TO 1956

Year	Expenditures (Millions of dollars)	Deflator (Index number, 1947 = 100)
1947	164,973	100.0
1948	177,609	105.7
1949	180,598	104.8
1950	194,026	106.2
1951	208,342	113.5
1952	218,328	115.3
1953	230,542	116.8
1954	236,557	117.7
1955	254,421	118.1
1956	267,160	120.1

SOURCE: *Survey of Current Business*, July, 1957.

12 SECULAR TREND—THE STRAIGHT LINE

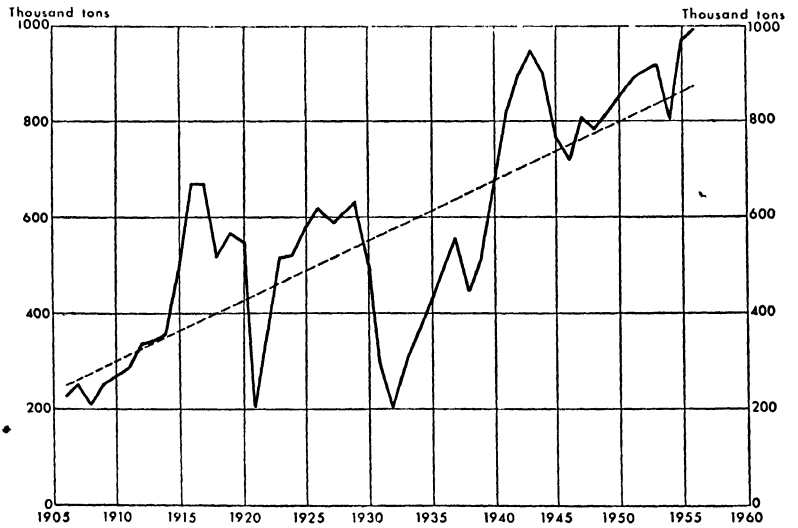
A brief discussion of the nature of secular trend was given in Chapter 11, and this chapter will be devoted primarily to the measurement of secular trend. Some additional discussion of the nature of long-term movements of statistical series will be given first as an introduction to the methods of measuring trend.

NATURE OF TREND

The secular trend of a time series is probably best represented graphically by a straight line or a smooth curve, since growth or decay may be assumed to proceed gradually without abrupt changes in direction. A straight line on an arithmetic chart indicates that the *amount* of change was the same each year. Zinc production shown in Chart 12-1 is an example of a series that has grown approximately the same amount each year. A

Chart 12-1

SMELTER PRODUCTION OF PRIMARY SLAB ZINC IN THE UNITED STATES 1906-1956



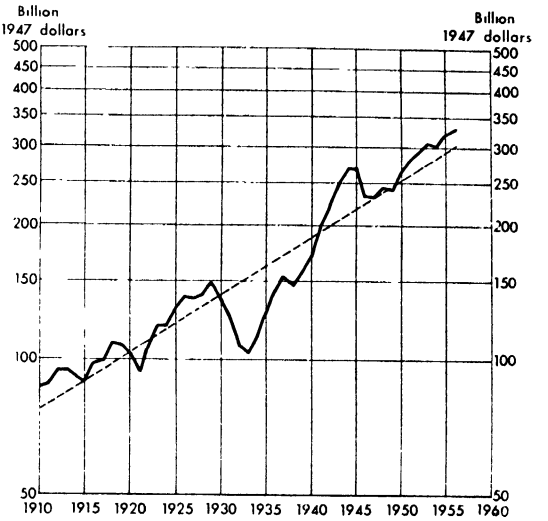
Sources of data 1906-1945 *Historical Statistics of the United States* 1946-1956 *Business Statistics, 1957*

straight line on a semi-logarithmic chart indicates that the *rate* of change was the same each year. Chart 12-2 shows that gross national product has grown at a constant annual rate of about 3 per cent.

Industries producing goods that are being replaced by competitive products may decline, as the trend in the consumption of anthracite coal shows in Chart 12-3 where two measurements of the trend are depicted. In this chart each straight-line trend declines the same amount each year.

Chart 12-2

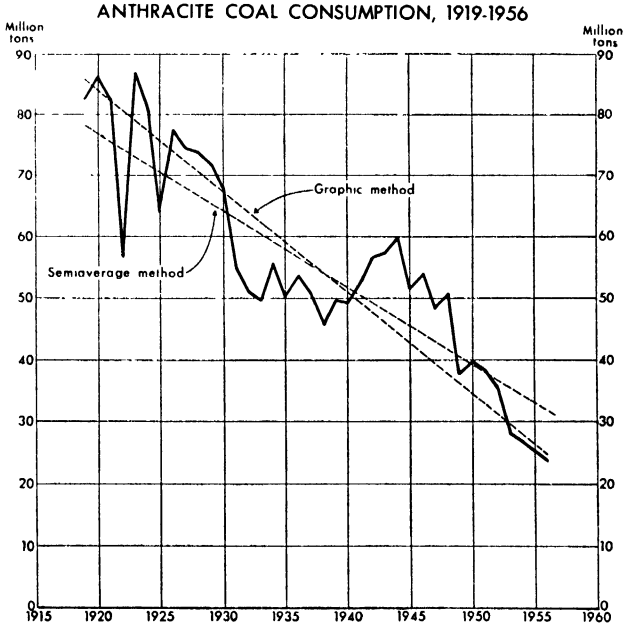
GROSS NATIONAL PRODUCT, 1910-1956



Sources of data 1910-1928 *Economic Almanac*
1929-1956 *Survey of Current Business, July, 1957*

Chart 12-3

ANTHRACITE COAL CONSUMPTION, 1919-1956



Source of data Tables 12-1 and 12-2

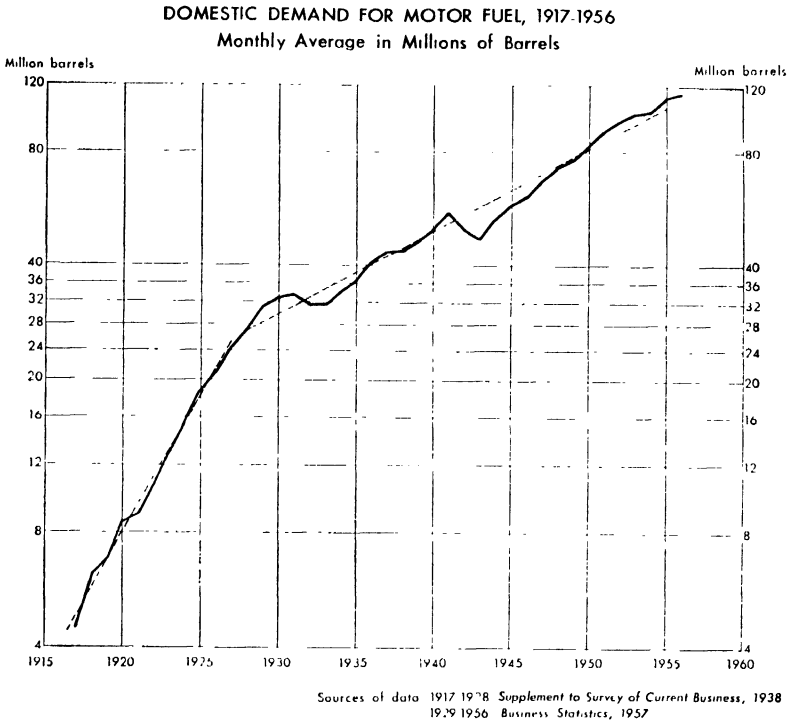
If neither the arithmetic nor the semilogarithmic straight line describes the underlying movement of the series, a curve on either an arithmetic or a semilogarithmic chart might be used. It is generally difficult for a constant rate of growth to be maintained over a long period of time; the more common pattern is for the rate of growth to decrease gradually. This means that the trend line on a semilogarithmic chart rises less and less sharply until it approaches a horizontal line. Chart 13-4 (page 306) shows the per capita consumption of rayon and acetate fibers. The secular trend has been rising at a decreasing rate; and although the trend is still rising, it would not be unreasonable to conclude that it is approaching the point where there will be no further growth in per capita consumption. This would mean that the market for rayon and acetate fibers had become practically saturated, although the trend in total consumption would continue upward as long as the population continued to grow.

Since the rate of growth usually changes slowly, several straight lines can often be used, each one covering a portion of the whole period. For example, it can be seen in Chart 12-4 that a straight line describes the trend in motor fuel demand reasonably well for the period 1917 to 1927; but from 1928 to 1956 the rate of growth was considerably less, and another straight line with a different slope was used. For the first period the growth was at the annual rate of 17.7 per cent, and for the second period, 5.2 per cent. Instead of using two lines with different slopes, it would be possible to use a curve that rose at a gradually decreasing rate. Probably it is more logical to describe the growth in motor fuel demand by a trend line that increases at a decreasing rate, but for many purposes it is desirable to have a measure of the average annual rate of growth for a shorter period. This is a useful summary device and makes possible a comparison of the rates of growth in the different periods. It will be seen also in the following discussion that the work of measuring the trend is much easier when a straight line is used instead of a curve.

In each of the charts shown in this chapter the trend line cuts between the high and low points of the line plotted from the actual data, with the result that the average of the trend values for a given series is approximately equal to the average of the actual values. In other words, the trend line represents the course the series might have taken if disturbing influences had not pushed it higher in some periods and lower in others. In this sense the trend represents an average value for the series, balancing high and low cyclical and seasonal values.

A different concept sometimes advanced is that the secular trend should represent full capacity, and deviations from trend would occur when operations fall below full capacity. Instead of the line representing actual data fluctuating above and below the trend line, the actual data line would

Chart 12-4



appear to hang from the trend line in loops. Periods of operation at full capacity would push the data line on a chart up to the trend line, and any operations at less than full capacity would pull the line below the trend. This concept of trend as full capacity is much less common than the treatment of trend as an average that cuts through the high and low points of the series. In the analysis that follows, it will be assumed that the line describing the trend graphically will cut between the high and low points of the data.

CHARACTERISTICS OF A STRAIGHT LINE

The straight-line trend will be represented by the equation

$$Y_c = a + bX. \quad \ldots\ldots\ldots\ldots \quad (32)$$

where a is the Y intercept, or the value of the Y variable when $X = 0$, and b represents the slope of the line, or the amount of change in the Y variable that is associated with a change of one unit in the X variable. Y_c is used to designate the trend values to distinguish the trend values from the actual Y values. Since the time series is plotted using time as the X variable, the

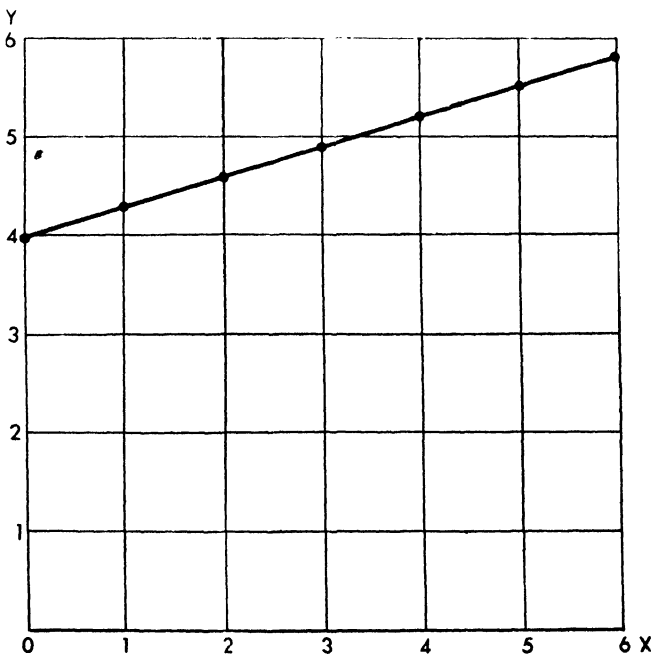
origin of the X axis is the year zero. Obviously it is inconvenient to use an origin so far from the present, so it is normal practice to choose a recent year as the origin and assign it the value of zero. If 1900 were chosen as the origin, 1900 would have a value of "0," 1901 would be "1," and so on. It seems logical to place the origin at the first year in the series, but there are situations in which some other year is more convenient to use in making calculations. The choice of origin will have no effect on the results, so it is governed entirely by the convenience of its use in making calculations.

When the years have been assigned values, beginning with zero, the graph of a linear trend is plotted as in Chart 12-5. For the equation $Y_c = 4.0 + .3X$, the following tabulation gives the values of Y that are associated with values of X from 0 to 6. These values are shown graphically by the straight line on Chart 12-5.

X	Y
0	4.0
1	4.3
2	4.6
3	4.9
4	5.2
5	5.5
6	5.8

Chart 12-5

A STRAIGHT LINE



MEASURING SECULAR TREND

In this section, the following methods of measuring secular trend will be described:

1. Graphic method.
2. Method of semiaverages.
3. Method of least squares.

Graphic Method

The simplest method of determining the trend values of a time series is to draw through the data a straight line that appears to describe the underlying, long-term movement in the series and ignores the movements of a cyclical nature that reverse themselves after a short period. If the trend is being determined from annual data, there will be no seasonal fluctuations to obscure the underlying trend movement. If monthly data were used, it would be necessary to avoid being influenced by the regularly recurring seasonal fluctuations; but the measurement of secular trend seldom makes use of data classified on the basis of a time period shorter than a year.

Chart 12-3 illustrates a linear trend computed graphically. The annual consumption of anthracite coal was plotted on the arithmetic chart and the trend line drawn on the chart so that it followed the gradual decline in consumption, following a path below the peaks of consumption but above the low years. It is essential that the trend line follow the underlying course of the actual data without being influenced by the forces that cause the alternating periods of expansion and contraction. Since anthracitic coal consumption appears to have declined by the same amount per year throughout the period since 1919, a straight line is a good description of the underlying movement in the series.

The trend value of 85.5 for 1919 can be read off the chart. The value of the trend in the origin year is the a value in equation (32); so by letting 1919 be the origin, $a = 85.5$. Again reading off the chart, it is found that the trend line was 24.8 in 1956, a decline of 60.7 Y units (millions of tons of anthracite coal) in 37 years. The decline per year was 1.64 ($60.7 \div 37 = 1.64$). This means that $b = -1.64$, and with this information, the equation of the trend line is summarized as follows:

$$Y_c = 85.5 - 1.64X$$

Origin: 1919.
X unit: 1 year.
Y unit: Millions of tons of anthracite coal.

From the equation of the trend line, the trend value for any year can be computed by substituting the X value representing the year for which the

trend value is wanted. For example, the trend value for 1929 is computed as follows, since $X = 10$ for the year 1929:

$$Y_c = 85.5 - (1.64)(10) = 85.5 - 16.4 = 69.1$$

For 1956, $X = 37$, and the trend value is as follows:

$$Y_c = 85.5 - (1.64)(37) = 85.5 - 60.7 = 24.8$$

The trend value for each year from 1919 to 1956 may be read off the chart or computed from the trend equation. These trend values are given in Table 12-1.

Table 12-1

ANTHRACITE COAL CONSUMPTION IN THE UNITED STATES: 1919 TO 1956

TREND VALUES COMPUTED BY THE GRAPHIC METHOD

Year (1)	(2) X	Consumption (Million tons) (3) Y	Trend values (Million tons) (4) Y_c	Year (1)	(2) X	Consumption (Million tons) (3) Y	Trend values (Million tons) (4) Y_c
1919....	0	81.5	85.5	1939....	20	49.7	52.7
1920....	1	85.8	83.9	1940....	21	49.0	51.1
1921....	2	82.0	82.2	1941....	22	52.7	49.4
1922....	3	56.8	80.6	1942....	23	56.5	47.8
1923....	4	86.9	78.9	1943....	24	57.1	46.1
1924....	5	80.7	77.3	1944....	25	59.4	44.5
1925....	6	64.1	75.7	1945....	26	51.6	42.9
1926....	7	77.2	74.0	1946....	27	53.9	41.2
1927....	8	74.7	72.4	1947....	28	48.2	39.6
1928....	9	73.7	70.7	1948....	29	50.2	37.9
1929....	10	71.5	68.1	1949....	30	37.7	36.3
1930....	11	67.6	67.5	1950....	31	39.9	34.7
1931....	12	58.4	65.8	1951....	32	37.0	33.0
1932....	13	50.5	64.2	1952....	33	35.3	31.4
1933....	14	49.6	62.5	1953....	34	28.0	29.7
1934....	15	55.5	60.9	1954....	35	26.9	28.1
1935....	16	51.1	59.3	1955....	36	22.3	26.5
1936....	17	53.2	57.6	1956....	37	23.8	24.8
1937....	18	50.4	56.0				
1938....	19	45.2	54.3	Total...	...	2,095.6	...

SOURCE: Bureau of Mines.

The graphic determination of a trend line is the simplest method that can be used, and there are many cases in which a satisfactory trend line can be located in this manner. This is particularly true when the deviations from the underlying trend are small, with the result that the path of the trend is clearly defined and very little doubt exists as to its proper location.

Instead of locating the line entirely by inspection, the average of the original data may be used as the trend value at the middle of the time period. It appears to be a reasonable condition that the average of the actual data should equal the average of the trend values. The average of the straight-line values will be the value midway between the first and the last values. This device was used in locating the trend line graphically in Chart 12-3. The average yearly consumption of anthracite coal for the 38-year period was 55.1 million tons, and this was plotted as the trend value at the middle of the time period. Since there were 38 years, 1919 to 1956, the middle of the period was between 1937 and 1938. With one point on the line established in this manner, it was necessary to decide only upon the slope of the trend line, with the result that one subjective element in the graphic solution was eliminated.

Method of Semiaverages

Even when the graphic method locates the average level of the trend line at the average of the actual data, the slope is determined by the judgment of the person drawing the line. Since one of the major contributions of statistical methods is the substitution of objective measurements for subjective judgments, the graphic method is not held in high regard. The use of the average of actual data to locate one point on the trend line substitutes measurement for one subjective decision, but an element of judgment still remains in the determination of the slope of the line. The method of semiaverages uses a simple objective method to compute the slope of the line as well as to locate its level.

The data for which the trend values are to be computed are divided into two equal periods, and the average is computed for each period. The annual consumption of anthracite coal from 1919 to 1956 shown in Table 12-2 is divided into two periods of 19 years each, 1919 to 1937 and 1938 to 1956. The total consumption for the first half is designated S_1; the total consumption for the second period is designated S_2. The average annual consumption for each period is computed by dividing each of the subperiod totals by 19, the number of years in each of the subperiods. The average annual consumption for the first period of 19 years was 66.9 million tons and for the second period, 43.4 million tons. The middle year of the first period is 1928, and the trend value is the average for the period, 66.9 million

tons. The average for the second period, 43.4 million tons, is the trend value for the middle year of the second period, 1947.

With two points on the straight line located, it is a matter of simple arithmetic to compute the slope of the line. The trend fell from 66.9 in 1928 to 43.4 in 1947, a decline of 23.5 million tons. Since the decline occurred over a period of 19 years, the annual decline was 1.24 million tons. This is the b value of the trend equation, and is computed as follows:[1]

$$b = \frac{43.4 - 66.9}{19} = \frac{-23.5}{19} = -1.24$$

Either 1928 or 1947 can be used as the origin; if 1928 is designated the origin, $a = 66.9$, since the trend value for 1928 is 66.9. With 1928 the origin, the equation of the trend is

$$Y_c = 66.9 - 1.24X$$

Origin: 1928.
X unit: 1 year.
Y unit: Millions of tons of anthracite coal.

Since 1928 is the origin on the time scale (X axis), the years prior to 1928 are minus; for example, 1927 is -1; 1926 is -2; and so on. To compute the trend values, the proper value of X is substituted in the trend equation and the value of Y_c determined. The trend value for each year is given in Table 12-2.

The trend values computed by the method of semiaverages are plotted on Chart 12-3. It will be noted that the trend at the middle of the period is the same as that established by the graphic method, but that the slope of the line is somewhat steeper for the graphic method. There is no conclusive method for deciding which slope best describes the long-term movement of the series, and the final decision is a matter of judgment as to the line that best fits the data in the sense of describing the underlying long-term movement. The obvious weakness of the graphic method of locating the trend is that it does leave the slope of the line to the judgment of the one who locates it, although using the average for the total period as the trend value at midrange locates the level of the line in an objective manner.

[1]The following formula may be used to compute the slope of the trend line. It is to be used only when the period is an even number of years.

S_1 = sum of the Y values for the first half of the period.

S_2 = sum of the Y values for the second half of the period.

$$b = \frac{4(S_2 - S_1)}{N^2} = \frac{4(824.4 - 1{,}271.2)}{38^2} = \frac{-1{,}787.2}{1{,}444} = -1.24$$

Table 12-2

ANTHRACITE COAL CONSUMPTION IN THE UNITED STATES: 1919 TO 1956

TREND VALUES COMPUTED BY METHOD OF SEMIAVERAGES

Year (1)	(2) X	Consumption (Million tons) (3) Y	Subperiod totals (4)	Trend values (Million tons) (5) Y_c	Year (1)	(2) X	Consumption (Million tons) (3) Y	Subperiod totals (4)	Trend values (Million tons) (5) Y_c
1919	−9	81.5		78.1	1938	10	45.2		54.5
1920	−8	85.8		76.8	1939	11	49.7		53.3
1921	−7	82.0		75.6	1940	12	49.0		52.0
1922	−6	56.8		74.3	1941	13	52.7		50.8
1923	−5	86.9		73.1	1942	14	56.5		49.5
1924	−4	80.7		71.9	1943	15	57.1		48.3
1925	−3	64.1		70.6	1944	16	59.4		47.1
1926	−2	77.2		69.4	1945	17	51.6		45.8
1927	−1	74.7		68.1	1946	18	53.9		44.6
1928	0	73.7		66.9	1947	19	48.2		43.4
1929	1	71.5		65.7	1948	20	50.2		42.1
1930	2	67.6		64.4	1949	21	37.7		40.9
1931	3	58.4		63.2	1950	22	39.9		39.6
1932	4	50.5		61.9	1951	23	37.0		38.4
1933	5	49.6		60.7	1952	24	35.3		37.1
1934	6	55.5		59.5	1953	25	28.0		35.9
1935	7	51.1		58.2	1954	26	26.9		34.7
1936	8	53.2		57.0	1955	27	22.3		33.4
1937	9	50.4	$S_1 = 1{,}271.2$	55.7	1956	28	23.8	$S_2 = 824.4$	32.2
					Total	...	2,095.6	...	...

SOURCE: Bureau of Mines.

The method of semiaverages is more objective in that the slope of the line is determined by the difference between the average of the first half of the series and the average of the last half. This method of determining the slope of the best-fitting trend line relies on the use of the average of each half of the series as typical of the level of that portion of the data. The discussion of averages in Chapter 8 established this principle, and so it seems logical to accept the method of semiaverages as a valid one for computing the secular trend of a series in which the change with respect to time is linear.

Method of Least Squares

The method of least squares is a widely used method of fitting a curve to data and is the most popular method of computing the secular trend of time series. The method locates the trend value for the middle of the time period at the average for the data to which the trend line is being fitted. This means that the least squares method gives the same value at midrange as the two methods previously described. The slope of the line, however, is computed differently, so all of the other trend values will be different from those given by the other two methods.

The basic problem of selecting a method of fitting a trend line is deciding on a criterion for measuring goodness of fit. If the points to which the line is being fitted will not all fall on a straight line, no straight line fits perfectly; so it becomes a problem of defining what is meant by the line of "best fit." The actual values of the data will deviate from the trend values in all or most of the years. One criterion that might be used to select the best-fitting line might be: Is the sum (disregarding signs) of the deviations of the actual values from the trend values a minimum? If so, this line might be regarded as the line of best fit since, in total, the trend values come closer to the actual values than any other line.

Another criterion might be: Is the sum of the squared deviations a minimum? This test of best fit would have an advantage over the criterion that minimizes the sum of the deviations without regard to signs since it is not strictly logical to ignore the signs. This is similar to the situation encountered in choosing between the average deviation and the standard deviation as a measure of dispersion. It was stated in Chapter 9 that ignoring the signs of the deviations from the mean was illogical, and a similar statement seems appropriate in considering the deviations from a trend. The most common criterion for selecting a straight line to show the trend is: Is this the line from which the squared deviations of the actual values from trend values are a minimum? The name *least squares* is derived from the use of this criterion.

The method of least squares gives the most satisfactory measurement of the trend of a series when the distribution of the deviations is approximately normal; but judgment must be used in deciding that such conditions are present. If there are some extremely large deviations from the trend, the least squares method will tend to give too much weight to these deviations and result in a trend that does not follow the general course of the series. Since time series are subject to extreme fluctuations, the method of least squares should not be considered a logical method that can be used in every situation. The only real criterion for the selection of a method of measuring trend is the judgment as to how well the trend line follows

the general movement of the series. However, when the least squares method is appropriate, it is probably the most satisfactory method of making the calculations. The least squares method also has the virtue of being impersonal, a feature that some people consider to be very necessary.

The computations needed to determine the straight-line trend by the method of least squares have been reduced to formulas. The use of these formulas is illustrated below by computing the trend for the set of numbers in Table 12-3 that constitutes a time series covering eleven years. The data represent the domestic demand for motor fuel in the United States for the years 1946 to 1956. The data and the least squares straight-line trend are shown graphically in Chart 12-6. The trend during these eleven years was upward in approximately a straight line, with very little deviation from a straight line.

When using the method of least squares to compute a trend line, it is convenient to use the middle of the time series as the origin. If the series consists of an odd number of years, the origin is at the middle of the middle year; if an even number of years is used, the origin falls between the two middle years. Although the method works equally well for both situations, the details of calculation for an even number of years differ slightly from those used for an odd number of years. Following the computation of a trend line for an odd number of years, an illustration will be given for the method applied to an even number of years using the 12 years, 1945 to 1956.

Odd number of years. The problem of fitting a trend by the method of least squares requires the determination of the values a and b for the equation $Y_c = a + bX$ that satisfy the criterion that the sum of the squared differences between the Y_c and the Y values be a minimum.

When the origin is taken at the midrange for the series, the following formula gives the value of a:

$$a = \frac{\Sigma Y}{N} \quad \ldots\ldots\ldots\ldots \quad (33)$$

This is the mean of the values for which the trend is being computed, and gives the same trend value at the middle of the series as the two methods discussed previously.

For the motor fuel series given in Table 12-3, the trend value at the origin (1951) is 88.6 (remembering that the original data were given in millions of barrels).

$$a = \frac{\Sigma Y}{N} = \frac{975}{11} = 88.64$$

Table 12-3

COMPUTATION OF STRAIGHT-LINE TREND (ODD NUMBER OF YEARS)

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Year (1)	(2) X	(3) X^2	Average monthly demand (Million bbls.) (4) Y	(5) XY	Trend values (Million bbls.) (6) Y_c
1946	− 5	25	61.3	− 306.5	61.5
1947	− 4	16	66.3	− 265.2	66.9
1948	− 3	9	72.6	− 217.8	72.4
1949	− 2	4	76.1	− 152.2	77.8
1950	− 1	1	82.9	− 82.9	83.2
1951	0	0	90 8	0	88.6
1952	1	1	96.4	96.4	94.1
1953	2	4	100.5	201.0	99.5
1954	3	9	103.2	309.6	104.9
1955	4	16	110.1	440.4	110.3
1956	5	25	114.8	574.0	115.8
Total	...	110	975.0	596.8	975.0

SOURCE: *Business Statistics, 1957.*

The computation of the slope of the trend line that gives the minimum squared deviations is made by the following equation when the origin of the X variables is at midrange:

$$b = \frac{\Sigma XY}{\Sigma X^2} \quad \ldots\ldots\ldots\ldots \quad (34)$$

In the discussion of the characteristics of the straight line on page 280, it was pointed out that it is convenient to use a recent year as "0" rather than using the years as the values of the X variable. In computing the least squares trend values, the origin on the X scale (the point where $X = 0$) is located at midrange. In this problem involving an odd number of years, the origin is the middle year (1951 = 0). The year 1950 is − 1, 1949 is − 2, etc. The years since 1951 are + 1, + 2, etc. These values are shown in Table 12-3 in Column (2) designated X. The X values are an arithmetic progression (in this case with a common difference of 1) since time is an arithmetic progression. It will be noted that the sum of the X values is always zero.

ΣXY is the sum of the products of the individual X and Y values in Table 12-3, computed in Column (5); ΣX^2 is the sum of the values in Column (3), which were computed by squaring the values of X. These amounts[2] were substituted in equation (34) and the value of b computed as follows:

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{596.8}{110} = 5.425$$

With the a and b values computed, the trend equation may be written

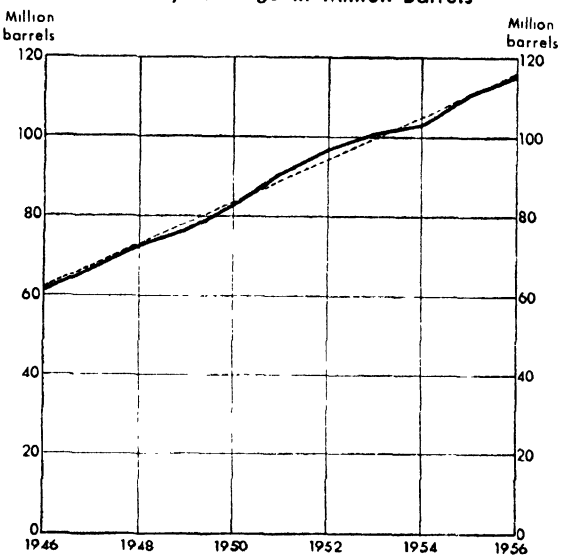
$$Y_c = 88.64 + 5.425X$$

Origin: July 1, 1951.
X unit: 1 year.
Y unit: Average monthly demand for motor fuel (millions of barrels).

Since the trend value for July 1, 1951, is 88.6, the trend value a year later will be 5.425 more, or 94.1. Each successive trend value later than 1951, the origin, is found by adding 5.425; and each trend value earlier than 1951, by subtracting 5.425. The trend values computed and entered in Column (6) of Table 12-3 are plotted in Chart 12-6, on which the original data are also plotted. Instead of computing the trend values by successively adding or subtracting the b value, it is possible to compute the trend value for any given year by substituting

Chart 12-6

DOMESTIC DEMAND FOR MOTOR FUEL, 1946-1956
Monthly Average in Million Barrels



Source of data: Table 12-3

[2]The following formula may be used to compute the value of ΣX^2, with N equal to the number of years to which the trend is fitted. This formula can be used only when the trend is being fitted to an odd number of years.

$$\Sigma X^2 = \frac{N(N^2-1)}{12}$$

When $N = 11$, the following computation gives the value of ΣX^2.

$$\Sigma X^2 = \frac{11(121-1)}{12} = 110$$

the proper X value in the trend equation. For example, the trend value for 1946 may be computed as follows:

$$\text{Since } X = -5$$
$$Y_c = 88.64 + (5.425)(-5) = 61.5$$

Even number of years. When the number of years to which the trend line is to be fitted is even, only a slight modification of the device given in Table 12-3 is necessary. The midrange now falls between two years instead of on the middle of the middle year. The arbitrary values assigned to the years are odd numbers, which makes the common difference 2 instead of 1. Column (2) of Table 12-4 shows these values: 1945 is -11, 1946 is -9, and so on. The origin, 0, is between 1950 and 1951, or January 1, 1951.

Table 12-4

COMPUTATION OF STRAIGHT-LINE TREND (EVEN NUMBER OF YEARS)

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1945 TO 1956

Year (1)	(2) X	(3) X^2	Average monthly demand (Million bbls.) (4) Y	(5) XY	Trend values (Million bbls.) (6) Y_c
1945	−11	121	58.0	−638.0	56.6
1946	− 9	81	61.3	−551.7	62.0
1947	− 7	49	66.3	−464.1	67.3
1948	− 5	25	72.6	−363.0	72.7
1949	− 3	9	76.1	−228.3	78.1
1950	− 1	1	82.9	− 82.9	83.4
1951	1	1	90.8	90.8	88.8
1952	3	9	96.4	289.2	94.1
1953	5	25	100.5	502.5	99.5
1954	7	49	103.2	722.4	104.8
1955	9	81	110.1	990.9	110.2
1956	11	121	114.8	1,262.8	115.5
Total	...	572	1,033.0	1,530.6	1,033.0

SOURCE: *Business Statistics, 1957.*

The arithmetic mean of the Y values is the trend value at the origin, or January 1, 1951, in this case. Therefore,

$$a = \frac{1{,}033.0}{12} = 86.08$$

The value of b is found by using equation (34). It should be noticed that the common difference of the arithmetic progression assigned to the X values is 2, which means that one X equals one-half year. $\Sigma XY = 1{,}530.6$, and $\Sigma X^2 = 572$; so the value of b is computed as follows[3]:

$$b = \frac{1{,}530.6}{572} = 2.676$$

The trend equation is:

$$Y_c = 86.08 + 2.676X$$

Origin: January 1, 1951.

X unit: ½ year.

Y unit: Average monthly demand for motor fuel (million of barrels).

The calculation of the trend for an even number of years is the same as for any other straight line. The individual X values are substituted in the trend equation and the Y_c value computed. For 1951, the X value is 1 and the following calculation gives the trend value:

$$Y_c = 86.08 + (2.676)\,(1) = 88.76$$

The X value for 1952 is 3 and the calculation of the trend value is made by substituting the value 3 in the trend equation as follows:

$$Y_c = 86.08 + (2.676)\,(3) = 86.08 + 8.03 = 94.10$$

The trend value for each year can be computed in this manner; although after one value has been determined, the easiest method of computing the remaining values is to add the annual increment for the later years and subtract it for the earlier. The yearly increase in the trend line is two times the b coefficient since b is the amount of change in Y when one X equals one-half year. Successive additions of 5.352 (two times 2.676) to

[3]The following formula may be used to compute the value of ΣX^2, with N equal to the number of years to which the trend is fitted. This formula can be used only when the trend is being fitted to an even number of years.

$$\Sigma X^2 = \frac{N(N^2-1)}{3}$$

When $N = 12$, the following computation gives the value of ΣX^2.

$$\Sigma X^2 = \frac{12\,(144-1)}{3} = 572$$

88.76, the trend value for 1951, give the trend value for the years 1951 to 1956. Successive subtractions of 5.352 from the 1951 trend value give the values for earlier years.

The computation of the straight line for an even number of years used data on demand for motor fuel for one year more than was used for computing the trend in Table 12-3. A comparison of the trend values in Tables 12-3 and 12-4 shows that the additional year makes very little difference. Since the trend values computed from the data for 12 years differ so little from those plotted in Chart 12-6, another chart is not shown.

USE OF LINEAR TREND

A measure of secular trend gives useful information about the change in the statistical series being studied, if the primary interest is in the long-term changes represented by the trend line. This interest may be either in determining how the series has grown in the past, or in making a forecast of the secular trend. In both cases as accurate a measure of the trend as possible is needed, since a forecast must necessarily be based on the record of past change. Many industries, particularly those that require a large fixed investment of capital, need a forecast of trend in making capital budgets and the subject of forecasting secular trend is considered in Chapter 21. Many trend forecasts, however, are not linear; in fact, it is quite likely that any measurement or forecast of trend that covers any considerable period will depict a change in amount of growth experienced per year. This means that a nonlinear trend should be used instead of a straight line. A series of straight lines can be fitted to segments of the total period, if a single straight line will not describe the long-term movement satisfactorily; but the general preference in such situations is to use a curve instead of a straight line.

A different use of the trend line is to adjust a series to take out the effect of the long-term trend in order to isolate nontrend fluctuations, such as the business cycle. A straight line is particularly effective in making this kind of adjustment, since a good fit of the data is all that is required of the trend line. When using a trend to measure long-term change for the purpose of forecasting this movement, it seems reasonable to require that the trend line be a rational description of the long-term movement. For long periods of time, the straight line is frequently difficult to defend as a logical formula to measure the growth or decline of a business or industry. But when used as a base from which to measure other types of economic change, the straight line is very effective whenever it fits the series reasonably well. Types of nonlinear trends will be described in the following chapter.

STUDY QUESTIONS

1. What is meant by the statement that the measurement of secular trend is in many aspects a problem of averages?

2. What are the advantages and disadvantages of computing a measure of secular trend graphically? arithmetically?

3. Discuss the problem of deciding upon the proper period of time to which a trend line should be fitted.

4. If the trend is being fitted to a time series that has a very pronounced, erratic fluctuation, such as that caused by an extended strike, would you consider leaving this year out of the calculation of the trend? If it did seem advisable to leave such an unusual year out of the calculation, how would you fit a straight line by the method of least squares?

5. What are the advantages and disadvantages of the method of semiaverages for computing a straight-line trend as compared to the method of least squares?

6. Discuss the problem of deciding which of several computed trend lines is the best measure of the long-term growth in a series.

7. What are the major forces that cause economic time series to exhibit a long-term movement that the statistician calls "secular trend?"

8. The method of semiaverages, described for a time series consisting of an even number of years, permits dividing the series into two equal parts. Usually it is possible to vary the period used for fitting the trend enough to insure having an even number of years. However, if it were very important that all of a period consisting of an odd number of years be used, describe how you would fit a trend line by the method of semiaverages.

9. Give an example of a reason why it might be necessary to use an odd number of years in fitting a trend line by the method of semiaverages. (Obviously any series with an odd number of years can be made into one with an even number of years by dropping one year, presumably the earliest.)

PROBLEM

The table on the next page represents the summary from an extensive study on the consumption of raw materials in the United States. These series may be used as the basis for the measurement of secular trend by any of the methods described in this chapter. The instructions that follow may be applied to any method of computing the trend for any of the nine series given in the table.

(a) Plot the series on an arithmetic chart.
(b) Derive the equation of the trend line and compute the trend values for all of the years.
(c) Plot the trend values on the chart with the actual values plotted in (a).
(d) How well does the straight line fit the data?

APPARENT CONSUMPTION OF INDUSTRIAL RAW MATERIALS IN THE UNITED STATES: 1923 TO 1952[a]

(MILLIONS OF 1935—1939 DOLLARS)

Year	Total	Agricultural non-foods	Fishery, wildlife products	Forest products	Total minerals	Iron and ferro-alloys	Other metals	Mineral fuels	Other non-metallic minerals
1923	$6,001	$1,552	$143	$1,220	$3,085	$217	$282	$2,235	$350
1924	5,680	1,528	157	1,179	2,815	176	246	2,044	349
1925	5,953	1,680	154	1,195	2,924	213	267	2,063	382
1926	6,220	1,708	167	1,177	3,168	228	302	2,225	413
1927	5,928	1,438	172	1,137	3,181	205	272	2,276	428
1928	6,133	1,645	190	1,117	3,180	199	296	2,246	440
1929	6,487	1,611	213	1,148	3,515	242	328	2,467	478
1930	5,739	1,456	206	1,016	3,060	193	217	2,219	421
1931	5,380	1,704	195	861	2,620	105	219	1,962	332
1932	4,282	1,270	172	747	2,093	36	111	1,716	230
1933	4,481	1,160	163	794	2,363	66	199	1,846	251
1934	4,272	798	164	802	2,509	81	206	1,939	284
1935	5,159	1,206	199	850	2,903	107	452	2,043	301
1936	5,708	1,256	233	931	3,288	185	376	2,307	420
1937	6,954	2,194	234	967	3,558	232	396	2,476	455
1938	5,686	1,601	196	889	3,000	96	323	2,178	403
1939	6,316	1,815	218	952	3,331	180	321	2,360	470
1940	7,145	2,028	234	991	3,892	253	453	2,675	510
1941	8,223	2,380	243	1,092	4,508	355	640	2,890	623
1942	8,326	2,439	185	1,100	4,602	400	588	2,940	674
1943	7,948	1,987	210	1,042	4,709	402	596	3,097	615
1944	8,451	2,399	236	1,033	4,783	358	499	3,349	577
1945	7,936	1,977	248	977	4,734	335	466	3,361	573
1946	8,123	2,130	296	1,083	4,615	274	338	3,347	656
1947	8,601	2,142	238	1,117	5,104	300	422	3,630	752
1948	9,788	2,786	270	1,175	5,558	376	477	3,893	812
1949	8,821	2,392	256	1,074	5,100	329	502	3,489	780
1950	9,488	2,166	256	1,212	5,853	415	578	3,944	916
1951	9,848	2,148	241	1,220	6,239	472	504	4,238	1,024
1952	10,514	2,638	237	1,242	6,398	431	641	4,256	1,070

[a]Apparent consumption is defined as primary production plus imports minus exports. Agricultural foods and gold are excluded. Components will not necessarily add to totals because of rounding.

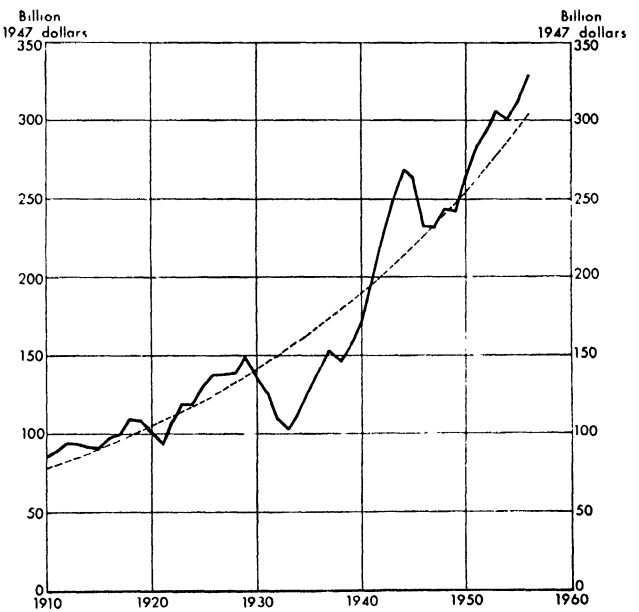
SOURCE: Vivian E. Spencer and Charles A. R. Wardwell, *Raw Materials in the United States Economy, 1900-1952*, U. S. Bureau of the Census, Working Paper No. 1, Preliminary, 1954, pp. 78-79.

13 SECULAR TREND—NONLINEAR

The discussion in Chapter 12 was devoted chiefly to the computation of straight-line trends, but some recognition was given to the fact that many times a straight line will not describe accurately the long-term movement of a time series. Chart 12-2 (page 277) depicted the trend of gross national product as a straight line on semilogarithmic paper, since the growth seemed to be at a constant rate. This same series and the same trend values are shown on an arithmetic grid in Chart 13-1, where the trend is seen to be a curve rather than a straight line. Chart 12-4 (page 279) shows the domestic demand for motor fuel on a semilogarithmic chart. Two straight lines on the semilogarithmic chart are needed to describe the trend in this series: one for the period 1917 to 1927, and the other for the period 1928 to 1956, when the rate of growth was much slower than for the earlier years.

Chart 13-1

GROSS NATIONAL PRODUCT, 1910-1956



Source of data 1910 1928 Economic Almanac
1929 1956 Survey of Current Business, July, 1957

STRAIGHT LINE FITTED TO LOGARITHMS BY METHOD OF LEAST SQUARES

This chapter will describe the methods of computing the logarithmic straight line — a nonlinear trend on an arithmetic chart, but a straight line on a semilogarithmic chart. Since the most common method of computing the line is to fit a straight line to the logarithms of the data, the subject might logically have been discussed in the previous chapter as a linear trend. However, it seems better to consider the linear trend to be a straight line on an arithmetic chart, and so it will be described in this chapter.

When a straight line on a semilogarithmic chart seems to describe the trend of a series, it is possible to compute a trend by fitting a straight line to the logarithms of the data by the method of least squares. Instead of computing the trend to the data (Y), as in Table 12-3 (page 288), the logarithms of the data are found, and a straight line is fitted to the series $\log Y$ instead of to the series Y. The trend values computed by this method will be the logarithms of the trend values and could be related to the logarithms of the data. It is more useful, however, to find the antilogs which will give trend values that can be compared with the actual values.

This method is illustrated in Table 13-1, where a straight line is fitted to the logarithms of the data. The first step is to enter the logarithms of the Y series in Column (5), headed "Log Y," and from this column compute $\Sigma(\log Y)$ and $\Sigma(X \cdot \log Y)$.

With the origin of the equation taken at 1942, the middle year, the formula for the a value will be simply the mean of the logarithms of the Y values.

$$a = \frac{\Sigma(\log Y)}{N} \quad \text{(35)}$$

$$a = \frac{50.23148}{29} = 1.73212$$

The formula for the b coefficient is the following:

$$b = \frac{\Sigma(X \cdot \log Y)}{\Sigma X^2} \quad \text{(36)}$$

$$b = \frac{44.65912}{2{,}030} = .02200$$

$\text{Log } Y_c = 1.73212 + .02200X$
Origin: July 1, 1942.
X unit: 1 year.
Y unit: Average monthly demand (millions of barrels).

Table 13-1

COMPUTATION OF SEMILOGARITHMIC STRAIGHT LINE

AVERAGE MONTHLY DOMESTIC DEMAND FOR MOTOR FUEL: 1928 TO 1956

Year (1)	(2) X	(3) X^2	Average monthly demand (Million barrels) (4) Y	(5) $\log Y$	(6) $X \cdot \log Y$	(7) $\log Y_c$	Trend values (Million barrels) (8) Y_c
1928..	−14	196	27.4	1.43775	−20.12850	1.42412	26.6
1929..	−13	169	31.3	1.49554	−19.44202	1.44612	27.9
1930..	−12	144	32.9	1.51720	−18.20640	1.46812	29.4
1931..	−11	121	33.6	1.52634	−16.78974	1.49012	30.9
1932..	−10	100	31.5	1.49831	−14.98310	1.51212	32.5
1933..	−9	81	31.7	1.50106	−13.50954	1.53412	34.2
1934..	−8	64	34.2	1.53403	−12.27224	1.55612	36.0
1935..	−7	49	36.2	1.55871	−10.91097	1.57812	37.9
1936..	−6	36	40.1	1.60314	−9.61884	1.60012	39.8
1937..	−5	25	43.3	1.63649	−8.18245	1.62212	41.9
1938..	−4	16	43.6	1.63949	−6.55796	1.64412	44.1
1939..	−3	9	46.3	1.66558	−4.99674	1.66612	46.4
1940..	−2	4	49.1	1.69108	−3.38216	1.68812	48.8
1941..	−1	1	55.6	1.74507	−1.74507	1.71012	51.3
1942..	0	0	49.1	1.69108	0	1.73212	54.0
1943..	1	1	47.4	1.67578	1.67578	1.75412	56.8
1944..	2	4	52.7	1.72181	3.44362	1.77612	59.7
1945..	3	9	58.0	1.76343	5.29029	1.79812	62.8
1946..	4	16	61.3	1.78746	7.14984	1.82012	66.1
1947..	5	25	66.3	1.82151	9.10755	1.84212	69.5
1948..	6	36	72.6	1.86094	11.16564	1.86412	73.1
1949..	7	49	76.1	1.88138	13.16966	1.88612	76.9
1950..	8	64	82.9	1.91855	15.34840	1.90812	80.9
1951..	9	81	90.8	1.95809	17.62281	1.93012	85.1
1952..	10	100	96.4	1.98408	19.84080	1.95212	89.6
1953..	11	121	100.5	2.00217	22.02387	1.97412	94.2
1954..	12	144	103.2	2.01368	24.16416	1.99612	99.1
1955..	13	169	110.1	2.04179	26.54327	2.01812	104.3
1956..	14	196	114.8	2.05994	28.83916	2.04012	109.7
Total..	...	2,030	...	50.23148	44.65912	...	...

SOURCE: *Business Statistics, 1957.*

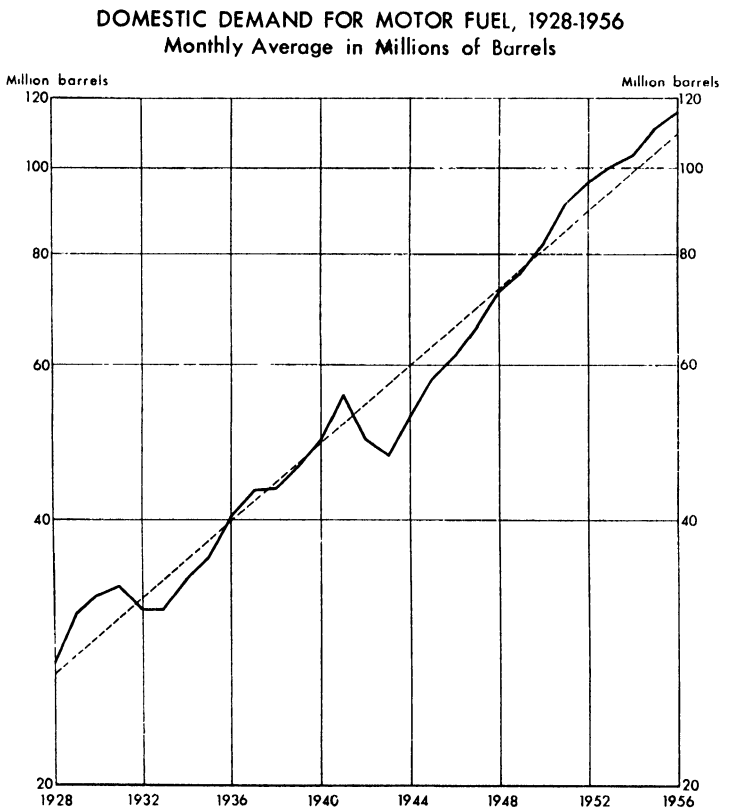
These two equations are merely equations (33) and (34) after substituting log Y for Y in the computation of the summations. When these coefficients are used to compute trend values, they will yield the logarithms of the trend values, since they were calculated from the logarithms of the original data.

The amounts in Column (7), headed "Log Y_c," are computed in the same manner as the amounts in the last column of Table 12-3 (page 288). The a value is log Y_c for 1942, the middle year of the series. Log Y_c for

each succeeding year is found by adding the value of *b*, and for each earlier year, by subtracting. The trend values in Column (8) are derived by finding the antilogs for the logarithms in the preceding column.

The trend values are plotted along with the original data in Chart 13-2. The fact that it is a straight line on semilogarithmic paper means that it increases at a constant rate. This characteristic of the trend is of considerable significance. That each of the trend values computed in Table 13-1 is 5 per cent more than the preceding year can be verified by making the necessary computations. These computations are unnecessary, however, since the *b* value is the logarithm of the ratio of each year to the preceding year. In Table 13-1 the value of *b* was found to be .02200, which is the logarithm of 1.05. Adding .02200 to the logarithm of any trend value multiplies that trend value by 1.05, which is increasing it by 5 per cent. From the *b* value of any logarithmic trend the rate of increase for the trend can be determined.

Chart 13-2



Source of data Table 13-1

SECOND-DEGREE PARABOLA

The examples of trend lines given previously illustrate the fact that the trend may not continue in a given direction at either a constant amount or a constant rate. A series may have an upward trend, but the amount of the increase may become smaller and smaller. Not uncommonly the trend of such a series finally turns down.

The method of least squares has been applied in the preceding pages to the fitting of a straight line to the data, and a straight line to the logarithms of the data. Other types of curves can be fitted by the method of least squares, thus making possible the use of this method when a more flexible trend than a straight line is required. The simplest of these nonlinear trend lines is the *second-degree parabola*, the equation of which may be written in the form

$$Y_c = a + bX + cX^2 \quad \text{(37)}$$

When numerical values for a, b, and c have been derived, the trend value for any year may be computed by substituting in the equation the value of X for that year. The following formulas are used for computing these constants, on the condition that the years are represented by an arithmetic progression with the origin at midrange, as already demonstrated in Table 12-3 (page 288).

$$c = \frac{N \cdot \Sigma X^2 Y - \Sigma X^2 \Sigma Y}{N \cdot \Sigma X^4 - (\Sigma X^2)^2} \quad \text{(38)}$$

$$c = \frac{(57)\,(849{,}240.2) - (15{,}428)\,(3{,}564.2)}{(57)\,(7{,}513{,}436) - (15{,}428)^2}$$

$$c = \frac{-6{,}581{,}786.2}{190{,}242{,}668} = -.034597$$

$$a = \frac{\Sigma Y - c\,\Sigma X^2}{N} \quad \text{(39)}$$

$$a = \frac{3{,}564.2 - (-.034597)\,(15{,}428)}{57}$$

$$a = \frac{3{,}564.2 + 533.8}{57} = \frac{4{,}098.0}{57}$$

$$a = 71.894737$$

Formula (34), which was presented on page 288, is used to calculate the value of b.

$$b = \frac{-13{,}205.8}{15{,}428} = -.855963$$

The values to be substituted in these formulas are derived in Table 13-2 in the same manner as for a straight line[1]. The computations in Table 13-2 are for an odd number of years, so the years are assigned values of a progression with a common difference of 1. No example is given for an even number of years since no new principle is involved.

Table 13-2

COMPUTATION OF A SECOND-DEGREE PARABOLA TREND

ANTHRACITE COAL CONSUMPTION: 1900 TO 1956

Year (1)	(2) X	(3) X^2	(4) X^4	Consumption (Million tons) (5) Y	(6) XY	(7) X^2Y	(8) Y_c
1900..	−28	784	614,656	55.5	−1554.0	43512.0	68.7
1901..	−27	729	531,441	65.2	−1760.4	47530.8	69.8
1902..	−26	676	456,976	40.5	−1053.0	27378.0	70.8
1903..	−25	625	390,625	72.6	−1815.0	45375.0	71.7
1904..	−24	576	331,776	70.7	−1696.8	40723.2	72.5
1905..	−23	529	279,841	75.2	−1729.6	39780.8	73.3
1906..	−22	484	234,256	68.8	−1513.6	33299.2	74.0
1907..	−21	441	194,481	82.6	−1734.6	36426.6	74.6
1908..	−20	400	160,000	80.2	−1604.0	32080.0	75.2
1909..	−19	361	130,321	77.9	−1480.1	28121.9	75.7
1910..	−18	324	104,976	81.1	−1459.8	26276.4	76.1
1911..	−17	289	83,521	86.5	−1470.5	24998.5	76.4
1912..	−16	256	65,536	80.2	−1283.2	20531.2	76.7
1913..	−15	225	50,625	85.5	−1282.5	19237.5	76.9
1914..	−14	196	38,416	84.0	−1176.0	16464.0	77.1
1915..	−13	169	28,561	88.1	−1145.3	14888.9	77.2
1916..	−12	144	20,736	87.1	−1045.2	12542.4	77.2
1917..	−11	121	14,641	94.1	−1035.1	11386.1	77.1
1918..	−10	100	10,000	92.8	−928.0	9280.0	77.0
1919..	−9	81	6,561	81.5	−733.5	6601.5	76.8
1920..	−8	64	4,096	85.8	−686.4	5491.2	76.5
1921..	−7	49	2,401	82.0	−574.0	4018.0	76.2
1922..	−6	36	1,296	56.8	−340.8	2044.8	75.8
1923..	−5	25	625	86.9	−434.5	2172.5	75.3
1924..	−4	16	256	80.7	−322.8	1291.2	74.8
1925..	−3	9	81	64.1	−192.3	576.9	74.2
1926..	−2	4	16	77.2	−154.4	308.8	73.5
1927..	−1	1	1	74.7	−74.7	74.7	72.7
1928..	0	0	0	73.7	0	0	71.9
1929..	1	1	1	71.5	71.5	71.5	71.0
1930..	2	4	16	67.6	135.2	270.4	70.1
1931..	3	9	81	58.4	175.2	525.6	69.0
1932..	4	16	256	50.5	202.0	808.0	67.9
1933..	5	25	625	49.6	248.0	1240.0	66.8
1934..	6	36	1,296	55.5	333.0	1998.0	65.5

Table 13-2 continued on next page.

[1]The value of ΣX^4 may be computed from the following formula:

$$\Sigma X^4 = \frac{3N^5 - 10N^3 + 7N}{240}$$

Table 13-2 (Continued)

Year (1)	(2) X	(3) X^2	(4) X^4	Consumption (Million tons) (5) Y	(6) XY	(7) X^2Y	(8) Y_c
1935..	7	49	2,401	51.1	357.7	2503.9	64.2
1936..	8	64	4,096	53.2	425.6	3404.8	62.9
1937..	9	81	6,561	50.4	453.6	4082.4	61.4
1938..	10	100	10,000	45.2	452.0	4520.0	59.9
1939..	11	121	14,641	49.7	546.7	6013.7	58.3
1940..	12	144	20,736	49.0	588.0	7056.0	56.7
1941..	13	169	28,561	52.7	685.1	8906.3	54.9
1942..	14	196	38,416	56.5	791.0	11074.0	53.1
1943..	15	225	50,625	57.1	856.5	12847.5	51.3
1944..	16	256	65,536	59.4	950.4	15206.4	49.4
1945..	17	289	83,521	51.6	877.2	14912.4	47.4
1946..	18	324	104,976	53.9	970.2	17463.6	45.3
1947..	19	361	130,321	48.2	915.8	17400.2	43.2
1948..	20	400	160,000	50.2	1004.0	20080.0	41.0
1949..	21	441	194,481	37.7	791.7	16625.7	38.7
1950..	22	484	234,256	39.9	877.8	19311.6	36.3
1951..	23	529	279,841	37.0	851.0	19573.0	33.9
1952..	24	576	331,776	35.3	847.2	20332.8	31.4
1953..	25	625	390,625	28.0	700.0	17500.0	28.9
1954..	26	676	456,976	26.9	699.4	18184.4	26.3
1955..	27	729	531,441	22.3	602.1	16256.7	23.6
1956..	28	784	614,656	23.8	666.4	18659.2	20.8
Total..	...	15,428	7,513,436	3,564.2	−13,205.8	849,240.2	...

SOURCE: Bureau of Mines.

GROWTH CURVES

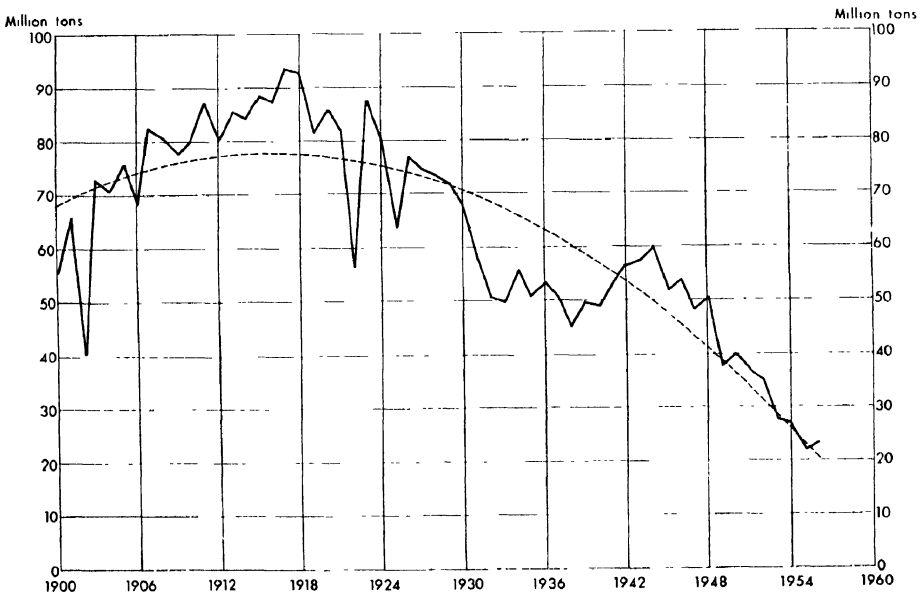
The use of a straight line fitted to the actual data or a straight line fitted to the logarithms of the data provides a satisfactory trend line for many time series, and when fitted to relatively short periods, it proves adequate for most situations. However, it does not seem logical for economic time series to continue to grow indefinitely either by a constant amount or a constant rate per year. It is not unusual for the rate of growth in the sale of a new product to slow down gradually as the product penetrates more and more nearly the complete market. This situation seems to be particularly appropriate for the per capita consumption of consumer goods. A new product is introduced and per capita consumption is at first very small; but as the product gains acceptance, the consumption per capita increases rapidly.

When this type of series is plotted on an arithmetic chart, it takes the shape of an elongated S, similar to the shape of the ogive or cumulative frequency distribution. At first the growth is very slow; but as the product is accepted, the demand increases by a greater amount each year. Finally, as the market becomes more and more fully developed, the amount of growth each year becomes less. The curve continues to grow more and more slowly, approaching an upper limit but not reaching it. When plotted on a semilogarithmic chart, the curve grows at a very rapid rate at first, with the rate of growth decreasing continuously. Even in the early stages the rate of growth declines from year to year since the greatest relative growth occurs at the beginning of the series when the amounts involved are small, and large percentages of increase are more easily achieved than when the amounts involved are larger.

The growth curves do not reach a maximum and turn down in the manner of the second-degree parabola in Chart 13-3. In some situations this failure to turn down is not an accurate description of events, as illustrated by the history of anthracite coal. Consumption increased at a steadily decreasing rate and finally the amount of consumption began to decrease. It would be possible to fit a curve that was the reverse of the growth curve, declining at a steadily decreasing rate; but this would necessitate splitting the series into two portions. When a product meets competition from

Chart 13-3

ANTHRACITE COAL CONSUMPTION, 1900-1956



Source of data Table 13-2

another product that gradually encroaches on its market, it can easily happen that the rate of growth continues to slow down until the rate becomes zero and then negative. The growth curves will not be applicable in such a situation; but as long as the series continues to grow at a decreasing rate, such curves are a logical description of this growth.

A number of different growth curves have been used to measure secular trend, but the best known are the Pearl-Reed and the Gompertz curves. The Pearl-Reed curve has been widely used by Raymond Pearl and L. J. Reed in the study and forecasting of population growth. The Gompertz curve has probably been more widely used as a measure of the trend of economic time series and will be described here as representative of this type of trend line.

A simple approximate method of fitting a Gompertz curve resembles the method of semiaverages described in Chapter 12 as a method of fitting a straight line. The series is divided into three equal subperiods, which requires that the length of the period be a number divisible by three, in the same manner as the period chosen for the method of semiaverages must be divisible by two. The average of the trend values for each of the subperiods equals the average of the actual values for that subperiod, in the same way that in the method of semiaverages the average of the trend values for each subperiod is equal to the average of the actual values for that subperiod. The method of fitting the Gompertz curve utilizes the logarithms of the data, and the equation derived is the equation for the values of $\log Y_c$:

$$\log Y_c = \log a + (\log b)c^x. \qquad (40)$$

The value of X represents the year, with the origin at the first of the period ($X = 0$).

The Gompertz curve is fitted to the data in Table 13-3, representing the per capita consumption of rayon and acetate fibers in the United States for the years 1921 to 1956. Per capita data are used to remove from the series the growth due to the increase in population. Since population growth can be described by the growth curves, it is not actually necessary to remove from the series the effect of population growth on consumption; but it does simplify the problem by removing one causal factor. Chart 13-4 shows that consumption increased at a gradually decreasing rate; this type of growth is best suited to the use of the growth curves.

The method of fitting the Gompertz curve described here computes the values of c, $\log b$, and $\log a$ in the order given. The values of S_1, S_2, and S_3 are taken from Table 13-3, where they were computed by finding the summations of $\log Y$ for each of the three subperiods. The number of years in the subperiods is designated by n, and care must be taken to distinguish between the values of n and N. The total number of years to which the

Table 13-3

FITTING A GOMPERTZ CURVE TO PER CAPITA CONSUMPTION OF RAYON AND ACETATE FIBERS IN THE UNITED STATES: 1921 TO 1956

Year (1)	(2) X	Per capita consumption (Pounds) (3) Y	(4) *Log Y*	Subperiod totals (5)	(6) c^x	(7) *(Log b) c^x* *Log b = −2.032647*	(8) *Log Y_c* *(7) + Log a* *Log a = 1.31601*	(9) *Y_c antilog of (8)*
1921..	0	.18	9.25527 −10		1.00000	−2.03265	9.28336 −10	.19
1922..	1	.22	9.34242 −10		.95015	−1.93132	9.38469 −10	.24
1923..	2	.29	9.46240 −10		.90278	−1.83503	9.48098 −10	.30
1924..	3	.37	9.56820 −10		.85778	−1.74356	9.57245 −10	.37
1925..	4	.50	9.69897 −10		.81500	−1.65661	9.65940 −10	.46
1926..	5	.51	9.70757 −10		.77437	−1.57402	9.74199−10	.55
1927..	6	.84	9.92428 −10		.73577	−1.49556	9.82045−10	.66
1928..	7	.83	9.91908 −10		.69908	−1.42100	9.89501−10	.79
1929..	8	1.09	.03743		.66424	−1.35017	9.96584−10	.92
1930..	9	.96	9.98227 −10		.63113	−1.28286	.03315	1.07
1931..	10	1.27	.10380		.59967	−1.21892	.09709	1.25
1932..	11	1.24	.09342	S_1 = 7.09511 − 10	.56978	−1.15816	.15785	1.44
1933..	12	1.72	.23553		.54137	−1.10041	.21560	1.64
1934..	13	1.55	.19033		.51438	−1.04555	.27046	1.86
1935..	14	2.02	.30535		.48874	−.99343	.32258	2.10
1936..	15	2.50	.39794		.46437	−.94390	.37211	2.35
1937..	16	2.35	.37107		.44122	−.89684	.41917	2.63
1938..	17	2.52	.40140		.41923	−.85215	.46386	2.91
1939..	18	3.46	.53908		.39833	−.80966	.50635	3.20
1940..	19	3.60	.55630		.37847	−.76930	.54671	3.52
1941..	20	4.37	.64048		.35960	−.73094	.58507	3.85
1942..	21	4.54	.65706		.34167	−.69449	.62152	4.18
1943..	22	4.73	.67486		.32464	−.65988	.65613	4.53
1944..	23	5.02	.70070	S_2 = 5.67010	.30846	−.62699	.68902	4.89
1945..	24	5.43	.73480		.29308	−.59573	.72028	5.25
1946..	25	6.11	.78604		.27847	−.56603	.74998	5.63
1947..	26	6.76	.82995		.26459	−.53782	.77819	6.00
1948..	27	7.73	.88818		.25140	−.51101	.80500	6.38
1949..	28	6.57	.81757		.23887	−.48554	.83047	6.77
1950..	29	8.79	.94399		.22696	−.46133	.85468	7.16
1951..	30	8.15	.91116		.21565	−.43834	.87767	7.54
1952..	31	7.63	.88252		.20490	−.41649	.89952	7.93
1953..	32	7.55	.87795		.19469	−.39574	.92027	8.32
1954..	33	7.11	.85187		.18498	−.37600	.94001	8.71
1955..	34	8.59	.93399		.17576	−.35726	.95875	9.09
1956..	35	7.15	.85431	S_3 = 10.31233	.16700	−.33945	.97656	9.47

SOURCE: *Textile Organon*, March, 1957.

curve is fitted is represented by N, so $n = \frac{N}{3}$. The values needed for fitting the curve are:

$$n = 12,\ S_1 = 7.09511 - 10,\ S_2 = 5.67010, \text{ and } S_3 = 10.31233.$$

The value of c is computed by the following equation:

$$c^n = \frac{S_3 - S_2}{S_2 - S_1} \quad \text{(40a)}$$

$$c^n = \frac{10.31233 - 5.67010}{5.67010 - (7.09511 - 10)} = \frac{4.64223}{8.57499} = .54137$$

$$\log c = \frac{\log .54137}{12} = \frac{9.73349 - 10}{12} = 9.97779 - 10$$

$$c = .95015$$

The value of $\log b$ is computed from the value of c already computed and the summations from Table 13-3.

$$\log b = \frac{(S_2 - S_1)(c - 1)}{(c^n - 1)^2} \quad \text{(40b)}$$

$$\log b = \frac{(8.57499)(-.04985)}{(-.45863)^2}$$

$$\log b = \frac{-.427549}{.210341} = -2.032647$$

The value of $\log a$ is computed as follows:

$$\log a = \frac{1}{n}\left(S_1 - \frac{S_2 - S_1}{c^n - 1}\right) \quad \text{(40c)}$$

$$\log a = \frac{1}{12}\left([7.09511 - 10] - \frac{8.57499}{-.45863}\right)$$

$$\log a = \frac{1}{12}(7.09511 - 10 + 18.69697) = 1.31601$$

The equation of the Gompertz curve fitted to the per capita consumption of rayon and acetate fibers in the United States is as follows:

$$\log Y_c = 1.31601 + (-2.032647)\ .95015^x$$

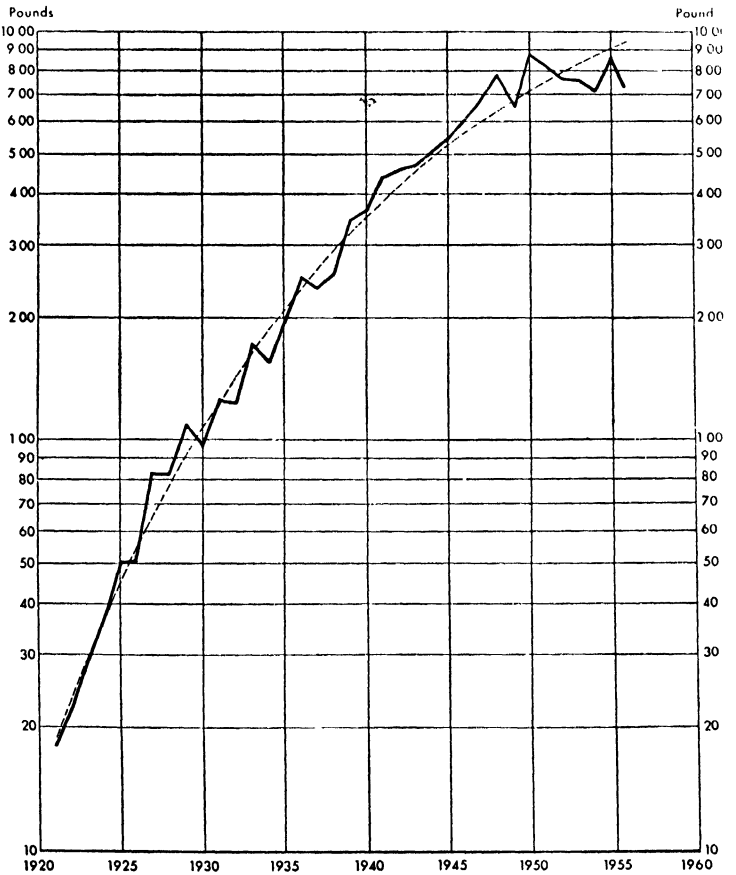
Origin: July 1, 1921.
X unit: 1 year.
Y unit: Per capita consumption of rayon and acetate fibers (in pounds).

The trend value for each year may be computed by substituting successive values of X in the above equation. It is important that the computations be performed systematically, and Columns (6) through (9) in Table 13-3 are

given as an example of an efficient method of recording the results. The first step is to compute c^x for all the values of X from 0 to 35 and enter the values in Column (6). Any number to the zero power is 1, so the value of c for 1921 is entered as 1.00000. The value of c for 1922 is $c^1 = .95015^1 = .95015$. The value of c for 1923 is $c^2 = .95015^2 = .90278$. Successive multiplication by c (.95015) gives the values in Column (6) for c^x. The value of log Y_c can be computed in two operations by multiplying the value of c^x in Column (6) by log b (-2.032647), and then adding the value of log a (1.31601). The values of (log b) c^x are given in Column (7). Column (8) was computed by adding 1.31601 to Column (7). The values of the trend, Y_c, are given in Column (9) and are plotted as the trend line on Chart 13-4. The trend line rises at a gradually decreasing rate approaching the value of a. As the value of X increases, c^x becomes smaller and smaller

Chart 13-4

PER CAPITA CONSUMPTION OF RAYON AND ACETATE FIBERS IN THE UNITED STATES, 1921-1956



Source of data Table 13-3

but does not reach zero. If c^x could become zero, the trend value would become the value of a; however, the trend can only approach the value of a as a limit. Since $\log a = 1.31601$, $a = 20.7$ pounds of rayon and acetate fiber consumption per capita. This is the upper limit of per capita rayon and acetate fiber consumption in the United States, as projected by this equation. The use of such an equation to forecast secular trend will be discussed in Chapter 21.

For the Gompertz curve to be used, it is necessary that the rate of growth be decreasing, or that the value of $(S_3 - S_2)$ be less than the value of $(S_2 - S_1)$. If $(S_3 - S_1) = (S_2 - S_1)$, it would indicate that the series was increasing at a constant rate; and a straight line could appropriately be fitted to the logarithms of the data. If $(S_3 - S_2) > (S_2 - S_1)$, it means that the series has been growing at an increasing rate. However, growth at an increasing rate is most likely to occur in the early stages of development. The fact that an increasing rate of growth for any extended length of time is an unusual situation in business means that a trend with a decreasing rate of growth is generally fairly logical. In general, the period covered should be long enough for the industry or business to have reached a considerable degree of maturity.

CHOICE OF TREND LINE

There is no positive objective test that can be applied to a line to determine whether or not it is a satisfactory measure of the secular trend of a series. Within limits it is possible to decide by inspection of a chart whether a given line is a reasonable measure of the trend. No matter what other method is used to decide if a trend is satisfactory, the plotted line must give a satisfactory fit to a graph of the series to which the trend is fitted.

It is usually true that a number of lines will fit the data and describe the secular trend reasonably well. In such a case it is necessary to make a choice among them, and the decision may well be made on the basis of the method of computation used. The method of least squares is probably the most favored of all those described in the preceding pages. The line from which the square of the deviations is a minimum usually fits the data in a satisfactory manner, though not always. It is essential that the type of line be one that will fit the data, but often none of those that have been described will do so. If none of them will fit satisfactorily, it is possible to use a different period; the choice of the period is somewhat arbitrary. However, the movements, other than trend, most likely to affect the data are cyclical fluctuations; and so the data should cover one or more complete cycles. It is desirable to begin and end the period with years that are neither extremely high nor extremely low with respect to the cycle. But

even with these specifications, there is a certain amount of choice possible in determining the period to be used, and this choice is a factor in determining the goodness of fit.

If no line that fits reasonably well can be determined by the method of least squares, the other methods may be tried. The graphic method permits the statistician to place the trend line where he thinks it should go; and for this reason, the graphic method is criticized by many. But since a trend line that does not fit in a manner satisfactory to the statistician is not acceptable and will not be used, there is a possibility that even the computed methods are not completely free from a certain subjective element.

If a least squares line fitted to a logical period gives a satisfactory fit, it probably should be used in preference to the other methods. But, in reality, any line that is a good fit of the data and appears to describe the way they have grown or declined, is probably a satisfactory measure of secular trend. It is difficult to get away from the goodness of fit as observed on a graph as the basis for deciding on a trend line.

The trend lines fitted by the method of least squares or the method of semiaverages have an inflexibility that often makes them a poor fit of the series. For example, an unusually high or low year at the end of the series has a pronounced effect on the least squares trend line. This is inevitable, since the equations used minimize the squared deviations from the line. A large increase or decrease in any year, and particularly in several years, will pull the line in the direction of the extreme amounts. In doing so, it will tilt the straight line along its whole length.

MAKING USE OF A MEASURE OF SECULAR TREND

The executives of a growing business should have a clear picture of that growth as part of their basic information. The secular trend of a series gives a better picture of the growth than the actual data, since the actual data contain the effect of the business cycle on the series, as well as the effect of the secular trend.

If it is desired to compare the growth of two series, the best comparison can be made between the trends rather than between the two series themselves. This is particularly true if any pronounced cyclical fluctuations are present in the data.

Although it is not commonly done, it would emphasize the changes in the trends of two series if the actual data were not plotted on the chart to make easier the comparison of the two trend lines. For example, the cyclical fluctuations in anthracite coal consumption could be ignored if only the trend line were plotted. If a comparison of the trends of anthracite and bituminous coal were made, the trend values might be more significant than

the actual consumption of the two kinds of coal. In tracing the changes in the trend of consumption of different kinds of food, it might be easier to plot the trend lines instead of the actual consumption figures for each year. In this way the attention could be focused on the underlying shifts in demand and not obscured by the short-term variations in demand.

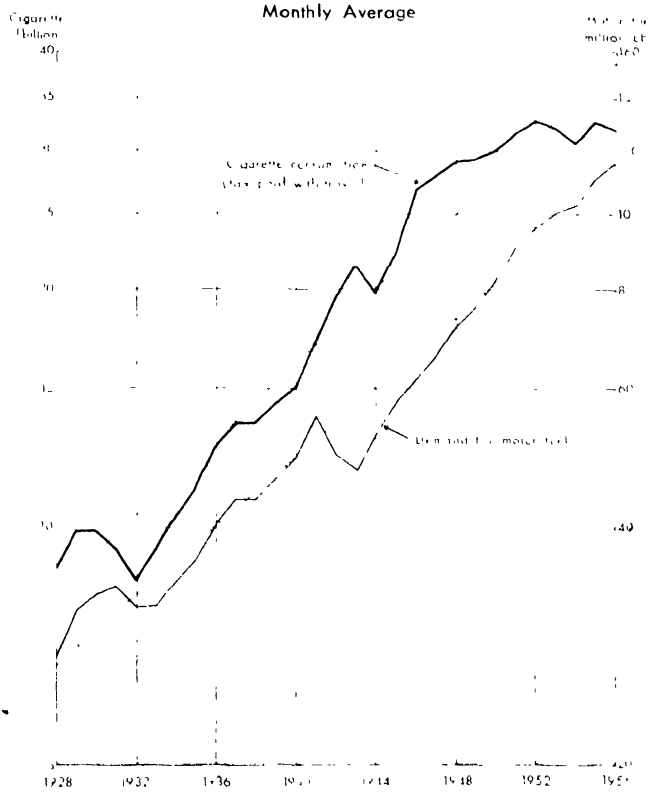
The following pages will discuss some of the ways in which the measures of secular trend can be used in making business decisions.

Comparison of Two Logarithmic Trend Lines

Since the rates of increase of two logarithmic straight lines are percentages, they can be compared even though the actual values of the trends are entirely different and not comparable. For example, if we want to compare the rate of growth of cigarette consumption with the rate of growth of the demand for motor fuel, the rates of increase of the trend lines can be used. Chart 13-5 shows cigarette consumption and the demand for motor

Chart 13-5

CIGARETTE CONSUMPTION AND DEMAND FOR MOTOR FUEL, 1928-1956
Monthly Average



fuel for 1928 to 1956. The equations of the two straight lines fitted to the logarithms of the two series are as follows:

Cigarette consumption

$$\log Y_c = 1.25042 + .02360X$$

Origin: July 1, 1942.
X unit: 1 year.
Y unit: Average monthly consumption (billions of cigarettes).

Demand for motor fuel

$$\log Y_c = 1.73212 + .02200X$$

Origin: July 1, 1942.
X unit: 1 year.
Y unit: Average monthly demand (million of barrels).

It was shown on page 298 that the rate of increase in the logarithmic straight line is represented by the b value of the trend equation. From the above equations the following rates of increase are found.

Cigarette consumption: b equals .02360, which is the logarithm of 1.056. This represents a 5.6 per cent increase.
Demand for motor fuel: b equals .02200, which is the logarithm of 1.052. This represents a 5.2 per cent increase.

These two trend lines represent the growth in the two industries during the decades of the late twenties through the mid-fifties. The two percentages compare the rates of growth of the two series, and show that these industries grew at nearly the same rate. The slopes of straight lines fitted to the logarithms of the data rather than to the original data can always be compared in this manner. It will be observed that the two trend lines on Chart 13-5 are very nearly parallel, which means that the two lines are rising at practically the same rate.

Instead of comparing the trends of two different series, the trend lines may be used to compare the rate of growth of the same series in two different periods. Chart 12-4 (page 279) shows the demand for motor fuel from 1917 to 1956. It is apparent that the rate of growth was much greater in the first part of the period from 1917 to 1927, than in the later part, from 1928 to 1956. The trend lines plotted on the chart are straight lines fitted to the logarithms of the two series, demand for motor fuel, 1917 to 1927, and demand for 1928 to 1956. The equations of the two trend lines are:

Trend equation, 1917 to 1927

$$\log Y_c = 1.04645 + .07084X$$

Origin: July 1, 1922.
X unit: 1 year.
Y unit: Average monthly demand for motor fuel, 1917 to 1927 (millions of barrels).

Trend equation, 1928 to 1956

$$\log Y_c = 1.73212 + .02200X$$

Origin: July 1, 1942.
X unit: 1 year.
Y unit: Average monthly demand for motor fuel, 1928 to 1956 (millions of barrels).

The b values in the equation represent the rates of increase of the trends, which are as follows:

1917 to 1927: b equals .07084, which is the logarithm of 1.177. This represents a 17.7 per cent increase.
1927 to 1941: b equals .02200, which is the logarithm of 1.050. This represents a 5.2 per cent increase.

Division of Trend into Its Components

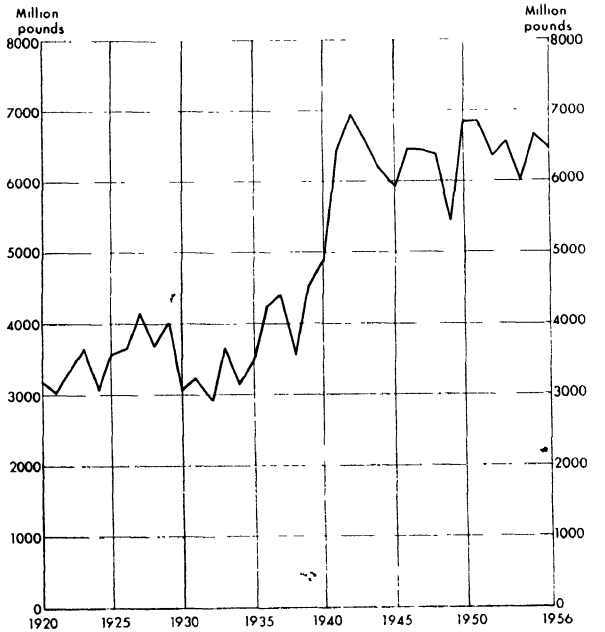
It is frequently possible to separate the secular trend into component elements, each of which may explain part of the change due to trend. Many business series in the United States show an upward trend for the obvious reason that the number of consumers has increased. Some of these can be related to the size of the population, while others should be compared with a more specific measure of the size of the consuming market. For example, the upward trend in the consumption of textile fibers is at least partly a result of the increase in population. The trend in the consumption of motor fuel may be related to the number of motor vehicles in use, since the motor vehicle is the consuming unit.

When data on consumption of textile fibers are reduced to a per capita basis, the amount of trend due to the increase in population is removed. Charts 13-6 and 13-7 on the following page show respectively the total consumption of textile fibers in pounds and the per capita consumption of textile fibers.

Chart 13-8 shows the trend in motor fuel consumption by motor vehicles from 1929 to 1955 and breaks this trend into the two components: number of motor vehicles and consumption of fuel per vehicle. This analysis makes clear that the rapid rise in fuel consumption has been caused primarily by the increase in the number of vehicles registered. In fact, since 1946 the consumption per vehicle has remained practically unchanged and nearly all the increase in total consumption has resulted from more cars in use. Using the trend instead of the actual data makes it possible to ignore the decline during the depression years of the thirties and the curtailment caused by the war. Both of these declines were temporary, due to specific causes other than trend.

Chart 13-6

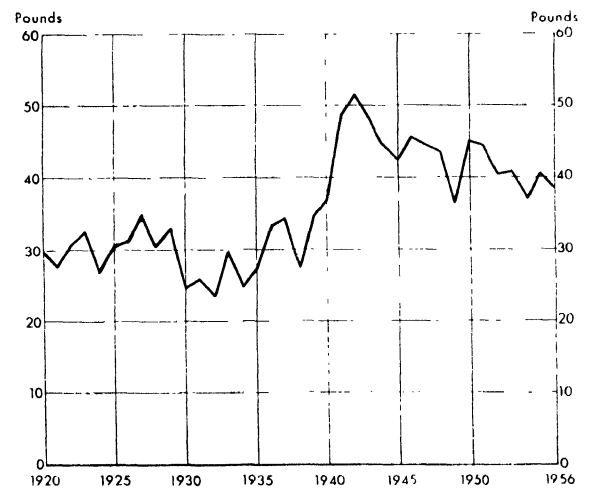
TEXTILE FIBER CONSUMPTION BY MILLS IN THE UNITED STATES, 1920-1956



Source of data: Textile Organon, March, 1957

Chart 13-7

PER CAPITA TEXTILE FIBER CONSUMPTION IN THE UNITED STATES, 1920-1956

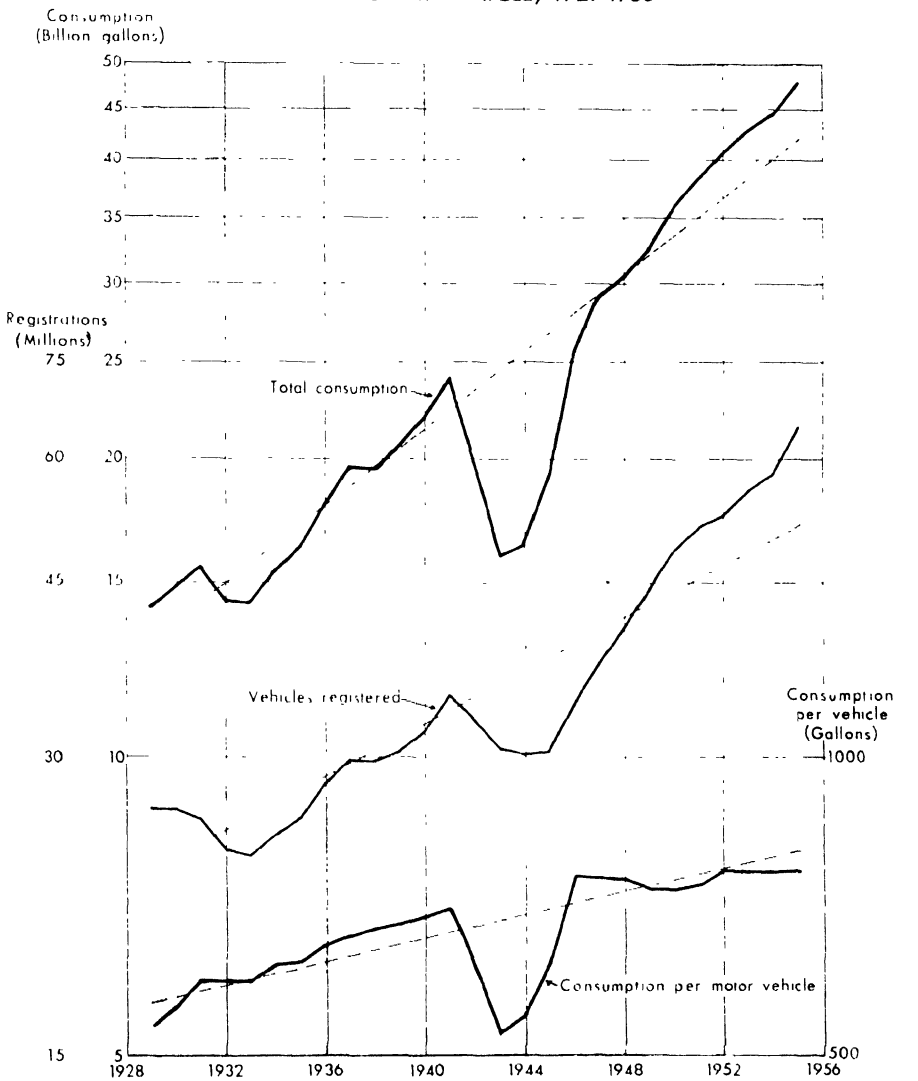


Source of data: Textile Organon, March, 1957
Statistical Abstract, 1957

The trend in consumption of motor fuel by motor vehicles increased at an average rate of 4.2 per cent per year. Motor vehicle registrations increased an average of 2.9 per cent per year, while consumption per vehicle increased 1.3 per cent, less than half the rate of increase in motor vehicle registrations.

Chart 13-8

CONSUMPTION OF MOTOR FUEL BY MOTOR VEHICLES, NUMBER OF MOTOR VEHICLES REGISTERED, AND CONSUMPTION OF FUEL PER MOTOR VEHICLE, 1929-1955



Source of data Bureau of Public Roads annual report, *Highway Statistics*, 1956

FORECASTING SECULAR TREND

When the long-term movements in a series are important to the businessman, it usually follows that a forecast of these movements is needed. Any consideration of the investment possibilities of an industry, for example, should take into account its past secular trend and its probable growth or decline.

In an industry requiring a large fixed investment of capital, it is necessary to make a forecast of the future trend of the business when plans are under consideration for replacing old equipment or buying new. If an electric power company is planning the installation of new generators, it is advisable to forecast the course of the consumption of current during the life of the new generators. If a pronounced growth is anticipated, it would probably be good business to install equipment with a larger capacity than needed at present.

It is too common practice for businessmen to base a decision as to expansion of manufacturing capacity on the current demand for their product; but if the increase in demand has been due primarily to a cyclical expansion, they should know that from the fundamental nature of the business cycle it can be expected that the demand will not continue at the same high level for very long. An understanding of the nature of secular trend and a conscious forecast of it are the most logical bases for the expansion of the fixed capital of a business. With this forecast as a guide, additions to plant and equipment can be made at the most advantageous times, which will usually be when demand for the product of the concern is below the peak. At such times prices are lowest and the expansion can be carried out under the most favorable conditions. The fact that demand at this time is not sufficient to utilize the full capacity of the plant is not the most important fact to consider; the forecast of the long-term trend of the business should be the guiding factor. As long as cyclical fluctuations are present, it is inevitable that the concern will operate well below peak capacity at certain times. These cyclical fluctuations represent a separate problem which the businessman must take into account; but when the problem is a forecast of the long-term trend, it is highly important not to permit the results to be influenced by the cyclical movements.

The contribution of statistical methods to the forecasting of secular trend consists primarily in the measurement of events in the past. Since the only basis for the forecast of the future is a knowledge of past happenings, the first step in the forecast of secular trend is normally the measurement of the trend in the immediate past.

The use of an equation to describe the trend of a series has suggested that a simple method of forecasting trend would be by *extrapolation,* that

is, projection of this fitted trend into the future. This process should be used only when it has been decided that the trend in the future will follow the same pattern as it did in the period to which the trend line was fitted. If the trend grew the same amount each year, a straight-line trend would be projected with the same increase each year in the future that it had shown in the past. If a straight line were fitted to the logarithms of the data, it would mean that the *rate* of increase was the same each year; and if this line were extrapolated, it would mean that the trend would continue to increase at the same *rate*.

It is obvious that there is no logical reason for assuming that the increase or decrease in the trend for a relatively short period (or even for a long period) in the past will continue for any fixed time in the future. Many different factors would normally have an effect on the growth or decline of a series, and it is reasonable to expect that these factors will not stay the same and that their importance with respect to each other will not remain fixed.

Projecting a past trend assumes that the factors causing the change in the secular trend *will continue to have the same influence in the future as in the past*. A more logical approach to the problem of forecasting trend is to determine as accurately as possible what factors caused the growth or decline, to consider the development of any new factors, and then to evaluate the future effect of each of these. Often the separate components may be forecast with more accuracy than the composite trend, since fewer causal factors would be affecting each component.

The first step in applying this method obviously would be the measurement of the secular trend, using one or more of the methods described earlier. It would be important, if possible, to measure the trend in the several elements as was done on pages 311 to 313 for the trend of motor fuel demand. It might even be possible to separate the trend of one or more of the components into subcomponents, which might be forecast individually. Each situation would be a separate problem, and the use of statistical methods would be chiefly to measure certain aspects of the problem and to break it into as many and as simple factors as possible. From this point the forecast would not use statistical methods so much as economic analysis of the factors in the specific problem.

In addition to the uses described on the preceding pages, a measure of the secular trend is used in determining the cyclical fluctuations in a time series. This is an important use of a measure of secular trend, but it will not be discussed at this point because it properly belongs with the description of methods of isolating cyclical fluctuations. The measurement of cyclical fluctuations is discussed in Chapter 15.

STUDY QUESTIONS

1. There is a distinction made by some statisticians between what they call "rational" trends and "empirical" trends. From the definitions of these two terms, what would you conclude this distinction to mean?

2. It has been said that the growth curves can be forecast more logically than either the straight line or the parabola. What reason can you give in support of such a statement?

3. Why does the straight line fitted to the logarithms of the data permit the comparison of the b values of two equations fitted to series in entirely different units? (For example, in the text the comparison was made between the b values of the equations representing cigarette consumption and demand for motor fuel.)

4. If you used a computed trend to forecast the trend of a series, you would be making some assumptions. What are some of these basic assumptions?

5. How do you explain the fact that cigarette consumption and motor fuel demand increased at so nearly the same rate, as shown in Chart 13-5 (page 309)?

6. What criterion should be used to decide whether or not to fit a nonlinear trend line to a series? (Remember that it is always possible to break the series into segments small enough so that a straight line could be made to fit.)

7. What characteristics must a time series possess for the Gompertz curve to be fitted by the method described in the text?

PROBLEMS

1. Sketch the type of curve represented by each of the following equations:
   (a) Log $Y_c = 2.10958 + .00484X$
   (b) Log $Y_c = 1.90472 - .01382X$
   (c) $Y_c = ab^x$ or $Y_c = (127.3)\,(1.052)^x$
   (d) $Y_c = 157.3 + 13.25X + .198X^2$
   (e) $Y_c = 98.4 - 2.19X + .013X^2$

2. The time series given in the problem at the end of Chapter 12, page 294, were plotted on arithmetic charts; and straight lines were used to describe the secular trend. This problem considers the use of nonlinear trends to improve the description of the long-term trend.
   (a) The first step in the analysis should be to plot the series on semilogarithmic paper to determine whether a straight line to the logarithms would give a better fit than the straight line fitted in the previous chapter.
   (b) Fit a straight line to the logarithms for the series that seem to justify using this type of trend.
   (c) Decide which type of line serves as the best description of the trend.
   (d) State the criterion you use to determine which line serves as the best measure of secular trend.

3. The following table gives the number of motor vehicles registered in the United States and the population of the United States since 1901. These series may be used to fit three Gompertz curves for:

(a) Motor vehicle registrations.
(b) Population.
(c) Motor vehicle registrations per thousand individuals.

MOTOR VEHICLE REGISTRATIONS AND ESTIMATED POPULATION: 1901 TO 1956

Year	Motor vehicle registrations (Thousands)	Population (Thousands)	Year	Motor vehicle registrations (Thousands)	Population (Thousands)
1901.....	14	77,585	1931.....	25,862	124,149
1902.....	23	79,160	1932.....	24,133	124,949
1903.....	33	80,632	1933.....	23,877	125,690
1904.....	55	82,165	1934.....	24,954	126,485
1905.....	79	83,820	1935.....	26,230	127,362
1906.....	108	85,437	1936.....	28,172	128,181
1907.....	143	87,000	1937.....	29,706	128,961
1908.....	198	88,709	1938.....	29,443	129,969
1909.....	312	90,492	1939.....	30,615	131,028
1910.....	469	92,407	1940.....	32,035	132,122
1911.....	640	93,868	1941.....	34,472	133,402
1912.....	944	95,331	1942.....	32,579	134,860
1913.....	1,258	97,227	1943.....	30,500	136,739
1914.....	1,763	99,118	1944.....	30,086	139,397
1915.....	2,491	100,549	1945.....	30,638	139,928
1916.....	3,618	101,966	1946.....	33,946	141,389
1917.....	5,119	103,266	1947.....	37,360	144,126
1918.....	6,160	103,203	1948.....	40,556	146,631
1919.....	7,577	104,512	1949.....	44,140	149,188
1920.....	9,239	106,466	1950.....	48,567	151,683
1921.....	10,494	108,541	1951.....	51,292	154,360
1922.....	12,274	110,055	1952.....	52,616	157,028
1923.....	15,102	111,950	1953.....	55,593	159,643
1924.....	17,613	114,113	1954.....	57,876	162,409
1925.....	19,941	115,832	1955.....	62,020	165,271
1926.....	22,053	117,399	1956.....	64,437	168,091
1927.....	23,140	119,038			
1928.....	24,512	120,501			
1929.....	26,503	121,770			
1930.....	26,532	123,188			

SOURCE: *Statistical Abstract of the United States, 1957.*

14 SEASONAL VARIATION

Any time series that shows a persistent tendency for certain months always to be particularly high or low probably contains a definite seasonal movement. This seasonal variation can be measured by statistical methods developed for this specific purpose. The measures of seasonal variation are usually referred to as *indexes of seasonal variation.* Because seasonal fluctuations in modern business are so pronounced, it is important that the business executive understand the significance and use of these indexes and the methods by which they are computed.

INDEXES OF SEASONAL VARIATION

Probably the simplest method of showing seasonal variation is to express the figure for each month as a percentage of the total for the year. For example, the number of freight cars loaded in the Southwestern District shows a pronounced seasonal variation, resulting primarily from the movement of crops. August, September, October, and November are decidedly above the remaining months of the year. Or, to state it in another way, more than one twelfth of the total freight cars are loaded in each of these months, and less than one twelfth of the total in each of the remaining eight months.

Computing these percentages rather precisely results in a table of percentages such as Table 14-1 on the opposite page. If there were no seasonal variation in the data, each of these monthly percentages would be $8\frac{1}{3}$, which is just one twelfth of the total for the year. The amount that each of them differs from $8\frac{1}{3}$ per cent expresses the effect of the seasonal factor on the data.

Another method which perhaps makes it easier to see the effect of seasonal variation on the different months is to multiply each of the percentages by 12, which makes their total 1,200. If there were no seasonal variation, each of the monthly percentages would be 100, with a total of 1,200 for the year. The amount by which any one month differs from 100 is the effect of seasonal variation on that month. Thus the data of Table 14-1 would be written as in Table 14-2.

Table 14-1

PERCENTAGE OF YEARLY TOTAL FREIGHT CAR LOADINGS IN EACH MONTH: 1947 TO 1956

SOUTHWESTERN DISTRICT

Month	Percentage of yearly total	Month	Percentage of yearly total
January	7.83	July	8.17
February	7.75	August	8.58
March	7.92	September	9.33
April	8.00	October	9.84
May	7.83	November	9.00
June	7.92	December	7.83
		Total	100.00

SOURCE: Unpublished study in the Bureau of Business Research, The University of Texas, April, 1957.

Table 14-2

INDEX OF SEASONAL VARIATION FOR FREIGHT CAR LOADINGS 1947 TO 1956

SOUTHWESTERN DISTRICT

Month	Index no.	Month	Index no.
January	94	July	98
February	93	August	103
March	95	September	112
April	96	October	118
May	94	November	108
June	95	December	94
		Total	1,200

SOURCE: Table 14-1.

This means that December being 94 per cent, the effect of the seasonal variation on December is – 6 per cent. The difference between each percentage and 100 is the effect that seasonal variation has on each month. A method of showing this even more clearly is to express each month, not as a percentage but by giving the amount, plus or minus, by which a particular month deviates from 100. The above index would then be

expressed as in Table 14-3. These percentages show the amount each month is above or below a "normal" month, which is another way of saying a month in which no seasonal variation is present.

Table 14-3

INDEX OF SEASONAL VARIATION FOR FREIGHT CAR LOADINGS 1947 TO 1956

SOUTHWESTERN DISTRICT

Month	Index no.	Month	Index no.
January	− 6	July	− 2
February	− 7	August	3
March	− 5	September	12
April	− 4	October	18
May	− 6	November	8
June	− 5	December	− 6
		Total	0

SOURCE: Table 14-1.

These three methods of expressing the index give exactly the same information, expressing it in each case in a slightly different manner. The second method, as illustrated in Table 14-2, is the one used most extensively by statisticians.

The three indexes were computed from monthly data ranging over a period of years, and represent the average fluctuation in the monthly data for that period. This average variation is taken as typical of the effect that seasonal variation has on the series. Specific instructions on how to compute an index of seasonal variation are given later in this chapter.

Probably the simplest method of arriving at an index of seasonal variation would be for an individual who is familiar with a particular industry or business to estimate the percentage of the total business of the year normally transacted in each month. These percentages would be an index of seasonal variation, if made to total 100 as the illustration given in Table 14-1. The accuracy of the index would depend entirely upon the knowledge of the individual making the estimate, and could not be expected to approach any high degree of precision. The owner of a retail store would know that December is the month in which his store sells the largest amounts of goods, but he probably would not know whether it is typically 40, 50, or 60 per cent above the average month's sales.

A method almost as simple would be to choose a year in which the seasonal variation was thought to be typical of all years, and express each

month of that year as a percentage of the total for that year. This would constitute an index expressed in the first form described above. Its accuracy would depend upon how representative the particular year actually was. There would be considerable danger that no one year was highly representative, so the method cannot be recommended. The basic reason for substituting statistical measures for opinions and general impressions is the fact that properly computed measures have a better chance of being accurate.

MEASURING SEASONAL VARIATION — A PROBLEM OF AVERAGES

The basic approach to the measurement of the effect of seasonal forces on a time series has been the use of averages. The idea of a seasonal pattern is in reality based on the concept of an average. The seasonal forces do not have exactly the same effect in different years. For example, a late spring will cause the seasonal rise in retail trade to come later than in other years. It is not to be expected that an absolutely rigid seasonal pattern will hold in any business; but if the effects of the seasonal factors are reasonably similar from year to year, it is valid to compute an average seasonal pattern. The problem of dispersion in data and the effect of dispersion on the use of an average are also involved, which means that some method of measuring the dispersion must be used to evaluate any measure of seasonal variation. The more the variations in the seasonal patterns from year to year, the less typical, and therefore the less reliable, the average.

In Chart 14-1 the domestic demand for motor fuel is plotted by months for the years 1946 to 1956. The pronounced seasonal fluctuation is easily seen, and it appears that the pattern is regular enough to justify computing an index of seasonal variation. The simplest method of measuring the seasonal variation is to find the average January value over a period of years, the average February value, and so on for the twelve months. The average demand for motor fuel is computed for each month, using the most recent eleven values for each month. All of these values are shown in Table 14-4. The arithmetic mean of the February demand figures is 71.7 million barrels, while the August demand averages 97.0 million barrels. These two months represent the months of lowest and highest demand for motor fuel, and we may assume that the two averages measure the relative strength of the demand in the two months. Average demand in the other ten months may in the same manner be taken as measures of the seasonal level of those months. The twelve means may together be used as an index of the seasonal fluctuations. To be of maximum usefulness these twelve

Table 14-4

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

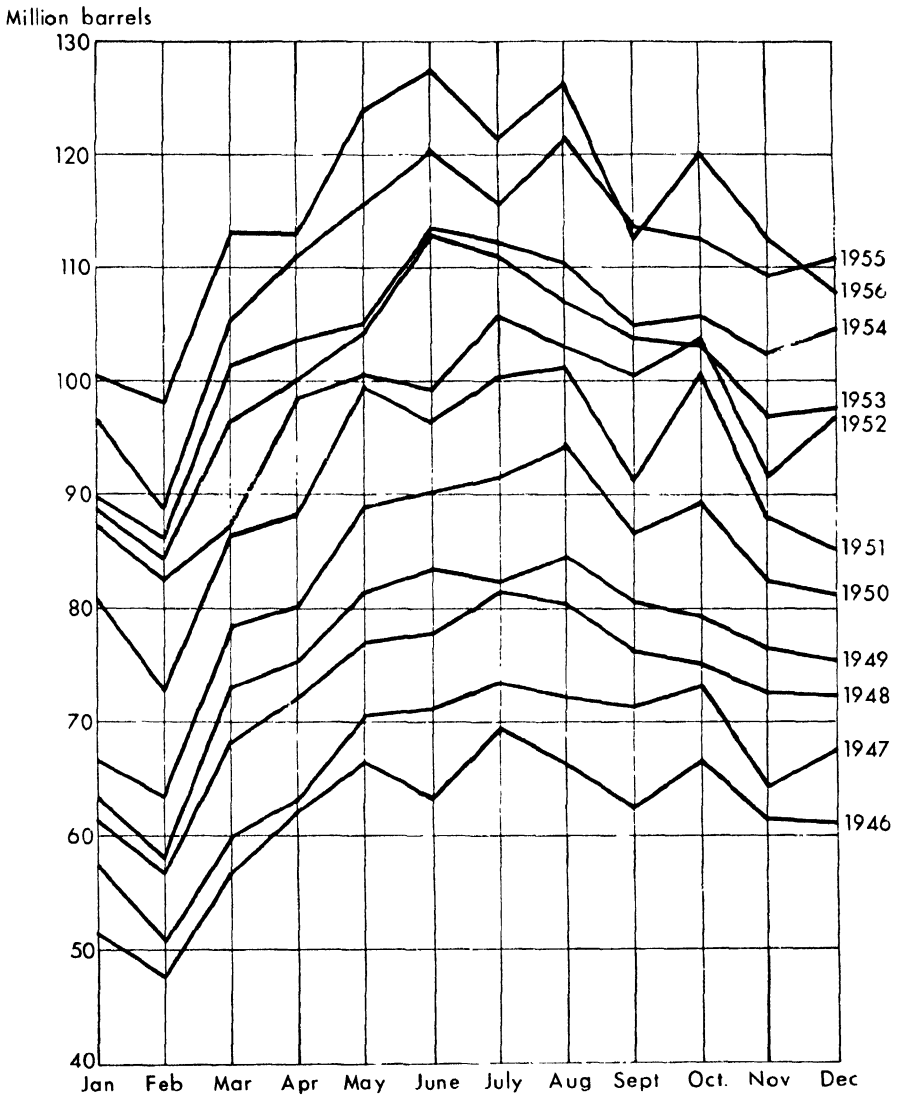
TABLE OF MONTHLY FIGURES IN MILLIONS OF BARRELS

Year and month (1)	Demand (Million barrels) (2)	Year and month (1)	Demand (Million barrels) (2)	Year and month (1)	Demand (Million barrels) (2)
1946 Jan.	51.7	1950 Jan.	66.9	1954 Jan.	89.9
Feb.	47.7	Feb.	63.4	Feb.	86.2
March	56.7	March	78.7	March	101.5
April	62.1	April	80.3	April	103.9
May	66.8	May	89.0	May	104.4
June	63.3	June	90.2	June	113.0
July	69.1	July	91.7	July	112.2
Aug.	66.7	Aug.	94.5	Aug.	110.2
Sept.	62.3	Sept.	86.8	Sept.	104.7
Oct.	66.6	Oct.	89.1	Oct.	105.6
Nov.	61.3	Nov.	82.7	Nov.	102.4
Dec.	61.1	Dec.	81.1	Dec.	104.3
1947 Jan.	57.1	1951 Jan.	80.7	1955 Jan.	96.4
Feb.	50.6	Feb.	72.6	Feb.	88.5
March	60.0	March	86.7	March	105.7
April	63.3	April	87.3	April	111.1
May	70.9	May	99.4	May	115.7
June	71.2	June	96.3	June	120.7
July	73.4	July	100.5	July	115.7
Aug.	72.1	Aug.	101.1	Aug.	121.8
Sept.	71.4	Sept.	91.3	Sept.	113.4
Oct.	73.3	Oct.	100.5	Oct.	112.6
Nov.	64.1	Nov.	88.0	Nov.	109.2
Dec.	67.5	Dec.	85.2	Dec.	111.0
1948 Jan.	61.3	1952 Jan.	87.1	1956 Jan.	100.6
Feb.	56.5	Feb.	82.3	Feb.	98.1
March	68.2	March	87.2	March	113.1
April	72.2	April	98.8	April	113.0
May	77.2	May	100.5	May	124.1
June	78.0	June	99.3	June	127.4
July	81.5	July	105.5	July	121.2
Aug.	80.3	Aug.	103.0	Aug.	126.2
Sept.	76.1	Sept.	100.2	Sept.	112.7
Oct.	75.2	Oct.	103.9	Oct.	120.1
Nov.	72.5	Nov.	91.6	Nov.	112.6
Dec.	72.1	Dec.	96.8	Dec.	108.2
1949 Jan.	63.1	1953 Jan.	88.1		
Feb.	58.0	Feb.	84.6		
March	73.3	March	96.7		
April	75.3	April	100.1		
May	81.7	May	104.2		
June	83.4	June	112.6		
July	82.1	July	111.0		
Aug.	84.7	Aug.	106.8		
Sept.	80.8	Sept.	103.6		
Oct.	79.3	Oct.	103.4		
Nov.	76.3	Nov.	96.9		
Dec.	75.6	Dec.	97.7		

SOURCE: *Survey of Current Business.*

Chart 14-1

DOMESTIC DEMAND FOR MOTOR FUEL
IN THE UNITED STATES, 1946-1956



Source of data. Table 14-4

means should be reduced to one of the forms given in Tables 14-1 to 14-3. This has been done in Column (3) of Table 14-5, where the mean for each month is expressed as a percentage of the mean of the twelve monthly means.

Table 14-5

INDEX OF SEASONAL VARIATION

BASED ON THE ARITHMETIC MEAN OF ACTUAL MONTHLY DATA

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Month (1)	Mean monthly demand (Millions of barrels) (2)	Index of seasonal variation (Average month = 100) (3)
January	76.6	86.5
February	71.7	80.9
March	84.4	95.2
April	88.0	99.2
May	94.0	106.1
June	96.0	108.3
July	96.7	109.1
August	97.0	109.5
September	91.2	102.9
October	93 6	105.6
November	87.1	98.2
December	87.3	98.5
Average	88.6	100.0

SOURCE: Table 14-4.

SPECIFIC SEASONALS

In the preceding section an index of seasonal variation was computed by simply averaging the domestic demand for motor fuel for each month over a period of eleven years. It was assumed that the differences between the averages for the various months were due entirely to seasonal forces. However, since trend and cyclical fluctuations are also present in this series, the assumption was probably not correct. It is usually more accurate to isolate the effect of the seasonal forces on each month before averaging a number of years to secure a measure of the typical seasonal variation.

The measurement of the effect of the seasonal forces on a given month will result in a figure called a *specific seasonal.* The different methods of computing these specific seasonals give a number of different methods of computing an index of seasonal variation. The methods of averaging the specific seasonals are the same no matter how the specific seasonals were derived; so the difference between methods is chiefly a difference in the ways of securing the values to be averaged.

Ratio to Twelve-Month Moving Average

The most widely used method of measuring the seasonals is some form of the ratio to a twelve-month moving average or twelve-month moving total. The basic principle of this method is the fact that an average of the twelve months of a year cannot be affected by the seasonal influences, since each of the twelve months was included in the total. For example, the demand for motor fuel for the twelve months of 1947 is shown in Table 14-6.

Table 14-6

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES, 1947

Month	Demand (Million barrels)
January	57.1
February	50.6
March	60.0
April	63.3
May	70.9
June	71.2
July	73.4
August	72.1
September	71.4
October	73.3
November	64.1
December	67.5
Total	794.9
Monthly average	66.24

SOURCE: Table 14-4.

July demand was 73.4 million barrels, or 110.8 per cent of the average monthly demand of 66.24 million barrels. What can we assume to be the cause of July demand being about 11 per cent more than the average monthly demand for the year? It was stated on page 248 that the forces affecting a time series may be classified as secular trend, seasonal, cyclical, and random. It was just stated that the average monthly figure of 66.24 million barrels could not be influenced by seasonal variation, so it might be that the difference between 73.4 and 66.24 was due entirely to seasonal factors. The possibility of this will be explored further in the following pages.

Chart 11-5 (page 251) shows that the trend of demand has risen, but of the twelve months averaged, six were immediately prior to July and five

were immediately after. This means that the influence of the rising trend on the average monthly demand of 66.24 is approximately the same as on the July demand, which means that *the ratio between the two amounts is not affected by the rising trend.* This fact can be further emphasized by reference to the demand for motor fuel in 1955 in Table 14-7.

Table 14-7

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES, 1955

Month	Demand (Million barrels)
January	96.4
February	88.5
March	105.7
April	111.1
May	115.7
June	120.7
July	115.7
August	121.8
September	113.4
October	112.6
November	109.2
December	111.0
Total	1,321.8
Monthly average	110.2

SOURCE: Table 14-4.

The total of the twelve months of 1955 is 1,321.8 million barrels and the monthly average is 110.2 million barrels, which is almost double the 1947 demand. However, July, 1955, demand was 115.7, also nearly double July, 1947. The ratio of July, 1955, demand to the monthly average of 1955 is 105.0 per cent. This is not greatly different from the 110.8 per cent for 1947.

The question naturally arises at this point as to whether or not the rise in the 1955 demand over 1947 might not have been cyclical. The answer to this is that it does not matter for this purpose whether the rise was trend or cyclical. The combined effect of both trend and cycle was exerted on the monthly average for 1955 to raise it to 1,321.8 million barrels, but at the same time essentially the same forces were causing an increase in July, 1955, demand. This means that the ratio of July, 1955, to the monthly average for 1955 is not appreciably affected by either trend or cyclical fluctuations.

Random fluctuations, however, cannot be assumed to be disposed of in computing the ratio between the demand in July and the average month of the year, since such fluctuations might affect one month alone, and have very little influence on the average for the year. Nevertheless, it does not seem necessary to attempt to remove the effect of random fluctuations, because in averaging the ratios, it is assumed that the random factors normally will cancel out. By their nature the random fluctuations will sometimes increase the demand for a given month and at other times decrease it.

If the monthly average for the year can be used as the base of the ratio for July, can it be used as the base of the ratios for the other months of the year? It is important that July is near the center of the twelve months averaged, since the effect of the trend and cycle on the months preceding and following July balance each other.[1] But if the other months of the year were compared with the average month, the influence of the trend and the cycle might become more pronounced. To avoid running any risk of this, it is customary to compute a new monthly average of twelve months for comparison with August sales. The twelve months used for this ratio would be February through January of the next year. This total of sales also includes twelve different months, and is used exactly as the total of the months from January through December in comparison with July.

The work sheet (Table 14-8) on pages 328 to 330 represents a systematic method of carrying out the computations. A twelve-month total must be taken for each month for which data are to be used. The total for 1946 is written opposite July of that year. This amount is 735.4 (all calculations are in millions of barrels). The total to be written opposite August is the total of the twelve months of which August is the seventh; in other words, the months from February, 1946, to and including January, 1947. The total for these twelve months is 740.8. It is not necessary to add these twelve months, since the total is different from the one written opposite July only in that the January, 1947, value is used instead of January, 1946. In other words, the simplest method of getting the moving total for the next month is to subtract the January, 1946, amount and add the January, 1947, amount. To get the total opposite September it is necessary only to subtract the February, 1946, amount and add the February, 1947, amount. This procedure is continued, giving the amounts in Column (3).

On page 325 the total of twelve months was divided by 12 to secure the average monthly demand, and this average was then divided into July demand to secure the ratio. Since these computations are almost always

[1]The fact that there are six months preceding July and only five following cannot be avoided, since there are only twelve months in a year. This small lack of symmetry is not important in its effect, but it is discussed further on pages 331 and 332.

Table 14-8

COMPUTATION OF THE RATIO OF DEMAND TO THE TWELVE-MONTH MOVING AVERAGE CENTERED AT THE SEVENTH MONTH

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Year and month (1)	Demand (Million barrels) (2)	12-month moving total centered at 7th month (3)	Demand as percentage of 12-month moving average (4)
1946 Jan.	51.7		
Feb.	47.7		
March	56.7		
April	62.1		
May	66.8		
June	63.3		
July	69.1	735.4	112.8
Aug.	66.7	740.8	108.0
Sept.	62.3	743.7	100.5
Oct.	66.6	747.0	107.0
Nov.	61.3	748.2	98.3
Dec.	61.1	752.3	97.5
1947 Jan.	57.1	760.2	90.1
Feb.	50.6	764.5	79.4
March	60.0	769.9	93.5
April	63.3	779.0	97.5
May	70.9	785.7	108.3
June	71.2	788.5	108.4
July	73.4	794.9	110.8
Aug.	72.1	799.1	108.3
Sept.	71.4	805.0	106.4
Oct.	73.3	813.2	108.2
Nov.	64.1	822.1	93.6
Dec.	67.5	828.4	97.8
1948 Jan.	61.3	835.2	88.1
Feb.	56.5	843.3	80.4
March	68.2	851.5	96.1
April	72.2	856.2	101.2
May	77.2	858.1	108.0
June	78.0	866.5	108.0
July	81.5	871.1	112.3
Aug.	80.3	872.9	110.4
Sept.	76.1	874.4	104.4
Oct.	75.2	879.5	102.6
Nov.	72.5	882.6	98.6
Dec.	72.1	887.1	97.5
1949 Jan.	63.1	892.5	84.8
Feb.	58.0	893.1	77.9
March	73.3	897.5	98.0
April	75.3	902.2	100.2
May	81.7	906.3	108.2
June	83.4	910.1	110.0
July	82.1	913.6	107.8
Aug.	84.7	917.4	110.8
Sept.	80.8	922.8	105.1
Oct.	79.3	928.2	102.5
Nov.	76.3	933.2	98.1
Dec.	75.6	940.5	96.5

Table 14-8 (Continued)

Year and month (1)	Demand (Million barrels) (2)	12-month moving total centered at 7th month (3)	Demand as percentage of 12-month moving average (4)
1950 Jan.	66.9	947.3	84.7
Feb.	63.4	956.9	79.5
March	78.7	966.7	97.7
April	80.3	972.7	99.1
May	89.0	982.5	108.7
June	90.2	988.9	109.5
July	91.7	994.4	110.7
Aug.	94.5	1008.2	112.5
Sept.	86.8	1017.4	102.4
Oct.	89.1	1025.4	104.3
Nov.	82.7	1032.4	96.1
Dec.	81.1	1042.8	93.3
1951 Jan.	80.7	1048.9	92.3
Feb.	72.6	1057.7	82.4
March	86.7	1064.3	97.8
April	87.3	1068.8	98.0
May	99.4	1080.2	110.4
June	96.3	1085.5	106.5
July	100.5	1089.6	110.7
Aug.	101.1	1096.0	110.7
Sept.	91.3	1105.7	99.1
Oct.	100.5	1106.2	109.0
Nov.	88.0	1117.7	94.5
Dec.	85.2	1118.8	91.4
1952 Jan.	87.1	1121.8	93.2
Feb.	82.3	1126.8	87.6
March	87.2	1128.7	92.7
April	98.8	1137.6	104.2
May	100.5	1141.0	105.7
June	99.3	1144.6	104.1
July	105.5	1156.2	109.5
Aug.	103.0	1157.2	106.8
Sept.	100.2	1159.5	103.7
Oct.	103.9	1169.0	106.7
Nov.	91.6	1170.3	93.9
Dec.	96.8	1174.0	98.9
1953 Jan.	88.1	1187.3	89.0
Feb.	84.6	1192.8	85.1
March	96.7	1196.6	97.0
April	100.1	1200.0	100.1
May	104.2	1199.5	104.2
June	112.6	1204.8	112.2
July	111.0	1205.7	110.5
Aug.	106.8	1207.5	106.1
Sept.	103.6	1209.1	102.8
Oct.	103.4	1213.9	102.2
Nov.	96.9	1217.7	95.5
Dec.	97.7	1217.9	96.3

Table 14-8 (Concluded)

Year and month (1)	Demand (Million barrels) (2)	12-month moving total centered at 7th month (3)	Demand as percentage of 12-month moving average (4)
1954 Jan.	89.9	1218.3	88.5
Feb.	86.2	1219.5	84.8
March	101.5	1222.9	99.6
April	103.9	1224.0	101.9
May	104.4	1226.2	102.2
June	113.0	1231.7	110.1
July	112.2	1238.3	108.7
Aug.	110.2	1244.8	106.2
Sept.	104.7	1247.1	100.7
Oct.	105.6	1251.3	101.3
Nov.	102.4	1258.5	97.6
Dec.	104.3	1269.8	98.6
1955 Jan.	96.4	1277.5	90.6
Feb.	88.5	1281.0	82.9
March	105.7	1292.6	98.1
April	111.1	1301.3	102.5
May	115.7	1308.3	106.1
June	120.7	1315.1	110.1
July	115.7	1321.8	105.0
Aug.	121.8	1326.0	110.2
Sept.	113.4	1335.6	101.9
Oct.	112.6	1343.0	100.6
Nov.	109.2	1344.9	97.4
Dec.	111.0	1353.3	98.4
1956 Jan.	100.6	1360.0	88.8
Feb.	98.1	1365.5	86.2
March	113.1	1369.9	99.1
April	113.0	1369.2	99.0
May	124.1	1376.7	108.2
June	127.4	1380.1	110.8
July	121.2	1377.3	105.6
Aug.	126.2		
Sept.	112.7		
Oct.	120.1		
Nov.	112.6		
Dec.	108.2		

SOURCE: Table 14-4.

performed on a calculator, the ratio may be more easily computed by first multiplying the monthly demand by twelve, leaving this product in the machine, and dividing it by the total of twelve months' demand. The only amount that need be written down is the ratio, which is the only value wanted. That the operations are identical is shown by the following:

$$\text{Ratio} = \text{July demand} \div \frac{\text{1946 demand}}{12} = \text{July demand} \times \frac{12}{\text{1946 demand}}$$

$$\text{Ratio} = 69.1 \div \frac{735.4}{12} = 69.1 \times \frac{12}{735.4} = 1.128$$

The ratios in Column (4) of Table 14-8 represent the effect of the seasonal and random forces on each month from July, 1946, to July, 1956. The next step is to derive an average of the specific seasonals for each month. These computations are shown in a later section on pages 333 to 338.

Ratio to Two-Item Average of Twelve-Month Moving Average

On page 327 it was pointed out that when the twelve-month average was related to July, six of the months preceded July and five months were later than July. It would be more logical if there were the same number of months preceding and following the month with which the average was compared. If thirteen months were used, the seasonal influence would not all be removed. For example, if the thirteen months from January, 1947, through January, 1948, were used, the seasonal effect of January would be included twice; but the seasonal effects of the other months would be included only once However, an average of thirteen months might be computed by giving each of the two months that were farthest away from the center a weight of ½. For example July, 1947, demand would be compared with the average of the demand shown in Table 14-9.

Table 14-9

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES

JANUARY, 1947, TO JANUARY, 1948

Month and year		Demand (Million barrels)
1947	January (½ of 57.1)..	28.55
	February.....	50.6
	March.....	60.0
	April.....	63.3
	May.....	70.9
	June.....	71.2
	July.....	73.4
	August.....	72.1
	September.....	71.4
	October.....	73.3
	November.....	64.1
	December.....	67.5
1948	January (½ of 61.3)..	30.65
	Total.....	797.00
	Monthly average.....	66.42

Source: Table 14-4.

The ratio of July demand to the twelve-month average computed in Table 14-9 is 110.5 per cent. (Ratio $= \frac{73.4}{66.42}\ 100 = 110.5$) On page 325 this ratio was found to be 110.8 per cent, based on the twelve full months from January through December. Demand in January, 1947, was 7 per cent more than January, 1946, which suggests that a cyclical-trend[2] rise occurred during 1947. This means that with only *five months following July and six months preceding July*, the average of the twelve months did not reflect the full effect of the cyclical-trend rise. The monthly average was 66.42 compared to 66.24 for the twelve months of 1947, and the difference of .3 per cent between the two ratios was due to this difference in monthly averages. The specific seasonals computed by this method are a little more precise measures of the seasonal factor, but the increase in the accuracy of the index of seasonal variation is usually so small that it is unimportant. This method, however, is used by many statisticians in spite of the increased clerical work.

The easiest method of computing the ratios described above is shown in Table 14-10. First, a twelve-month moving total is taken and entered between the sixth and seventh months. The total for 1947 is 794.9, and it is entered between June and July. The total for February, 1947, through January, 1948, is 799.1 and is entered between July and August. These are the same totals that were found in Table 14-8. The next step is to add the total that was entered between June and July to the total between July and August. This new total is entered opposite July; it consists of twenty-four months; but January, 1947, has been included once; January, 1948, has been included once; and the months between have each been included two times. This gives a total double the total in Table 14-9. To compute the moving average, this total must be divided by 24 instead of 12.

$$\text{Monthly average} = \frac{1{,}594.0}{24} = 66.42 \text{ (the same as found in Table 14-9)}$$

If the ratios are to be computed on a calculator, it may be done directly from the totals. The demand for the month should be entered on the calculator by multiplying by 24, in the same manner as described on page 330, where the demand figure was multiplied by 12. The product should be left in the machine and divided by the sum of the two moving totals. For July, 1947, this calculation would be as follows:

$$\text{Ratio} = 73.4 \div \frac{1{,}594.0}{24}\ 100 = 73.4 \times \frac{24}{1{,}594.0}\ 100 = 110.5 \text{ per cent}$$

[2]Whether it was trend or cycle, or both, is not important in this connection.

Table 14-10

COMPUTATION OF THE RATIOS TO THE TWELVE-MONTH MOVING AVERAGE CENTERED

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 AND 1947

Month and year	Demand (Million barrels)	12-month moving total centered	Two-item moving total	Ratio to 12-month moving average
1946 January	51.7			
February	47.7			
March	56.7			
April	62.1			
May	66.8			
June	63.3			
July	69.1	735.4	1,476.2	112.3
August	66.7	740.8	1,484.5	107.8
September	62.3	743.7	1,490.7	100.3
October	66.6	747.0	1,495.2	106.9
November	51.3	748.2	1,500.5	98.0
December	61.1	752.3	1,512.5	97.0
1947 January	57.1	760.2	1,524.7	89.9
February	50.6	764.5	1,534.4	79.1
March	60.0	769.9	1,548.9	93.0
April	63.3	779.0	1,564.7	97.1
May	70.9	785.7	1,574.2	108.1
June	71.2	788.5	1,583.4	107.9
July	73.4	794.9	1,594.0	110.5
August	72.1	799.1		
September	71.4			
October	73.3			
November	64.1			
December	67.5			
1948 January	61.3			

SOURCE: Table 14-4.

AVERAGING SPECIFIC SEASONALS

It was pointed out on page 321 that the problem of measuring seasonal variation was a problem of averages. The previous section has described methods of isolating the specific seasonal fluctuations for individual years. This section will discuss the problem of how to average the specific seasonals that have been derived.

Table 14-8 gives the ratios to the twelve-month moving average centered at the seventh month, for the complete years 1947 through 1955, and parts of 1946 and 1956. It will be noted in Table 14-8 that the twelve-month moving average centered at the seventh month always ends five months before the end of the last year and begins six months after the first month of the first year.

Table 14-11

ARRAYS OF RATIOS TO TWELVE-MONTH MOVING AVERAGE CENTERED AT SEVENTH MONTH

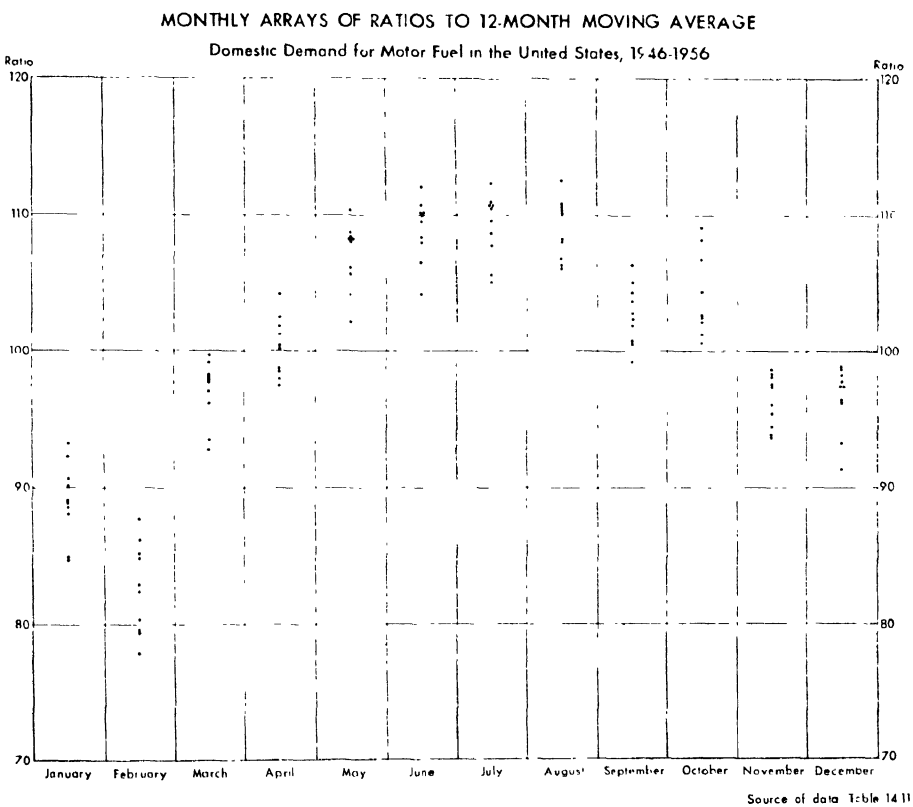
DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
	84.7	77.9	92.7	97.5	102.2	104.1	105.0	106.1	99.1	100.6	93.6	91.4
	84.8	79.4	93.5	98.0	104.2	106.5	105.6	106.2	100.5	101.3	93.9	93.3
	88.1	79.5	96.1	99.0	105.7	108.0	107.8	106.8	100.7	102.2	94.5	96.3
	88.5	80.4	97.0	99.1	106.1	108.4	108.7	108.0	101.9	102.5	95.5	96.5
	88.8	82.4	97.7	100.1	108.0	109.5	109.5	108.3	102.4	102.6	96.1	97.5
	89.0	82.9	97.8	100.2	108.2	110.0	110.5	110.2	102.8	104.3	97.4	97.5
	90.1	84.8	98.0	101.2	108.2	110.1	110.7	110.4	103.7	106.7	97.6	97.8
	90.6	85.1	98.1	101.9	108.3	110.1	110.7	110.7	104.4	107.0	98.1	98.4
	92.3	86.2	99.1	102.5	108.7	110.8	110.8	110.8	105.1	108.2	98.3	98.6
	93.2	87.6	99.6	104.2	110.4	112.2	112.3	112.5	106.4	109.0	98.6	98.9
Mean.....	89.01	82.62	96.96	100.37	107.00	108.97	109.16	109.00	102.70	104.44	96.36	96.62
Median.....	88.90	82.65	97.75	100.15	108.10	109.75	110.00	109.25	102.60	103.45	96.75	97.50
Mean of 4 central items..	89.10	82.63	97.63	100.15	107.63	109.50	109.85	109.23	102.70	104.03	96.65	97.33

SOURCE: Table 14-8.

If the percentages showing the specific seasonal for each month are to be averaged to find a typical seasonal, it is necessary that there be a significant concentration of the individual items for each month. If there is too much dispersion among the percentages, it is incorrect to use any average as typical of the whole group. The whole question of the dispersion in items being averaged was discussed in Chapter 9. In order to find out whether there is sufficient concentration to warrant averaging the specific seasonals, the ratios for each month were put into an array and a chart drawn of the array. Table 14-11 shows the arrays, and Chart 14-2 presents them graphically. If the chart shows enough concentration for an average to be significant, the next step is to find an average of the ratios for each month.

Chart 14-2



Arithmetic Mean

In the chapter on averages it was stated that the arithmetic mean is the best known and most generally useful average; the arithmetic mean of the ratios has therefore been computed first. The arithmetic mean of the ratios

for each month has been entered in Column (2) of Table 14-12. These means total 1,203.21, with an average monthly value of 100.27. On page 318 it was pointed out that the index of seasonal variation could be made to total 1,200, which would give a monthly average of exactly 100. In order to make the individual monthly indexes total 1,200, each one is reduced proportionately. The desired total of 1,200 is 99.73 per cent of the actual total of 1,203.21; so the level of the index is reduced by multiplying each monthly value by .9973. The adjusted index of Column (3) in Table 14-12 totals 1,200.

Table 14-12

COMPUTATION OF INDEX OF SEASONAL VARIATION

ARITHMETIC MEAN OF RATIOS

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Month (1)	Mean of ratios to 12-month moving average (2)	Index of seasonal variation Average month = 100 (3)
January	89.01	88.8
February	82.62	82.4
March	96.96	96.7
April	100.37	100.0
May	107.00	106.7
June	108.97	108.7
July	109.16	108.9
August	109.00	108.7
September	102.70	102.4
October	104.44	104.2
November	96.36	96.1
December	96.62	96.4
Total	1,203.21	1,200.0

SOURCE: Table 14-11.

$$\text{Leveling factor} = \frac{1{,}200}{1{,}203.21} = .9973$$

Median

Since there is always danger that an extreme value will have an undue amount of influence on the arithmetic mean, the median is frequently used instead of the mean to average specific seasonals. In Table 14-13 the median

for each month has been entered from Table 14-11, and these twelve medians reduced to an index of seasonal variation in the same manner as in Table 14-12. It was stated on page 183 that the median cannot logically be used in further calculations. The fact, however, that the median may be more nearly typical than the mean frequently outweighs this objection.

Table 14-13

COMPUTATION OF INDEX OF SEASONAL VARIATION

MEDIAN OF RATIOS

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Month	Median of ratios to 12-month moving average	Index of seasonal variation Average month = 100
January	88.90	88.4
February	82.65	82.2
March	97.75	97.2
April	100.15	99.6
May	108.10	107.5
June	109.75	109.1
July	110.00	109.4
August	109.25	108.6
September	102.60	102.0
October	103.45	102.9
November	96.75	96.2
December	97.50	96.9
Total	1,206.85	1,200.0

SOURCE: Table 14-11.

$$\text{Leveling factor} = \frac{1,200}{1,206.85} = .9943$$

Modified Mean

Because there is always a possibility that a few unusually large or small items may unduly influence the mean, a modification is sometimes used to average specific seasonals. Since the number of specific seasonals being averaged is small, an exceptionally large or small one in only a few months could have considerable effect on the index of seasonal variation computed by the use of the arithmetic mean. The *modified mean* used is an arithmetic mean of only the central items, eliminating an equal number of the largest

and of the smallest values to be averaged. If fifteen specific seasonals were being averaged, the largest five and the smallest five might be eliminated, and the average based on the central five. This average has some of the characteristics of the median, since it will be less influenced by extremely large or small items than the mean; but since it is the arithmetic mean of several central items, it is less likely to be erratic than the median.

The choice of the number of central items to be used is arbitrary, but what number is taken makes very little difference in the result. Averaging the central four or five items is ordinarily satisfactory. In Table 14-14 the means of the four central ratios have been averaged. These modified means were then leveled to total 1,200 in the same way as in Tables 14-12 and 14-13.

Table 14-14

COMPUTATION OF INDEX OF SEASONAL VARIATION

MODIFIED MEAN OF RATIOS

DOMESTIC DEMAND FOR MOTOR FUEL
IN THE UNITED STATES: 1946 TO 1956

Month	Modified mean of ratios to 12-month moving average	Index of seasonal variation Average month = 100
January	89.10	88.6
February	82.63	82.2
March	97.63	97.1
April	100.15	99.6
May	107.63	107.0
June	109.50	108.9
July	109.85	109.3
August	109.23	108.7
September	102.70	102.2
October	104.03	103.5
November	96.65	96.1
December	97.33	96.8
Total	1,206.43	1,200.0

SOURCE: Table 14-11.

$$\text{Leveling factor} = \frac{1{,}200}{1{,}206.43} = .9947$$

EVALUATION OF THE DIFFERENT METHODS OF MEASURING SEASONAL VARIATION

Comparing the three indexes of seasonal variation from Tables 14-12, 14-13, and 14-14 indicates that the method of averaging the specific seasonals was not an important factor in this case; there is so little difference in the results that no one method could be chosen as outstandingly better. The method of isolating the specific seasonals, however, seems to have had more effect on the index. In Table 14-15 the index based on the means of the actual monthly items is compared with the index based on the means of the ratios to the twelve-month moving average; and it is seen that there is considerable variation between the two. This difference is especially marked in the months at the beginning and at the end of the year. It is

Table 14-15

COMPARISON OF TWO INDEXES OF SEASONAL VARIATION

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Month (1)	Based on means of actual monthly demand (2)	Based on means of ratios to 12-month moving average (3)	Difference between indexes (2) — (3)
January	86.5	88.8	— 2.3
February	80.9	82.4	— 1.5
March	95.2	96.7	— 1.5
April	99.2	100.0	— .8
May	106.1	106.7	— .6
June	108.3	108.7	— .4
July	109.1	108.9	.2
August	109.5	108.7	.8
September	102.9	102.4	.5
October	105.6	104.2	1.4
November	98.2	96.1	2.1
December	98.5	96.4	2.1

SOURCE: Tables 14-5 and 14-12.

significant that the index based on the means of the actual monthly items is below the index based on the means of the ratios to the twelve-month

moving average at the beginning of the year, and above it at the end of the year. It seems likely that this is the case because the means of the monthly items are affected by trend, while the ratios to the twelve-month moving average are not. The upward trend in the demand for motor fuel has tilted slightly upward the index of seasonal variation based on the means of monthly items.

Because of the greater accuracy of the specific seasonals when the ratios to the twelve-month moving average are used, this method is much to be preferred to averaging the actual monthly data. It is possible to remove the effect of the trend from the averages instead of from the specific seasonals before they are averaged, and methods of doing this have been worked out by statisticians. But the methods that use some form of the ratio to the twelve-month moving average are so generally satisfactory that they have become practically standard for measuring seasonal variation.

CHANGING SEASONAL PATTERN

Abrupt Changes

Seasonal fluctuations in business that result from customs, habits, or business practices may be changed abruptly when something happens to affect those habits or practices. A standard example was the change in the date of the New York Automobile Show in the winter of 1935–36. The seasonal fluctuations in the manufacture and sale of new automobiles are normally greatly influenced by the practice of introducing yearly models. Volume of production and sales fall off at the end of the model year and increase rapidly at the beginning of the new model year. The New York Automobile Show served as the official introduction of the new models, and the change of its date from January to November brought about a radical change in the whole seasonal pattern of the industry. In fact, one reason for favoring the change was the belief that the production of cars would be greater in the winter months, which would increase employment at a season of the year when many other industries were at a seasonal low point. Although new car models are not now all introduced at the same time, they usually make their appearance within a month or two of each other.

The problem of measuring such a sudden change in seasonal fluctuations is usually met by constructing a new index of seasonal variation beginning with the date of the change. This index can be compared with the index based on the data up to the date of the change, to measure the change in the seasonal pattern. However, until several years have elapsed it is difficult to do more than forecast what the seasonal pattern will be.

Gradual Change

In computing any index of seasonal variation by averaging specific seasonals, it is important that the period covered show no important shift in the seasonal pattern. When such a shift occurs gradually, it becomes necessary finally to compute a new index based on recent years. An example is shown in Table 14-16 and Chart 14-3. The two indexes of seasonal variation for demand for motor fuel are based on two periods. The first period covers the years 1930 to 1944, and the second, the years 1946 to 1956.

Some statistical organizations have had a policy of computing a new index of seasonal variation yearly by adding the most recent year and dropping the earliest, thus keeping the average based on an up-to-date group of specific seasonals.

Table 14-16

COMPARISON OF INDEXES OF SEASONAL VARIATION*: 1930 TO 1944 AND 1946 TO 1956

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES

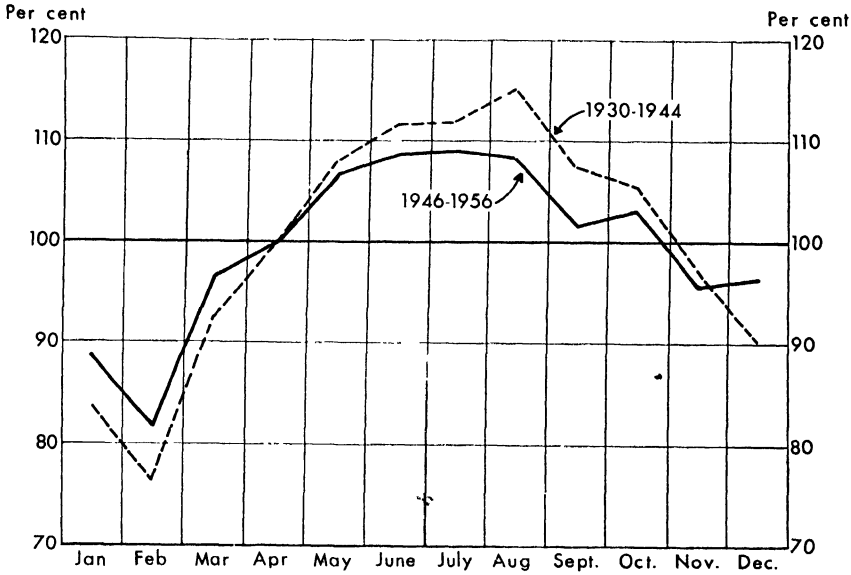
Month	Indexes of seasonal variation	
	1930 to 1944	1946 to 1956
January	83.5	88.6
February	76.7	82.2
March	92.6	97.1
April	99.8	99.6
May	107.7	107.0
June	111.6	108.9
July	112.0	109.3
August	115.2	108.7
September	107.5	102.2
October	105.8	103.5
November	97.0	96.1
December	90.6	96.8

*Indexes computed by method of ratios to 12-month moving average, using the modified mean (four central ratios).

SOURCE: 1930 to 1944 — Computed from data in *Survey of Current Business*.
1946 to 1956 — Table 14-14.

Chart 14-3

INDEXES OF SEASONAL VARIATION IN DEMAND FOR MOTOR FUEL IN THE UNITED STATES 1930-1944 AND 1946-1956



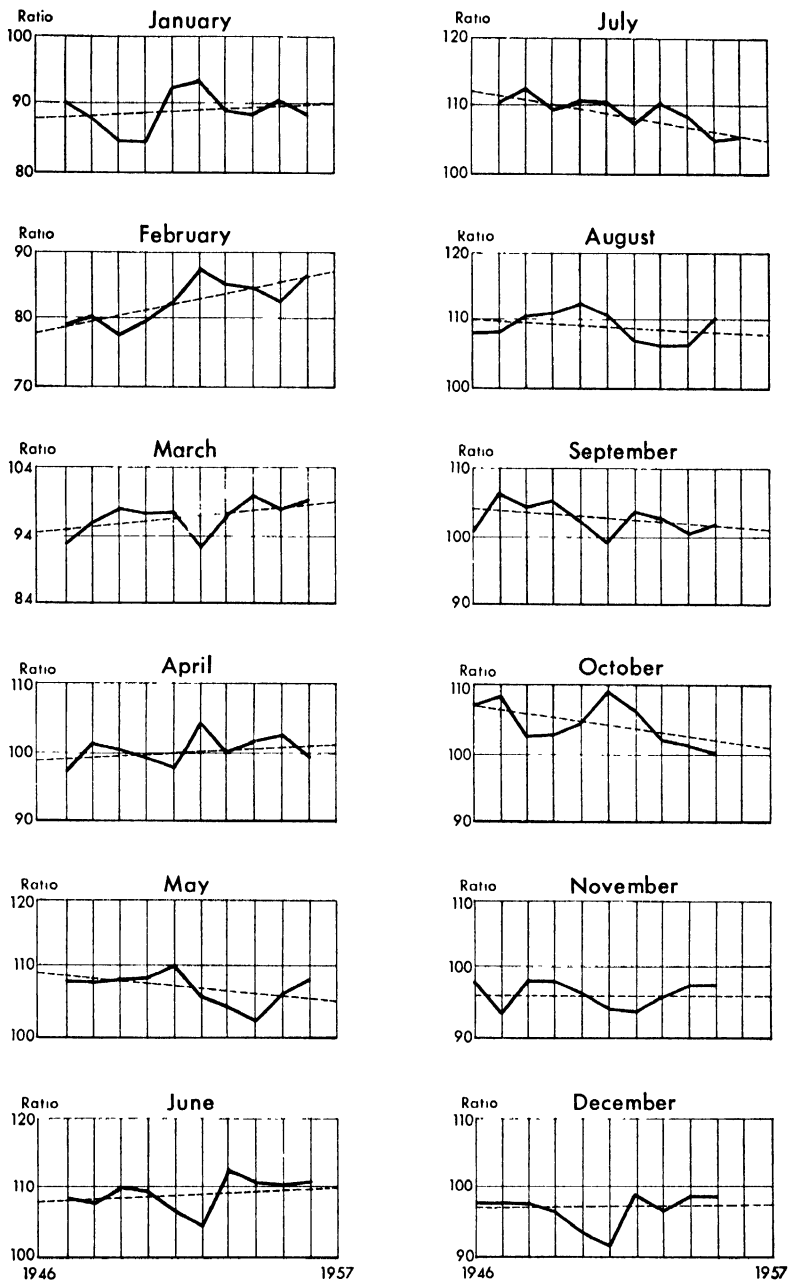
Source of data: Table 14-16

Because business is constantly changing, it is perhaps normal for the seasonal pattern to shift over a period of years. The improvement in roads and in automobiles has caused a gradual leveling of the seasonal fluctuations in the demand for motor fuel and the business of supplying it. Consumer buying habits for any product may change and cause a shift in the seasonal pattern of retail trade and the businesses directly affected by it. Many industries have made an effort to level out the seasonal fluctuations by trying to stimulate business in the seasonal low points, or by adding lines with different seasonal peaks and low points.

If there is reason to suspect that a shift has occurred in the seasonal pattern, a test should be made before averaging the specific seasonals. The simplest method of doing this is to plot the specific seasonals for each month chronologically. This has been done with the demand for motor fuel in Chart 14-4 for each of the twelve months. Inspection of these graphs makes it possible to decide whether there has been any shift in the seasonal behavior of a month. If the specific seasonals for any month show a persistent movement either upward or downward, a trend line should be used to describe the seasonal instead of an average of the specific seasonals.

Chart 14-4

COMPUTATION OF CHANGING SEASONAL INDEX
Domestic Demand for Motor Fuel in the United States, 1946-1956



Source of data: Table 14-8

Any of the methods given in Chapters 12 or 13 for computing the trend might be used, but the most common in this connection is the graphic method. In Chart 14-4 the trends have been computed graphically and the values read from the chart. These trend values, entered in Table 14-17, become the indexes of seasonal variation for the different years. After the values for the curve were read from the chart, they were entered in Table 14-17, and then adjusted until the indexes for each year totaled exactly 1,200, or very close to this amount. This adjustment is usually made by merely shifting the curves up or down slightly to make the totals for each year 1,200, rather than using any method of calculation.

Table 14-17

INDEXES OF SEASONAL VARIATION

DOMESTIC DEMAND FOR MOTOR FUEL, IN THE UNITED STATES: 1946 TO 1956

AVERAGE MONTH = 100

Month	1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957
January.....	*(88)	88	88	88	88	89	89	89	89	90	90	(90)
February....	(78)	79	80	80	81	82	83	84	84	85	86	(87)
March.....	(95)	95	95	96	96	97	97	97	98	98	98	(99)
April.....	(99)	99	99	99	100	100	100	100	101	101	101	(101)
May.....	(109)	108	108	108	107	107	107	106	106	105	105	(105)
June.....	(107)	108	108	108	108	109	109	109	109	110	110	(110)
July.....	(112)	112	111	110	110	109	109	108	107	107	106	(105)
August.....	110	110	109	109	109	109	109	108	108	108	(108)	(108)
September...	104	104	103	103	103	103	102	102	102	101	(101)	(101)
October.....	107	107	106	106	105	104	104	103	103	102	(102)	(101)
November...	96	96	96	96	96	96	96	96	96	96	(96)	(96)
December...	97	97	97	97	97	97	97	97	97	97	(98)	(98)

*The values in parentheses have been extrapolated.

SOURCE: Table 14-8.

MAKING USE OF A MEASURE OF SEASONAL VARIATION

One use of an index of seasonal variation is in making the monthly plans of a business. (An index of seasonal variation on a quarterly or weekly basis is sometimes used, but it is the most common practice to make such plans on a monthly basis.) In a business that suffers no seasonal variation, planning the operations for the coming year is much less com-

plicated than it is in a highly seasonal business when working capital needs fluctuate widely with different months and plans must be made to have adequate funds on hand when they are needed. Part of the basic information an executive should have about his business is the seasonal pattern of the whole industry, and the degree of resemblance or difference that his own business exhibits. Quotas, sales plans, advertising campaigns, financial budgets, and production schedules are all based on this information.

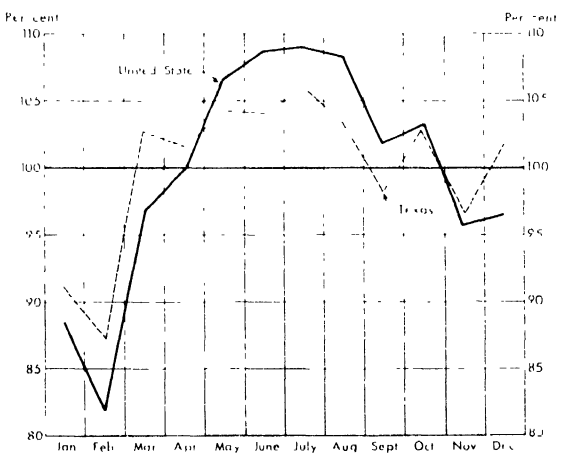
The seasonal pattern in advertising is usually made to coincide with the seasonal fluctuations in sales, perhaps with advertising running a little in advance. This would be done on the theory that the time to put forth the greatest sales effort is the part of the year when people want to buy the product. Sometimes an advertising campaign is undertaken to stimulate sales in the dull periods. Reducing prices in the seasons with reduced sales may be used as a method of increasing business.

How exactly a measure of seasonal variation needs to be computed depends upon the regularity with which the seasonal movement occurs and upon the importance of this factor in the operations of the business. A well-informed executive no doubt knows with considerable accuracy the seasonal pattern in the fluctuations of his business. Sometimes, however, he only thinks he knows; it is notoriously true that general impressions may be far wrong.

The consumption of many products shows a seasonal pattern which influences retail distribution. Gasoline is one of these commodities. Chart 14-5 shows the seasonal variation in the retail distribution of gasoline in all states and in Texas. There is a more pronounced variation in the United States total than in Texas. This doubtlessly results from the fact that automobiles can be used more generally the year round in Texas than in all of the United States.

Chart 14-5

INDEXES OF SEASONAL VARIATION IN DEMAND FOR MOTOR FUEL IN THE UNITED STATES AND TEXAS

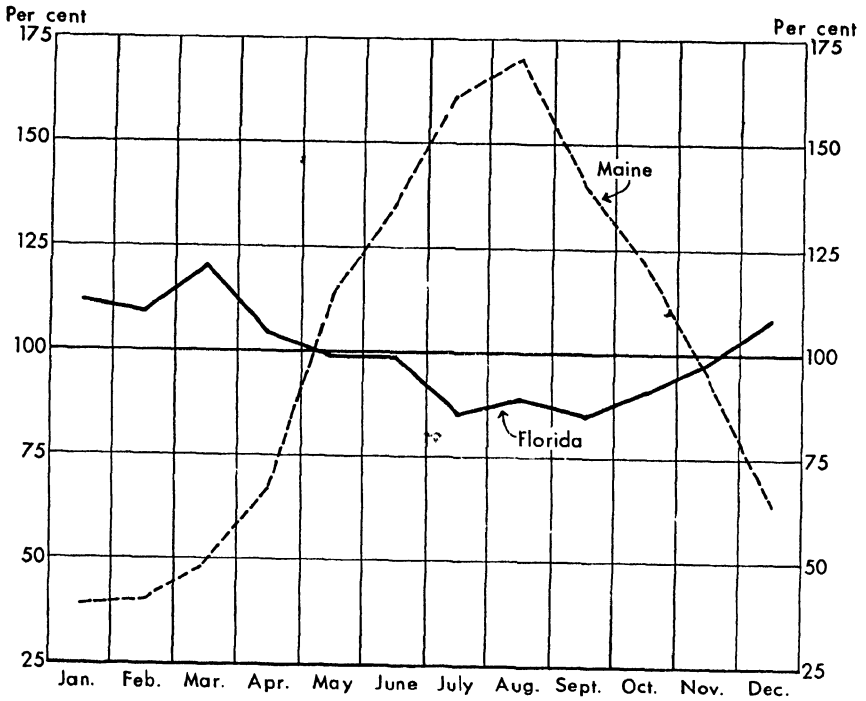


Source of data: U.S. Table 14-14
Texas: Bureau of Business Research, The University of Texas

Chart 14-6 shows the seasonal variation in gasoline consumption in Florida and Maine, one a winter resort and the other a summer resort state. The fluctuations in Maine are particularly striking.

Chart 14-6

INDEXES OF SEASONAL VARIATION IN GASOLINE CONSUMPTION IN FLORIDA AND MAINE

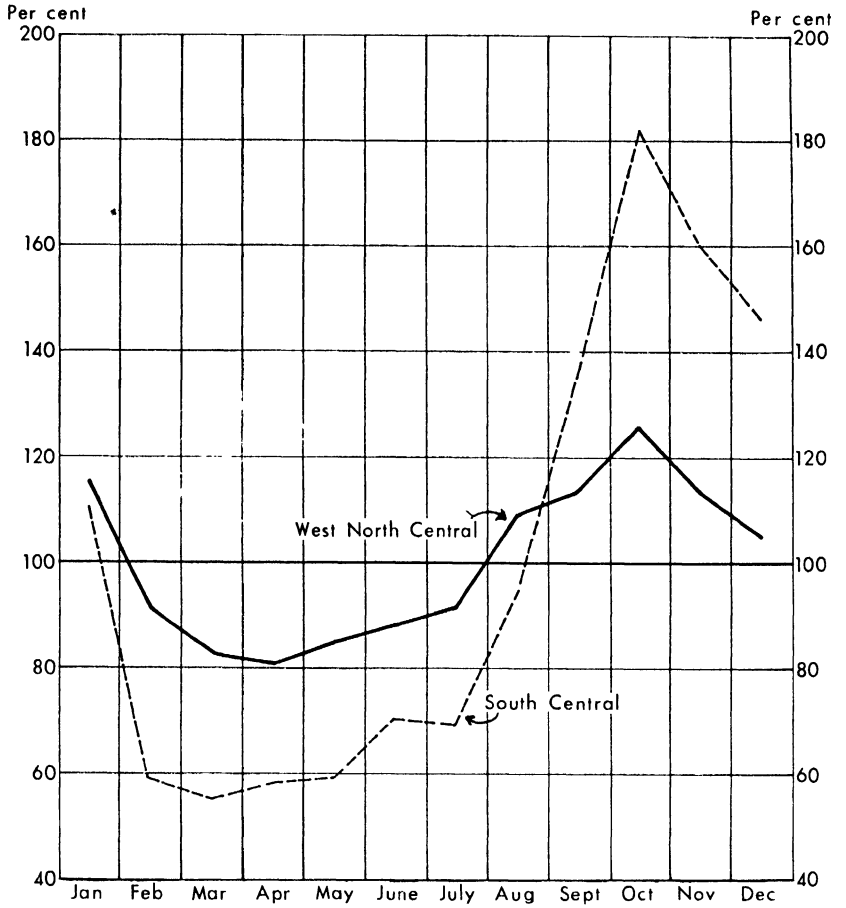


Source of data: Simon Kuznets, *Seasonal Variations in Industry and Trade*

Chart 14-7 shows the seasonal variation in farm income in the West North Central states and in the South Central states. The fact that the West North Central states produce a large percentage of the livestock and livestock products in the country tends to make for stability of farm income in the different months of the year. The farmers in the South Central states receive a predominant proportion of their income from cotton, which is sold as soon as it is picked. This causes the pronounced peak of income in the fall. Advertising and sales plans of companies selling in the farm market should be adjusted to the seasonal pattern of the farmers' income in different sections. In some sections, farm income is more evenly distributed throughout the year than in others. In this connection it may be enough for a sales executive to have fairly accurate general knowledge but it is better to have indexes showing this information in a specific, quantitative form.

Chart 14-7

INDEXES OF SEASONAL VARIATION IN FARM CASH INCOME IN WEST NORTH CENTRAL AND SOUTH CENTRAL REGIONS



Source of data Indexes computed from data published in *Farm Income Situation*

Measuring Differences in Seasonality

At times it is important to have a method of summarizing very concisely the extent of the seasonal variation in different series. Variation in employment is a serious problem for many industries. A study of this matter would ordinarily require some method of measuring the degrees of fluctuation in various industries. Table 14-18 shows the differences in Iowa, as revealed by a study of the labor market in that state.

The measurement of the degree of seasonality is merely an application of the measurements of dispersion discussed in Chapter 9. The second column shows the *range of the seasonal index for each industry*, that is,

Table 14-18

VARIABILITY OF EMPLOYMENT IN VARIOUS INDUSTRIAL GROUPS IN IOWA

Group	Range of seasonal index	Average deviation of seasonal index from mean
Manufacturing	17.4	4.30
Retail trade	9.7	1.78
Wholesale trade	15.0	3.92
Service industries	9.7	2.62
Agriculture	43.2	14.30
Mining	28.0	10.35
Construction	53.4	17.00
All industries	23.3	7.57

SOURCE: Herbert W. Bohlman, *Labor Market in Iowa: Characteristics and Trends* (Des Moines, Iowa, 1937), p. 31.

the difference between the largest and the smallest index numbers. The range is one measure of dispersion, but it is based on only two items and so it tells nothing about the scatter in the remaining items. The third column shows the *average deviation of the index numbers, computed from the mean of the twelve indexes for the year*. Since the mean of the twelve indexes is 100, it is easy to find the deviation of the individual months from the mean, and then to compute the mean of these deviations.

Table 14-19 illustrates the computation of the average deviation of the index from the mean, using the index of seasonal variation computed for the demand for motor fuel for the period 1946 to 1956. The range of this index is 27.1, the difference between the February index and the July index. The average deviation as computed in Table 14-19 gives the average amount that the indexes deviate from 100 and represents the degree of seasonality in the series.

By this method careful comparisons can be made between the seasonality in different series for which indexes of seasonal variation have been computed. The average deviation for the index of seasonal variation based on 1930 to 1944 demand for motor fuel is 10.0. The statement was made

Table 14-19

COMPUTATION OF THE AVERAGE DEVIATION OF THE INDEX OF SEASONAL VARIATION

DOMESTIC DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1956

Month	Index of seasonal variation 1946 to 1956	Deviations from the mean* (100)
January	88.6	11.4
February	82.2	17.8
March	97.1	2.9
April	99.6	.4
May	107.0	7.0
June	108.9	8.9
July	109.3	9.3
August	108.7	8.7
September	102.2	2.2
October	103.5	3.5
November	96.1	3.9
December	96.8	3.2
Total	1,200.0	79.2
Mean	100.0	6.6

*Sign ignored.

SOURCE: Table 14-14.

above that the seasonal variation in demand for motor fuel has been reduced, and a comparison with the standard deviation of 6.6 for the years 1946 to 1956 gives an objective measure of the extent of this reduction.

Using a Measure of Seasonal Variation in Forecasting

It was stated in Chapter 11 that one important reason for measuring the changes in time series is the need for forecasting. This is true of seasonal fluctuations in data, since all budgeting and planning that is done on a monthly basis must make provision for the variations from month to month due to the seasonal changes.

If the seasonal pattern shows no indication of undergoing a sudden change, it is usually safe to assume at a given date that the following year will be the same as the preceding year. If there is any indication that an abrupt change has occurred, this of course will not be true, and other methods must be employed to predict what the new seasonal pattern will be.

But a basic assumption in the computation of an index of seasonal variation was that the seasonal pattern is stable enough to be averaged in order to secure the typical seasonal. When this is true and there is no evidence that any pronounced change is occurring, this typical seasonal is a reasonable forecast of the variation for the coming year.

If the pattern shows gradual change, it will be somewhat more accurate to use a changing index, which means a different index each year. When the change from one year to the next is slight, however, it is usually sufficient to use the most recent index as a forecast of the following year. For example, the index for 1956 in Table 14-17 might be used as a forecast of the index for 1957.

When a sudden change in the seasonal pattern occurs, as with the change in the date of the introduction of new models in the automobile industry, any forecast of the new seasonal pattern will be largely an estimate until some time has elapsed and it is possible to measure the actual seasonal variations.

When a forecast of the index has been made, this index may be used to put the seasonal movement into the forecast for periods of less than a year. For example, if an oil company forecast sales of gasoline to be 600 million gallons in 1957, the index of seasonal variation could be used to break this down into a forecast of monthly sales. On the basis of our knowledge of the seasonal pattern of demand for motor fuel, we would not forecast that each month would be one-twelfth of the annual total; January typically accounts for much less than one-twelfth and July for much more than one-twelfth.

Using the index based on the average for 1946 to 1956, the calculations in Table 14-20 are made as a forecast of the monthly sales if the total for the year 1957 was 600 million gallons. If there were no seasonal variation, a reasonable forecast for each month would be 50 million gallons, or one-twelfth of the annual total. But January is expected to be only 88.6% of the average month, so the forecast for January is 88.6% of 50 million gallons or 44.3 million gallons. The forecast for each month is made in the same way, with the result that the total of 600 million gallons is divided among the months on the basis of the forecast of the seasonal pattern for 1957.

A somewhat different use of the index of seasonal variation in forecasting or estimating is made in the following case. Assume that a statewide highway planning survey is under way and data are being collected on the volume of traffic. After a year, the seasonal fluctuations on different types of roads are measured and indexes are computed. The traffic count is made with an automatic counting device, but the cost of leaving such a device on every road for twelve months is high. Instead of counting for a

Table 14-20

FORECAST OF MONTHLY SALES OF MOTOR FUEL FOR 1957, BASED ON ESTIMATED ANNUAL TOTAL OF 600 MILLION GALLONS

Month	Index of seasonal variation (Forecast for 1957)	Estimated monthly sales (million gallons) (Index of seasonal variation times 50 million)
January	88.6	44.3
February	82.2	41.1
March	97.1	48.6
April	99.6	49.8
May	107.0	53.5
June	108.9	54.4
July	109.3	54.6
August	108.7	54.3
September	102.2	51.1
October	103.5	51.8
November	96.1	48.1
December	96.8	48.4
Total	...	600.00

SOURCE: Table 14-14.

full year on every road, it would be satisfactory in many cases to count for only a month, and from this count estimate the total annual volume. However, since traffic probably varies with the seasons, the seasonal pattern should be taken into account in making the estimate.

Assume that the traffic on a particular road showed 3,685 cars for July. What would be a reasonable estimate for a year? Obviously it would not be satisfactory to multiply 3,685 by twelve, since traffic for July is undoubtedly higher than for the average month. The effect of the seasonal variation should be taken into account, and if the index for this particular type of road shows July traffic was typically 25 per cent above the average, it is possible to estimate the average monthly traffic from the July count. From this average the annual volume can then be computed.

The statement that the July traffic is 25 per cent above the average means that the index of seasonal variation for July is 125. The count of 3,685 cars represents 125 per cent of the average month, and the problem is to compute the average month.

$$125\% = 3{,}685$$
$$1\% = 3{,}685 \div 125 = 29.48$$

therefore

$$100\% = 29.48 \times 100 = 2{,}948$$

If the traffic in the average month is 2,948, the total traffic for the year is 2,948 times twelve, or 35,376.

An estimate of the volume or traffic for each month also could be made from the index of seasonal variation, in the same way as the estimates of the sale of motor fuel for the twelve months were made in Table 14-20.

ADJUSTING DATA FOR SEASONAL VARIATION

An index of seasonal variation is also used to remove from a time series the fluctuations due to seasonal factors. This is one of the most important uses of such an index, but it will be discussed in the following chapter as one of the steps in isolating the cyclical fluctuations in a series.

STUDY QUESTIONS

1. Why is the measurement of seasonal variation a problem of averages?

2. What factors should be considered in choosing the length of the period to use in computing an index of seasonal variation?

3. Define the term "specific seasonals."

4. What are the advantages of determining specific seasonals by expressing each month as a percentage of the twelve-month moving average instead of expressing all the months in a year as percentages of the average for that year?

5. What is the advantage of using as specific seasonals the ratios to the twelve-month moving average, centered by taking a two-item moving average, over the twelve-month moving average centered at the seventh month? How significant do you consider this advantage?

6. What are the relative advantages of the following methods of averaging specific seasonals to find the index of seasonal variation?

(a) Mean.
(b) Median.
(c) Modified mean.

7. How can one determine whether to use a changing index of seasonal variation instead of computing an average index for a number of years?

8. Assume that a manufacturing concern was trying to reduce the amplitude of the seasonal swings in its production rate by adding new products with different seasonal patterns. After a substantial volume of business has been built up for the new products, how would you determine whether or not the diversification of products had succeeded in reducing the seasonal swings of the business?

9. How accurately do you think a businessman can estimate the seasonal pattern prevailing in the demand for his product? For example, do you believe that the businessman has an accurate enough knowledge of the seasonal pattern (without making any calculations) to set up a monthly budget that takes into account the seasonal fluctuations?

PROBLEMS

1. The following table gives the average cold storage holdings of frozen fruits and frozen vegetables for the five-year period 1950 to 1954.

(a) From these averages compute indexes of seasonal variation for cold storage holdings of frozen fruits and frozen vegetables.

(b) Plot both indexes on the same chart and describe briefly the two seasonal patterns.

(c) Compare the amplitude of the seasonal fluctuations in the cold storage holdings of frozen fruits and frozen vegetables by computing the average deviation from the mean for the two indexes of seasonal variation. Summarize briefly what the measures of dispersion show.

AVERAGE COLD STORAGE HOLDINGS OF FROZEN FRUITS AND FROZEN VEGETABLES ON FIRST OF MONTH: 1950 TO 1954

(MILLION POUNDS)

Date	Frozen fruits[1]	Frozen vegetables
Jan. 1	314	507
Feb. 1	289	457
Mar. 1	257	409
Apr. 1	220	369
May 1	182	336
June 1	182	320
July 1	224	338
Aug. 1	307	400
Sept. 1	351	489
Oct. 1	365	573
Nov. 1	375	607
Dec. 1	361	591

[1]Exclusive of juices and purees.
SOURCE: *Agricultural Statistics, 1956.*

2. From the data given in the table below:

(a) Compute the index of seasonal variation for newsprint consumption in the United States for the years 1947 to 1956 by averaging ratios to the twelve-month moving average centered at the seventh month.

(b) Plot the index on an arithmetic chart.

(c) How do you explain the seasonal pattern in the consumption of newsprint?

(d) *Retain this solution for use in Problem 4 of Chapter 15.*

NEWSPRINT CONSUMPTION BY PUBLISHERS IN THE UNITED STATES: 1947 TO AUGUST, 1957
(THOUSANDS OF SHORT TONS)

Month	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957
Jan.	266	293	333	345	346	345	352	363	384	402	408
Feb.	258	308	309	333	337	349	346	346	365	398	387
Mar.	303	338	367	397	394	399	421	400	439	446	463
Apr.	297	343	369	404	411	393	409	415	432	462	442
May	303	349	392	402	403	404	430	422	455	464	466
June	293	327	350	376	365	380	381	384	422	422	434
July	264	292	313	337	333	330	340	338	378	389	374
Aug.	281	314	318	347	344	342	359	361	385	403	386
Sept.	300	337	357	374	381	380	385	388	425	435	
Oct.	339	382	399	421	405	426	430	437	479	477	
Nov.	338	364	379	408	403	417	428	420	462	468	
Dec.	322	364	372	398	388	387	388	408	419	444	
Average	297	334	355	379	376	379	389	390	420	434	

SOURCE: *Business Statistics*, 1951, 1955, and 1957 (Supplements to the *Survey of Current Business*). *Survey of Current Business*, October, 1957.

3. Stocks of evaporated milk held by manufacturers show a pronounced seasonal variation, as measured by the following index of seasonal variation based on the seven years 1950 to 1956.

INDEX OF SEASONAL VARIATION
MANUFACTURERS' STOCKS OF EVAPORATED MILK: 1950 TO 1956

Month	Index	Month	Index	Month	Index	Month	Index
Jan.	58.9	Apr.	51.0	July	144.4	Oct.	138.7
Feb.	43.1	May	90.7	Aug.	153.3	Nov.	116.3
Mar.	39.5	June	130.0	Sept.	152.0	Dec.	82.1

SOURCE: *Business Statistics*, 1951, 1955, and 1957 (Supplements to the *Survey of Current Business*).

The production of milk fluctuates widely with the seasons, but the consumption of fresh milk is distributed throughout the year much more evenly. The surplus of milk during the months of large production is absorbed in the manufacture of products that can be stored, thus shifting the seasonal surplus into stocks. This tends to put wide seasonal swings into the production of the products made from surplus milk. Evaporated milk is one of the products that can be stored, and production tends to take on a seasonal pattern that is determined by the availability of surplus milk. The seasonal pattern of the production of evaporated milk is revealed in the data given in the following table for each month from 1950 through 1956.

(a) Construct an index of seasonal variation by the twelve-month moving average method.

(b) On the same chart plot the index of seasonal variation in production and the index of seasonal variation in stocks.

(c) Summarize the relationship between the two seasonal patterns.

(d) Compare the amplitude of the seasonal fluctuations in stocks and production of evaporated milk by computing the average deviation from the mean for the two indexes of seasonal variation. Summarize briefly what the measures of dispersion show.

(e) Assume that a forecast has been made that total production of evaporated milk for the coming year will be 2,700 million pounds. Forecast the monthly production of evaporated milk that you would consider reasonable on the basis of the seasonal pattern determined above.

(f) *Retain this solution for use in Problem 5 of Chapter 15.*

PRODUCTION OF EVAPORATED MILK: 1950 TO 1956
(MILLIONS OF POUNDS)

Month	1950	1951	1952	1953	1954	1955	1956
Jan.....	169.8	180.8	156.0	169.7	163.4	164.3	164.5
Feb.....	183.9	189.7	162.0	159.1	156.1	174.2	170.9
Mar.....	241.0	257.5	204.6	202.4	194.5	229.5	209.0
April....	258.0	288.8	261.6	243.4	243.8	256.1	240.1
May.....	347.0	390.4	370.3	321.1	314.6	325.1	300.7
June.....	348.8	372.9	349.5	324.7	308.0	303.4	304.1
July.....	302.1	314.7	271.4	262.8	265.8	255.7	266.7
August..	284.3	264.8	277.0	227.4	238.7	226.2	232.6
Sept.....	232.4	195.6	242.3	170.5	186.3	184.6	189.1
Oct.....	202.0	168.8	206.2	162.3	159.0	164.8	164.2
Nov.....	159.0	131.3	166.2	152.5	150.6	143.3	146.5
Dec.....	156.3	141.1	172.9	157.4	153.2	152.5	152.2
Average.	240.4	241.4	236.7	212.8	211.2	215.0	211.7

SOURCE: *Business Statistics*, 1951, 1955, and 1957 (Supplements to the *Survey of Current Business*).

15 CYCLICAL FLUCTUATIONS IN TIME SERIES

From the standpoint of an individual firm, the cyclical fluctuations of business represent a phenomenon entirely beyond the control of the manager. Practically no business concern can avoid being affected by the alternating periods of prosperity and depression; thus, information on the general state of business is vital in managing an individual business.

IMPORTANCE OF BUSINESS CYCLES IN BUSINESS MANAGEMENT

When business conditions are approaching a peak of prosperity, inventories should be held down, expansion of plant equipment should be little or none at all, extension of credit should be made very carefully, and, in general, the individual firm should be put into a position to lose as little money as possible when the inevitable decline in prices and volume occurs. When recovery is beginning to get under way, it is time to increase inventories, make needed expansion and modernization of equipment, increase advertising budgets and sales efforts, and, in general, get the firm in shape to take advantage of improved conditions.

When a business firm is caught with a large, high-priced inventory in a period of declining prices, there is usually little to do but take a loss. If plant and equipment were enlarged and improved at high prices just before a decline in volume, the concern may not be able to survive. Usually the new equipment can be justified on the basis of a continuation of business at a level approaching the high point of prosperity. If volume falls off, the money spent for expansion may be a complete loss. If the expansion was financed with borrowed funds, the interest charges may be the factor that puts the concern into bankruptcy.

At the bottom of a depression, business is usually conducted on a hand-to-mouth basis; no future commitments are made until it is certain that recovery is under way. If recovery comes suddenly after a long depression, a concern may find itself unable to step up production fast enough to take full advantage of the increased demand. For example, the increase in farm income in 1935 stimulated the demand for farm equipment far beyond the forecasts of the manufacturers. Dealers who could deliver equipment, particularly tractors, were in a position to make sales and take full advantage of the recovery. Companies that could not make deliveries simply lost the sales they might otherwise have made.

The cyclical fluctuation in business is an ever-present force that must be constantly watched. Some time in the future the business cycle may be brought under control, but at present about all businessmen can do is to manage their firms with as complete knowledge as possible of what is happening, trying in this way to avoid as many losses as possible and to take advantage of as many chances as possible to make a profit. For example, an executive cannot prevent a period of prosperity from ending in a sudden drop in prices and business activity; but if he can foresee the event, he can get his firm into the best possible condition to go through this period. At the bottom of the depression he can get his business concern into a condition to earn as large a profit as possible when conditions do improve. Proper management enabled a few concerns to show a profit even during the depression of the 1930's. Others came through in sound condition, though they may not have made money every year.

Even though it seems that the cyclical swings in business are at present inevitable, a great deal of study has been made in recent years of the possibility of controlling these fluctuations. Enough progress has been made in the use of some devices, notably credit control, to make it desirable that the business statistician consider the effects that the efforts to stabilize business are having on the current level of activity and may be expected to have on the forecast of the future.

It is the purpose of this chapter to describe the statistical methods that may be employed to measure the purely cyclical fluctuations in a time series. This is the first step in supplying information on such fluctuations. The method may be applied to the internal statistical data of an individual company, to the data for an industry, or to series showing changes in any of the elements of business conditions that interest the businessman.

If an executive wants to see how the cyclical fluctuations in the sales of his concern compare with the cyclical fluctuations in the industry, it is necessary to secure a measure that expresses the cyclical fluctuations in the two series so that they can be compared. The same principle applies when comparing an individual business firm or an industry with general business conditions.

Importance of keeping informed. Probably the most common method of keeping informed on the condition of general business is reading the financial pages of newspapers and trade papers, and talking to others. Much of the businessman's conversation consists of discussions of the question, "How's business?" The results of such informal methods are not likely to be highly accurate.

Since the most definite and accurate information on the condition of business is in quantitative form, knowledge of what information is available

and what it means is of utmost importance in answering the question, "How's business?" Using this kind of information is the best method now available for keeping informed.

Importance of business forecasting. The whole difficult problem of business forecasting starts with the measurement of the state of business activity. The term *business forecasting* may, strictly speaking, mean forecasts of the (1) long-term growth of a business, (2) seasonal fluctuations, or (3) cyclical fluctuations. In forecasting these, complete information on the conditions prevailing in all phases of business activity is important; the measurement of the conditions in the past and the present, which must precede a forecast of the future, is best accomplished by the use of statistical data in time series. Methods of forecasting trend and cycle will be described in Chapter 21.

Many businessmen believe that it will always be impossible to forecast the business cycle, and the severe depression of the 1930's seems to lend support to this belief. But regardless of the difficulty of making accurate forecasts, executives must make plans and commitments in advance. The management of modern business is in its very essence forecasting future conditions and operating the business in the light of these forecasts.

An analysis of typical decisions made by executives will reveal that their action is based to a considerable extent on forecasts of the future, either implied or expressed. Assume, for example, that a businessman decides to expand his factory in a period when business is good and orders are coming in faster than existing facilities can take care of them. He may not set down definitely his belief as to the continuation of orders in the future at the existing rate; but if he goes ahead and expands the plant capacity, he assumes that future volume will justify his course of action. Since the action is based on a forecast, whether stated or not, it is correct to say that this executive made a forecast and acted on the basis of it. Recognition of this fact tends to emphasize the importance of forecasts in making such decisions, which in turn should bring about more careful forecasting. It is conceivable that many courses of action would not be taken if it were entirely clear in the minds of the men making the decision just what the forecasts implied.

ISOLATING CYCLICAL FLUCTUATIONS

The Problem of Measuring Cyclical Fluctuations

No completely satisfactory method of measuring directly the cyclical swings in a time series has been developed. The irregular nature of the fluctuations makes impossible any attempt to find an average cycle that could be used to represent the effect of the cycle on the series. For this

reason the methods used to measure seasonal variation cannot be used for cyclical fluctuations. The length of the seasonal movement is fixed at twelve months, and so when the specific seasonals for a number of years show approximately the same amplitude of fluctuation, they can be averaged to secure a typical seasonal pattern. The cyclical fluctuations, on the other hand, show a wide variation in the length of the cycle as well as in the amplitude of the variations. Business volume expands and contracts in an oscillating movement that varies greatly in the time required to make a complete cycle. Professor W. C. Mitchell computed the average length of the business cycle to be four years, but this average was computed from cycles ranging in length from one year to nine years.

The best approach to the problem of measuring the cyclical fluctuations in a time series has been the indirect method of removing the variation in the series that results from either seasonal forces or secular trend. The remaining fluctuations are considered to be cyclical and erratic movements. Frequently no attempt is made to separate these two types of fluctuation, since no very satisfactory technique has been developed.

In the following sections the isolation of cyclical fluctuations is described for several types of data. Since seasonal variation and secular trend can be measured with reasonable accuracy, it is possible to calculate what the fluctuations in a series would have been if the trend and seasonal had not been present. This technique is based on the assumption that the trend, seasonal, and cyclical variations are independent of each other and therefore the cyclical fluctuations would have been the same if there had been no trend and seasonal variation in the series. It is believed that this assumption is usually approximately correct, but the extent to which it is incorrect is reflected in the resulting measure of cyclical fluctuations.

Adjusting Annual Data for Secular Trend

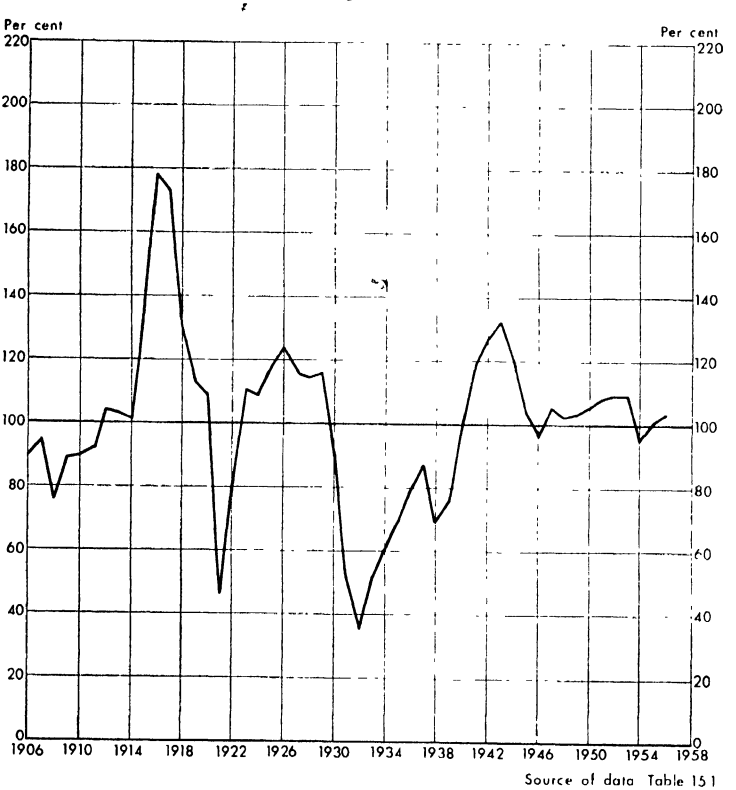
A time series consisting of period data on a yearly basis may show the effect of trend, cyclical, and erratic fluctuations but cannot show seasonal variations. Since each yearly amount includes all twelve months, any seasonal characteristics the data might possess are lost in the yearly total, and it is also usually true that a year is long enough for the erratic fluctuations to cancel out. Occasionally there will be nonrecurring events of sufficient magnitude for their effects to appear in annual data, but the minor factors that cause data classified on the basis of short periods to fluctuate erratically will ordinarily not influence annual data.

In an annual series that is affected only by trend and cyclical movements, the measurement of secular trend can be used to remove the effect of trend and leave only cyclical fluctuations. In Chart 12-1 on page 276 it can be seen that forces other than growth have influenced the production

of zinc in the different years. The fluctuations of the series about the line of trend have a wavelike swing that has given this movement the name *cyclical.* However, the trend has risen so rapidly that it obscures somewhat the cyclical fluctuations. These movements can be seen more clearly in Chart 15-1, in which the trend movement has been taken out, leaving only the cyclical fluctuations.

Chart 15-1

SMELTER PRODUCTION OF PRIMARY SLAB ZINC
IN THE UNITED STATES, 1906-1956



Source of data Table 15 1

The most commonly used method of separating the effect of the cycle from the trend is to express the data for each year as a percentage of the trend value for that year. These calculations appear in Table 15-1 for each year from 1906 to 1956, and the results are shown graphically in Chart 15-1. The logic underlying this method of separating the effects of trend and cycle is to assume that each movement is essentially independent of the other and that the forces making for the cyclical fluctuations have approximately

the same relative effect on the series regardless of the level of the trend. For example, the depression of 1921 pushed zinc production to a level 46 per cent of the trend, or 54 per cent below trend. In 1932 a still more severe depression resulted in a level of production 36 per cent of trend, or 64 per cent below trend. These relative deviations from trend represent chiefly the effect of the depression on production. If a comparison were made between the amounts by which the two years were below the trend value, the 1932 drop would be substantially larger than 1921 partly because the trend had increased during the eleven years — even the same percentage drop would represent a larger amount. This tendency for the cyclical fluctuations to maintain a somewhat consistent relative amplitude seems to justify using the relative fluctuations instead of the amounts of variation as a measure of the effect of the cycle on a series.

Table 15-1

SMELTER PRODUCTION OF PRIMARY SLAB ZINC IN THE UNITED STATES: 1906 TO 1956

ADJUSTMENT FOR SECULAR TREND

Year	Produc. (Th. tons)	Trend (Th. tons)	Percent. of trend	Year	Produc. (Th. tons)	Trend (Th. tons)	Percent. of trend
1906.....	225	250	90	1931.....	292	563	52
1907.....	250	262	95	1932.....	207	576	36
1908.....	210	275	76	1933.....	307	588	52
1909.....	256	287	89	1934.....	364	601	61
1910.....	269	300	90	1935.....	421	613	69
1911.....	287	312	92	1936.....	492	626	79
1912.....	339	325	104	1937.....	557	638	87
1913.....	347	338	103	1938.....	446	651	69
1914.....	353	350	101	1939.....	507	663	76
1915.....	490	363	135	1940.....	675	676	100
1916.....	668	375	178	1941.....	822	689	119
1917.....	670	388	173	1942.....	892	701	127
1918.....	518	400	130	1943.....	942	714	132
1919.....	466	413	113	1944.....	869	726	120
1920.....	463	425	109	1945.....	765	739	104
1921.....	201	438	46	1946.....	728	751	97
1922.....	354	450	79	1947.....	802	764	105
1923.....	510	463	110	1948.....	788	776	102
1924.....	517	475	109	1949.....	815	789	103
1925.....	573	488	117	1950.....	843	801	105
1926.....	618	501	123	1951.....	882	814	108
1927.....	593	513	116	1952.....	904	826	109
1928.....	603	526	115	1953.....	916	839	109
1929.....	625	538	116	1954.....	808	851	95
1930.....	498	551	90	1955.....	874	864	101
				1956.....	901	877	103

SOURCE: *Historical Statistics of the United States*, 1789–1945; *Business Statistics*, 1957.

Adjusting Monthly Data for Seasonal Variation

In a series with neither upward nor downward trend, the fluctuations will be the result of cyclical and erratic variations, when classified on a yearly basis. When classified into periods of less than a year, usually months or quarters, some of the variations will probably be due to seasonal as well as cyclical and erratic fluctuations. The measurement of the cyclical movements becomes a matter of first adjusting for seasonal variation and then, if possible, removing the erratic variations.

Seasonal swings commonly show a "wide" or "great" amplitude, with the result that the cyclical movements are badly obscured. For example, the demand for motor fuel is regularly lower in the winter than in the summer. When the increased demand in the spring brings about increased consumption, it might mean that business is getting better from the standpoint of the business cycle; but it might also mean that the increase is merely the usual seasonal rise. It might even be demonstrated that the rise is less than the normal seasonal increase in demand because the business cycle is in a period of downward movement.

Adjusting bank debits for seasonal variation. Bank debits represent a measure of business activity since they include a major portion of the payments made by businessmen and consumers. Since the portion of business transacted by the use of currency or coin is relatively small compared with the volume transacted by checks, the total debits made to individual accounts in banks represent the major portion of money spent during the period considered. Many types of payments, however, show a strong seasonal pattern, which means that bank debits reflect changes in the cyclical fluctuations of business and the seasonal variation.

For example, in early 1957 bank debits declined from $204,293 million in January to $178,049 million in February. This decline of 12.8 per cent in one month did not necessarily mean that business had suffered a severe cyclical setback, since the change could be seasonal. The index of seasonal variation for January was computed to be 103.3 and for February, 89.9. The normal decline in bank debits in February was, therefore, 13.4 points on the index of seasonal variation. A decline of 13.4 points is a decline of 13.0 per cent of the seasonal factor 103.3, the January value of the index of seasonal variation. This was very nearly the actual decline that occurred in 1957, so it would be reasonable to conclude that there was no significant change between January and February business volume due to the business cycle. In other words, the change was due almost entirely to seasonal factors operating on the two months, but this can be recognized only after the amount of seasonal variation present has been determined.

During the last half of 1949, business activity showed some cyclical decline, but in December, bank debits rose 18.0 per cent from November. Did this mean that the cyclical movement had turned up? According to the index of seasonal variation, December bank debits normally are 18.3 per cent higher than in November. If bank debits in 1949 increased only 18.0 per cent between November and December, *this was less than the normal seasonal increase and the cyclical movement was still downward.* In this case the seasonal movement completely obscured the cyclical movement, which was in the opposite direction.

These examples are typical of the difficulty of determining the cyclical movements unless some systematic method is used. The analyses in the two preceding paragraphs are accurate but somewhat cumbersome. When it is desired to separate the seasonal and the cyclical fluctuations over several months, a systematic method of adjusting for seasonal variation is used.

The index of seasonal variation for January bank debits was computed by the method of ratios to the twelve-month moving average and was found to be 103.3 per cent of the average month of the year. Bank debits in January, 1957, totaled $204,293 million and reflected the effect of the trend, the business cycle, and the seasonal pattern.

Because seasonal variation normally makes bank debits in January 3.3 per cent higher than the average month of the year, a portion of the volume of bank debits reflects the effect of seasonal forces. The index of seasonal variation measures the amount of influence exerted by the seasonal factors and can be used to measure what the January value would have been if no seasonal forces had been present. The bank debits for January, 1957, would have been somewhat smaller than $204,293 million if there had been no seasonal forces; since the seasonal factor has been measured to be 103.3, bank debits for January, 1957, were 103.3 per cent of what they would have been without the seasonal element present. In other words, the total bank debits of $204,293 million is 103.3 per cent and the problem is to compute the value of 100 per cent.

The computation to be made is as follows:

$$\begin{aligned} 103.3\% &= \$204{,}293 \text{ million} \\ 1\% &= \$204{,}293 \div 103.3 = \$1{,}977.7 \text{ million} \\ 100\% &= \$197{,}770 \text{ million} \end{aligned}$$

Bank debits of $204,293 million in January, 1957, are the equivalent of $197,770 million if no seasonal variation were present. In other words, bank debits in January have been *adjusted for the effect of seasonal variation.* February, 1957, bank debits adjusted by using the February index would be $198,050 million. The February adjusted bank debits increased only slightly, 0.1 per cent compared to a decrease of 12.8 per cent in the unadjusted bank debits.

The decrease of 12.8 per cent in the unadjusted bank debits represents the net effect of all of the factors on the series; the increase of 0.1 per cent represents the effect of all of the factors except seasonal variation. Secular trend, the business cycle, and erratic variations all register their influence on the series. Over as short a period as a month the trend usually does not have much effect. There appears to be a pronounced growth in bank debits, although the fact that prices have been rising is probably as important a factor as the growth in the volume of business transactions. Erratic factors are always present in monthly data and undoubtedly account for most of the sharp fluctuations that give a sawtooth appearance to so many graphs of time series.

An orderly method of adjusting a series for seasonal variation is shown in Table 15-2. The original data are entered in Column (2) and the index of seasonal variation, in Column (3). The index of seasonal variation may be expressed as a ratio to 1 instead of a percentage, since this eliminates the multiplication by 100 shown above, and the calculation for January, 1957, becomes simply

$$\$204{,}293 \div 1.033 = \$197{,}770$$

The adjusted figures in Column (4) represent the way bank debits would have been distributed over the months if all other forces affecting the data remained the same and there were no seasonal influences. The last digit to the right has been rounded; it is questionable whether even five digits should be used. The index of seasonal variation is only approximate, and there are no more significant figures in the adjusted series than in the index of seasonal variation.

In Table 15-2 the same index of seasonal variation is used to adjust the data for each year from January, 1947, to October, 1957. This is done on the assumption that the seasonal pattern has not shown any appreciable change over this period. The ten years 1947 to 1956 were used to compute the index of seasonal variation, but the index has been used to adjust the monthly data for 1957 also. The seasonal index can be used to adjust data beyond the period on which it is based on the assumption that it still measures the seasonal change even though the most recent years were not used in computing it. If this assumption cannot be made, it is not possible to adjust current data for seasonal variation.

If there is so much change in the seasonal pattern that the index cannot be projected for several years, it is preferable that a changing index be constructed according to the instructions given on pages 340 to 344. If a fixed index of seasonal variation is used beyond the period for which it was computed, it should be checked after the end of the year to see if

any substantial change in the index is required. Statistical organizations normally maintain a regular system for keeping the indexes of seasonal variation up to date, but it is generally not necessary to revise them every year.

The amounts in Column (4) are given in dollars, but they are merely abstract, computed quantities. The $197,770 million in January, 1957,

Table 15-2

INDEX OF BANK DEBITS IN 344 CITIES IN THE UNITED STATES: 1947 TO 1957

ADJUSTED FOR SEASONAL VARIATION

1947 TO 1949 = 100

Month and year (1)	Bank debits (Millions of dollars) (2)	Index of seasonal variation (3)	Bank debits adjusted for seasonal variation (Millions of dollars) (4)	Index adjusted for seasonal variation 1947 to 1949 = 100 (5)
1947				
Jan.	91,528	103.3	88,600	90
Feb.	80,224	89.9	89,240	91
March	92,216	105.4	87,490	89
April	85,770	98.4	87,160	89
May	86,493	97.7	88,530	90
June	92,381	103.9	88,910	90
July	91,656	98.2	93,340	95
Aug.	83,607	93.6	89,320	91
Sept.	90,873	96.9	93,780	95
Oct.	103,347	101.5	101,820	104
Nov.	91,486	96.8	94,510	96
Dec.	114,139	114.6	99,600	101
1948				
Jan.	102,929	103.3	99,640	101
Feb.	88,786	89.9	98,760	101
March	106,229	105.4	100,790	103
April	100,207	98.4	101,840	104
May	96,184	97.7	98,450	100
June	106,341	103.9	102,350	104
July	101,091	98.2	102,940	105
Aug.	96,634	93.6	103,240	105
Sept.	102,983	96.9	106,280	108
Oct.	105,064	101.5	103,510	105
Nov.	101,540	96.8	104,900	107
Dec.	119,488	114.6	104,270	106
1949				
Jan.	103,165	103.3	99,870	102
Feb.	88,534	89.9	98,480	100
March	107,808	105.4	102,280	104
April	97,288	98.4	98,880	101
May	97,588	97.7	99,890	102
June	106,178	103.9	102,190	104
July	96,049	98.2	97,810	100
Aug.	97,631	93.6	104,310	106
Sept.	99,412	96.9	102,590	104
Oct.	99,682	101.5	98,210	100
Nov.	97,707	96.8	100,940	103
Dec.	115,251	114.6	100,570	102

Table 15-2 (Continued)

Month and year (1)	Bank debits (Millions of dollars) (2)	Index of seasonal variation (3)	Bank debits adjusted for seasonal variation (Millions of dollars) (4)	Index adjusted for seasonal variation 1947 to 1949 = 100 (5)
1950				
Jan.	105,067	103.3	101,710	104
Feb.	94,377	89.9	104,980	107
March	113,300	105.4	107,500	109
April	100,325	98.4	101,960	104
May	110,342	97.7	112,940	115
June	117,445	103.9	113,040	115
July	108,706	98.2	110,700	113
Aug.	127,023	93.6	135,710	138
Sept.	121,469	96.9	125,360	128
Oct.	123,974	101.5	122,140	124
Nov.	121,368	96.8	125,380	128
Dec.	136,716	114.6	119,300	121
1951				
Jan.	136,271	103.3	131,920	134
Feb.	112,797	89.9	125,470	128
March	141,422	105.4	134,180	137
April	125,537	98.4	127,580	130
May	127,867	97.7	130,880	133
June	131,522	103.9	126,590	129
July	121,196	98.2	123,420	126
Aug.	122,421	93.6	130,790	133
Sept.	117,940	96.9	121,710	124
Oct.	135,644	101.5	133,640	136
Nov.	128,972	96.8	133,324	136
Dec.	140,965	114.6	123,010	125
1952				
Jan.	135,801	103.3	131,460	134
Feb.	125,159	89.9	139,220	142
March	136,275	105.4	129,290	132
April	134,110	98.4	136,290	139
May	133,000	97.7	136,130	139
June	139,740	103.9	134,490	137
July	137,313	98.2	139,830	142
Aug.	122,175	93.6	130,530	133
Sept.	136,048	96.9	140,400	143
Oct.	150,470	101.5	148,250	151
Nov.	127,647	96.8	131,870	134
Dec.	165,115	114.6	144,080	147

is not an actual amount of bank debits, but a computed amount used to represent what debits would be without the influence of seasonal factors. Since the series does not represent actual dollars, it is logical to express seasonally adjusted data as fixed-base relatives, as discussed in Chapter 6. These relatives are given in Column (5) of Table 15-2 and are plotted in Chart 15-2, using the monthly average for the 36 months — January, 1947, to December, 1949 — as 100. Until recently few seasonally adjusted time

Table 15-2 (Continued)

Month and year (1)	Bank debits (Millions of dollars) (2)	Index of seasonal variation (3)	Bank debits adjusted for seasonal variation (Millions of dollars) (4)	Index adjusted for seasonal variation 1947 to 1949 = 100 (5)
1953				
Jan.	145,919	103.3	141,260	144
Feb.	129,163	89.9	143,670	146
March	153,356	105.4	145,500	148
April	145,567	98.4	147,930	151
May	141,781	97.7	145,120	148
June	153,846	103.9	148,070	151
July	147,957	98.2	150,670	153
Aug.	134,386	93.6	143,570	146
Sept.	147,699	96.9	152,420	155
Oct.	149,606	101.5	147,400	150
Nov.	140,992	96.8	145,650	148
Dec.	168,596	114.6	147,120	150
1954				
Jan.	154,294	103.3	149,360	152
Feb.	141,926	89.9	157,870	161
March	171,260	105.4	162,490	165
April	154,661	98.4	157,180	160
May	149 722	97.7	153,250	156
June	163,407	103.9	157,270	160
July	154,859	98.2	157,730	160
Aug.	151,525	93.6	161,890	165
Sept.	149,907	96.9	154,700	157
Oct.	152,321	101.5	150,070	153
Nov.	156,843	96.8	162,030	165
Dec.	186,317	114.6	162,580	166
1955				
Jan.	163,378	103.3	158,160	161
Feb.	149,718	89.9	166,540	170
March	178,914	105.4	169,750	173
April	158,289	98.4	160,860	164
May	167,710	97.7	171,660	175
June	177,908	103.9	171,230	174
July	161,741	98.2	164,710	168
Aug.	167,358	93.6	178,800	182
Sept.	169,000	96.9	174,410	178
Oct.	175,807	101.5	173,210	176
Nov.	173,198	96.8	178,920	182
Dec.	200,523	114.6	174,980	178

series were published in concrete units, such as dollars or barrels, but were nearly always expressed as relatives. Practice has changed, however, and many examples of seasonally adjusted series in various units can be found in such sources as the *Survey of Current Business.*

Computing an index from relatives. Instead of adjusting the original data for bank debits, expressed in millions of dollars, it is possible to reduce the series to relatives and compute the index of seasonal variation from

Table 15-2 (Concluded)

Month and year (1)	Bank debits (Millions of dollars) (2)	Index of seasonal variation (3)	Bank debits adjusted for seasonal variation (Millions of dollars) (4)	Index adjusted for seasonal variation 1947 to 1949 = 100 (5)
1956				
Jan.	187,364	103.3	181,380	185
Feb.	162,107	89.9	180,320	184
March	189,793	105.4	180,070	183
April	176,760	98.4	179,630	183
May	185,584	97.7	189 950	193
June	186,540	103.9	179,540	183
July	181,284	98.2	184,610	188
Aug.	183,819	93.6	196,390	200
Sept.	167,154	96.9	172,500	176
Oct.	193,140	101.5	190,290	194
Nov.	185,207	96.8	191,330	195
Dec.	201,875	114.6	176,160	179
1957				
Jan.	204,293	103.3	197,770	201
Feb.	178,049	89.9	198,050	202
March	197,024	105.4	186,930	190
April	192,653	98.4	195,786	199
May	197,206	97.7	202,055	206
June	193,303	103.9	186,047	189
July	200,572	98.2	202,187	206
Aug.	190,539	93.6	203,567	207
Sept.	189,297	96.9	195,353	199
Oct.	...	101.5	...	...
Nov.	...	96.8	...	...
Dec.	...	114.6	...	...

SOURCE: *Survey of Current Business.*

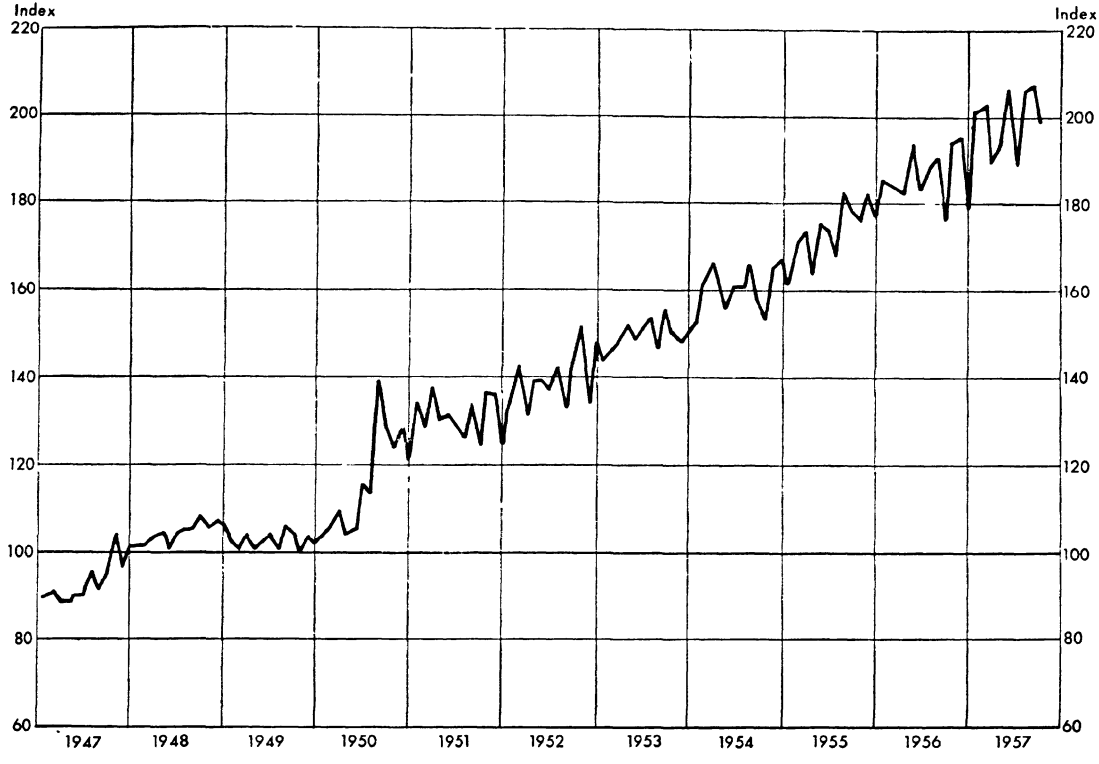
these relatives. The index of seasonal variation will be the same whether computed from the original data or from the relatives based on these data. It was demonstrated in Chapter 6 that expressing a series as relatives does not change the variations from period to period.

Table 15-3 gives the data for bank debits in millions of dollars and also as relatives (1947 to 1949 = 100) for the twelve months of 1956. The relatives are described as "Index, Unadjusted, 1947 to 1949 = 100." The index of seasonal variation is used to adjust this unadjusted index, which gives the same value for the index adjusted for seasonal variation as shown in Table 15-2. The advantage of constructing an index in the form of relatives is discussed in Chapter 6; and when there is any interest in the bank debits before the seasonal variation is eliminated, such an index is generally easier to interpret than a tabulation of the actual data on bank debits in dollars.

Chart 15-2

INDEX OF BANK DEBITS IN 344 CITIES IN THE UNITED STATES, 1947-1957

Adjusted for Seasonal Variation 1947-1949 = 100



Source of data, Table 15-2

Table 15-3

INDEX OF BANK DEBITS IN 344 CITIES IN THE UNITED STATES: 1956

ADJUSTED FOR SEASONAL VARIATION
1947 TO 1949 = 100

Month	Bank debits (Millions of dollars)	Index (Unadjusted) 1947 to 1949 = 100	Index of seasonal variation	Index (Adjusted for seasonal variation) 1947 to 1949 = 100
January.....	187,364	190.7	103.3	185
February.....	162,107	165.6	89.9	184
March.....	189,793	193.1	105.4	183
April.....	176,760	179.9	98.4	183
May.....	185,584	188.9	97.7	193
June.....	186,540	189.8	103.9	183
July.....	181,284	184.5	98.2	188
August.....	183,819	187.1	93.6	200
September.....	167,154	170.1	96.9	176
October.....	193,140	196 6	101.5	194
November.....	185,207	188.5	96.8	195
December.....	201,875	205.4	114.6	179

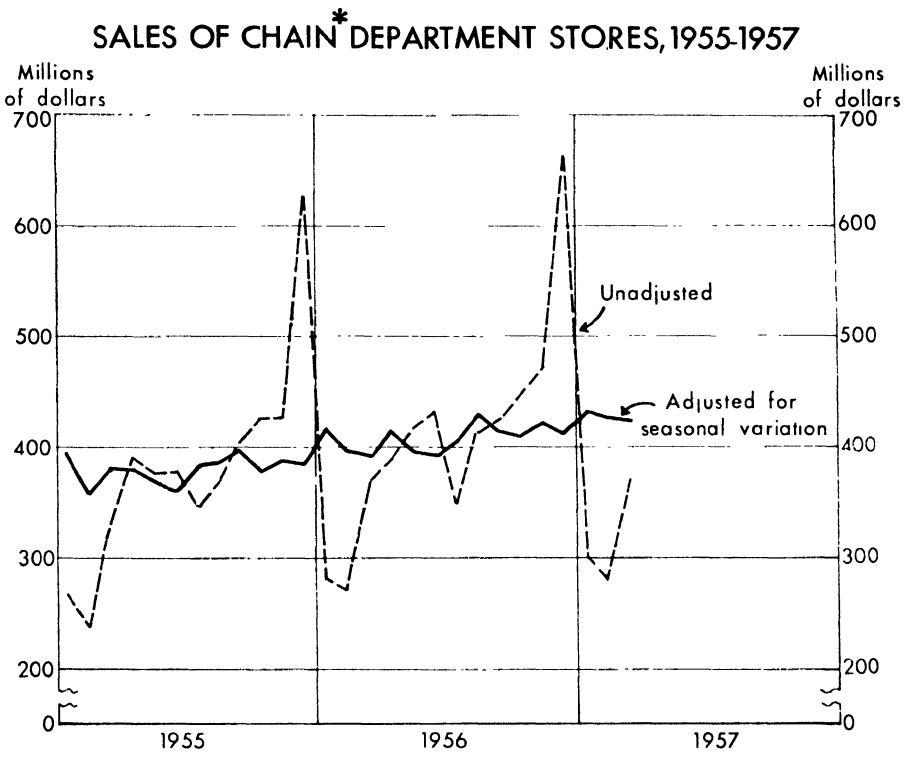
SOURCE: *Business Statistics*, 1957.

Using a changing index. A changing index of seasonal variation, which was described on page 340, may be used to adjust the series instead of the average index; in Table 15-4 the use of the changing index is illustrated. The arithmetic is the same in either case; that is, the original data are divided by the index of seasonal variation to remove the effect of seasonal variation.

Showing the seasonal pattern. The seasonal fluctuations in time series represent such a pronounced variation that most important business and economic time series are published in seasonally adjusted form. The removal of the effects of seasonal variation has become almost a routine operation on important time series, since it permits attention to be centered on the other kinds of variation.

Both the seasonally adjusted data and the actual unadjusted data may be plotted on the same chart, although in practice there is not often any reason for doing so. The difference between the curve representing the adjusted data and the unadjusted data shows the effect of the seasonal variation. A more direct method of showing the seasonal pattern is to plot the index of seasonal variation as was done in Chart 14-3 (page 342).

Chart 15-3



Source of data *Survey of Current Business, March, 1957,* and *May, 1957*

Chart 15-3 gives an adjusted and unadjusted series for the years 1955 to 1957 to show graphically the change made in a series by removing the effects of the seasonal fluctuations. The sales of chain department stores show wide fluctuations with the seasons, but after allowance for these variations has been made, the change is not nearly so great. Some erratic fluctuations are apparent, and a gradual rise suggests a cyclical expansion of sales.

Adjusting Monthly Data for Seasonal Variation and Trend

Monthly data can be adjusted for secular trend in the same manner as annual data were adjusted in Table 15-1, except that a trend value must be secured for each month and compared with the monthly data. If the series has no seasonal variation, the data adjusted for trend would then contain cyclical and erratic movements. But if there is a seasonal movement in the series, adjustment must be made for seasonal variation as well as trend in order to isolate the cyclical and erratic fluctuations.

Most monthly data do show a seasonal movement which calls for adjustment. In Table 15-4 the domestic demand for motor fuel is adjusted first for seasonal variation, and then these figures are further adjusted for trend. The percentages of trend in Column (6) are then smoothed by a three-month moving average to remove erratic fluctuations. These computations combine the procedures used in Tables 15-1 and 15-2. The index

Table 15-4

DEMAND FOR MOTOR FUEL IN THE UNITED STATES: 1946 TO 1957

ADJUSTED FOR SEASONAL VARIATION AND SECULAR TREND

Month and year (1)	Demand (Million barrels) (2)	Index of seasonal variation (3)	Demand adjusted for seasonal variation (Million barrels) (4)	Trend (Million barrels) (5)	Demand adjusted for trend and seasonal variation (6)	Three-month moving average (7)
1946						
Jan.....	51.7	88	58.8	59.0	100	...
Feb.....	47.7	78	61.2	59.5	103	101.0
March...	56.7	95	59.7	59.9	100	102.3
April....	62.1	99	62.7	60.4	104	101.7
May.....	66.8	109	61.3	60.8	101	100.7
June.....	63.3	107	59.2	61.3	97	99.3
July.....	69.1	112	61.7	61.7	100	98.0
Aug.....	66.7	110	60.6	62.2	97	97.7
Sept.....	62.3	104	59.9	62.6	96	97.3
Oct.....	66.6	107	62.2	63.1	99	98.7
Nov.....	61.3	96	63.9	63.5	101	99.3
Dec.....	61.1	97	63.0	64.0	98	100.0
1947						
Jan.....	57.1	88	64.9	64.4	101	99.3
Feb.....	50.6	79	64.1	64.9	99	99.0
March...	60.0	95	63.2	65.3	97	97.7
April....	63.3	99	63.9	65.8	97	97.7
May.....	70.9	108	65.6	66.2	99	98.3
June.....	71.2	108	65.9	66.7	99	98.3
July.....	73.4	112	65.5	67.2	97	97.7
Aug.....	72.1	110	65.5	67.6	97	98.3
Sept.....	71.4	104	68.7	68.1	101	99.3
Oct.....	73.3	107	68.5	68.5	100	99.3
Nov.....	64.1	96	66.8	69.0	97	99.0
Dec.....	67.5	97	69.6	69.4	100	99.0
1948						
Jan.....	61.3	88	69.7	69.9	100	100.0
Feb.....	56.5	80	70.6	70.3	100	100.3
March...	68.2	95	71.8	70.8	101	101.0
April....	72.2	99	72.9	71.2	102	101.0
May.....	77.2	108	71.5	71.7	100	100.7
June.....	78.0	108	72.2	72.1	100	100.3
July.....	81.5	111	73.4	72.6	101	100.7
Aug.....	80.3	109	73.7	73.0	101	101.0
Sept.....	76.1	103	73.9	73.5	101	99.3
Oct.....	75.2	106	70.9	73.9	96	99.3
Nov.....	72.5	96	75.5	74.4	101	98.7
Dec.....	72.1	97	74.3	74.8	99	98.3

Table 15-4 (Continued)

Month and year (1)	Demand (Million barrels) (2)	Index of seasonal variation (3)	Demand adjusted for seasonal variation (Million barrels) (4)	Trend (Million barrels) (5)	Demand adjusted for trend and seasonal variation (6)	Three-month moving average (7)
1949						
Jan.....	63.1	88	71.7	75.3	95	96.7
Feb.....	58.0	80	72.5	75.7	96	97.0
March...	73.3	96	76.4	76.2	100	98.3
April....	75.3	99	76.1	76.6	99	99.0
May.....	81.7	108	75.6	77.1	98	99.0
June.....	83.4	108	77.2	77.5	100	98.0
July.....	82.1	110	74.6	78.0	96	98.3
Aug.....	84.7	109	77.7	78.5	99	98.0
Sept.....	80.8	103	78.4	78.9	99	97.3
Oct.....	79.3	106	74.8	79.4	94	97.7
Nov.....	76.3	96	79.5	79.8	100	97.0
Dec.....	75.6	97	77.9	80.3	97	97.0
1950						
Jan.....	66.9	88	76.0	80.7	94	95.7
Feb.....	63.4	81	78.3	81.2	96	96.7
March...	78.7	96	82.0	81.6	100	98.0
April....	80.3	100	80.3	82.1	98	99.7
May.....	89.0	107	83.2	82.5	101	100.0
June.....	90.2	108	83.5	83.0	101	100.7
July.....	91.7	110	83.4	83.4	100	101.3
Aug.....	94.5	109	86.7	83.9	103	101.0
Sept.....	86.8	103	84.3	84.3	100	101.0
Oct.....	89.1	105	84.9	84.8	100	100.3
Nov.....	82.7	96	86.1	85.2	101	99.7
Dec.....	81.1	97	83.6	85.7	98	101.3
1951						
Jan.....	80.7	89	90.7	86.1	105	101.7
Feb.....	72.6	82	88.5	86.6	102	103.3
March...	86.7	97	89.4	87.0	103	101.7
April....	87.3	100	87.3	87.5	100	103.0
May.....	99.4	107	92.9	87.9	106	102.0
June.....	96.3	109	88.3	88.4	100	103.3
July.....	100.5	109	92.2	88.8	104	102.7
Aug.....	101.1	109	92.8	89.3	104	102.3
Sept.....	91.3	103	88.6	89.8	99	103.3
Oct.....	100.5	104	96.6	90.2	107	102.3
Nov.....	88.0	96	91.7	90.7	101	101.3
Dec.....	85.2	97	87.8	91.1	96	101.3
1952						
Jan.....	87.1	89	97.9	91.6	107	103.7
Feb.....	82.3	83	99.2	92.0	108	104.0
March...	87.2	97	89.9	92.5	97	103.7
April....	98.8	100	98.8	92.9	106	101.3
May.....	100.5	107	93.9	93.4	101	101 3
June.....	99.3	109	91.1	93.8	97	100.3
July.....	105.5	109	96.8	94.3	103	100.0
Aug.....	103.0	109	94.5	94.7	100	102.0
Sept.....	100.2	102	98.2	95.2	103	102.3
Oct.....	103.9	104	99.9	95.6	104	102.0
Nov.....	91.6	96	95.4	96.1	99	102.0
Dec.....	96.8	97	99.8	96.5	103	101.3

Table 15-4 (Continued)

Month and year (1)	Demand (Million barrels) (2)	Index of seasonal variation (3)	Demand adjusted for seasonal variation (Million barrels) (4)	Trend (Million barrels) (5)	Demand adjusted for trend and seasonal variation (6)	Three-month moving average (7)
1953						
Jan.....	88.1	89	99.0	97.0	102	102.7
Feb.....	84.6	84	100.7	97.4	103	102.3
March...	96.7	97	99.7	97.9	102	102.3
April....	100.1	100	100.1	98.3	102	101.0
May.....	104.2	106	98.3	98.8	99	101.7
June.....	112.6	109	103.3	99.2	104	102.0
July.....	111.0	108	102.8	99.7	103	102.0
Aug.....	106.8	108	98.9	100.1	99	101.0
Sept.....	103.6	102	101.6	100.6	101	99.7
Oct.....	103.4	103	100.4	101.1	99	99.7
Nov.....	96.9	96	100.9	101.5	99	99.0
Dec.....	97.7	97	100.7	102.0	99	99.0
1954						
Jan.....	89.9	89	101.0	102.4	99	99.3
Feb.....	86.2	84	102.6	102.9	100	99.7
March...	101.5	98	103.6	103.3	100	99.7
April....	103.9	101	102.9	103.8	99	98.0
May.....	104.4	106	98.5	104.2	95	97.7
June.....	113.0	109	103.7	104.7	99	98.0
July.....	112.2	107	104.9	105.1	100	98.7
Aug.....	110.2	108	102.0	105.6	97	98.0
Sept.....	104.7	102	102.6	106.0	97	96.7
Oct.....	105.6	103	102.5	106.5	96	97.7
Nov.....	102.4	96	106.7	106.9	100	98.7
Dec.....	104.3	97	107.5	107.4	100	99.7
1955						
Jan.....	96.4	90	107.1	107.8	99	98.3
Feb.....	88.5	85	104.1	108.3	96	98.0
March...	105.7	98	107.8	108.7	99	98.7
April....	111.1	101	110.0	109.2	101	100.3
May.....	115.7	105	110.2	109.6	101	100.7
June.....	120.7	110	109.7	110.1	100	99.7
July.....	115.7	107	108.1	110.5	98	100.0
Aug.....	121.8	108	112.8	111.0	102	100.3
Sept.....	113.4	101	112.3	111.4	101	100.7
Oct.....	112.6	102	110.4	111.9	99	100.3
Nov.....	109.2	96	113.7	112.4	101	100.3
Dec.....	111.0	97	114.4	112.8	101	100.3
1956						
Jan.....	100.6	90	111.8	113.3	99	100.0
Feb.....	98.1	86	114.1	113.7	100	100.0
March...	113.1	98	115.4	114.2	101	99.7
April....	113.0	101	111.9	114.6	98	100.7
May.....	124.1	105	118.2	115.1	103	100.3
June.....	127.4	110	115.8	115.5	100	100.7
July.....	121.2	106	114.3	116.0	99	99.7
Aug.....	126.2	108	116.9	116.4	100	98.0
Sept.....	112.7	101	111.6	116.9	95	98.3
Oct.....	120.1	102	117.7	117.3	100	98.3
Nov.....	112.6	96	117.3	117.8	100	97.7
Dec.....	108.2	98	110.4	118.2	93	98.3

Table 15-4 (Concluded)

Month and year (1)	Demand (Million barrels) (2)	Index of seasonal variation (3)	Demand adjusted for seasonal variation (Million barrels) (4)	Trend (Million barrels) (5)	Demand adjusted for trend and seasonal variation (6)	Three-month moving average (7)
1957						
Jan.....	109.2	90	121.3	118.7	102	96.3
Feb.....	96.7	87	111.5	119.1	94	97.3
March...	113.2	99	114.3	119.6	96	95.3
April....	115.8	101	114.7	120.0	96	97.3
May.....	124.3	105	118.4	120.5	98	95.0
June.....	121.6	110	110.5	120.9	91	97.0
July.....	130.5	105	124.3	121.4	102	...
Aug.....	...	108	...	121.8	...	...
Sept.....	...	101	...	122.3	...	...
Oct.....	...	101	...	122.7	...	...
Nov.....	...	96	...	123.2	...	...
Dec.....	...	98	...	123.7	...	...

SOURCE: *Survey of Current Business.*

of seasonal variation was computed by the method of ratios to the twelve-month moving average in Table 14-17 (page 344) using a changing seasonal index. The secular trend was computed for the years 1946 to 1956 in Table 12-3 (page 288).

Chart 15-4 shows the unadjusted demand and the trend line for the years 1946 to 1957. Chart 15-5 shows the demand adjusted for trend and seasonal variation, expressed as a percentage of the secular trend, and smoothed with a three-month moving average, which removes most of the erratic fluctuations. The fluctuations remaining after the series has been adjusted for trend and seasonal variation and smoothed by a three-month moving average can be considered the cyclical movements of the series.

Monthly Trend Values

The computation of the monthly trend values in Column (5) of Table 15-4 was not explained in Chapters 12 and 13 ("Secular Trend"), since normally the only occasion for deriving a trend value for each month is the desire to adjust the series for trend to isolate the cyclical fluctuations. The computation of monthly trend values is described in this section for the most important types of trends.

When the trend line has been located by the graphic method, the monthly trend values may be secured in the same manner as the annual trend values. The monthly figures to which the trend is to be fitted should

DEMAND FOR MOTOR FUEL IN THE UNITED STATES, 1946-1957

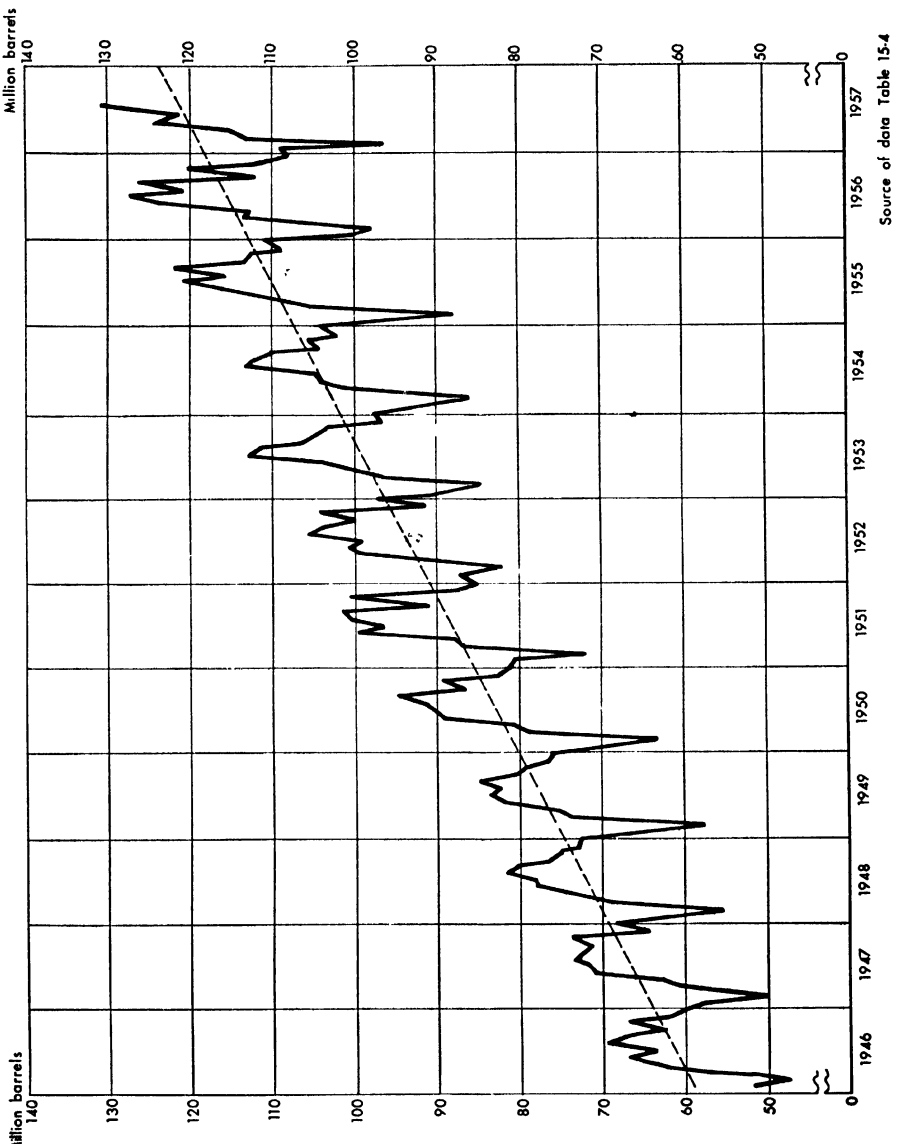
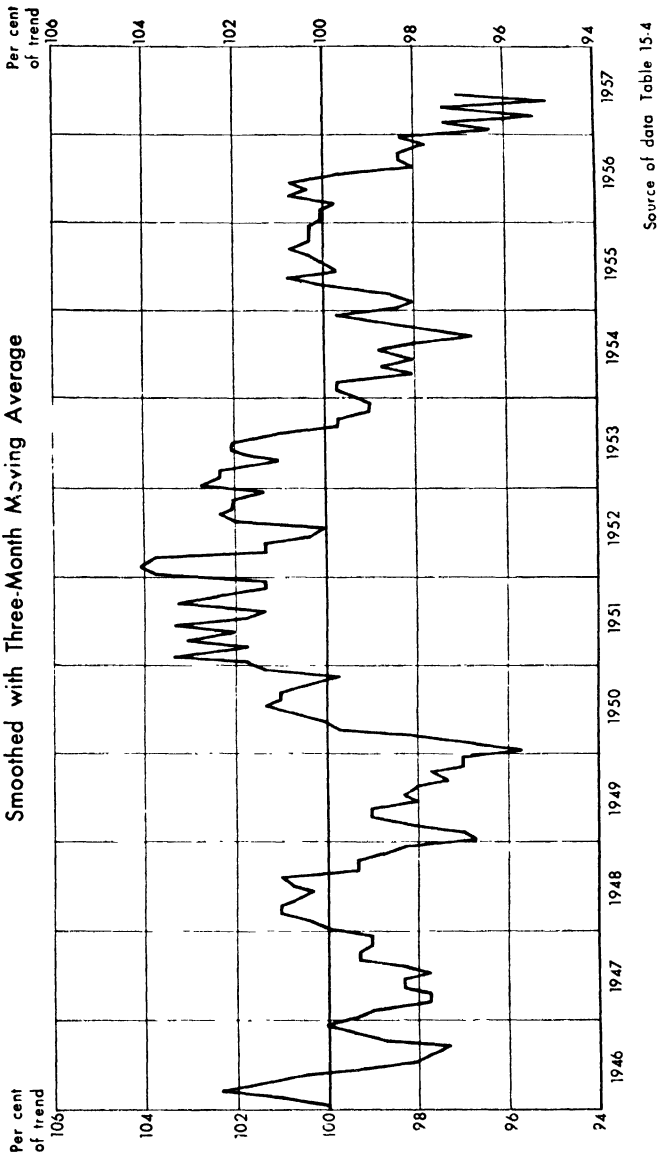


Chart 15-5

DEMAND FOR MOTOR FUEL IN THE UNITED STATES, 1946-1957

Adjusted for Trend and Seasonal Variation
Smoothed with Three-Month Moving Average



first be plotted, and the trend located graphically. The monthly values may then be read from the chart in the same manner as the annual values. For other methods of computation, annual trend values are ordinarily computed first and monthly values derived from the annual values. For a straight line fitted to the original data or to the logarithms of the data, the following steps are necessary to derive monthly trend values from annual.

1. The first step is to determine that the trend was fitted to average monthly data or, if fitted to total annual data, to divide the annual trend values by twelve to secure the average monthly trend values.
2. The second step is to determine the increase or decrease from one year to the next. One twelfth of this amount will be the increase or decrease in the trend *from one month to the next.*
3. The average monthly trend for a year is the level of the trend at July 1 of that year. The trend for the month of July, however, will be the level of the trend at July 15[1]. Since the rise or fall of the trend in a month was found in step 2 above, the change occurring in one-half a month can be computed by merely dividing by two. This amount added to the average monthly trend for the year will give the trend value at July 15.
4. After the trend for one month has been found, the trend for all other months can be computed by adding or subtracting the monthly change, determined in step 2. This statement applies only to the straight line, fitted either to the data or to the logarithms. For any trend other than a straight line, the monthly increase or decrease will *not* be the same for all months. This is discussed further below.
5. The accuracy of the calculations can be checked by computing the average of the twelve monthly trend values for each year. These averages should agree exactly with the average monthly trend values previously computed.

In computing the monthly trend values for the domestic demand for motor fuel, for instance, step 1 consisted merely of verifying that the trend values in Table 12-3 (page 288) were computed from a series of average monthly data and therefore represented average monthly trend values.

Step 2 was to find the amount by which the trend increased in one year. Since the straight line was fitted by the method of least squares, the b constant of the trend equation represents the yearly increase. On page 289, b was found to equal 5.425, and the monthly increase in the trend was therefore one twelfth this amount, or .452.

In step 3 the trend for the year 1946 was found to be 61.5, and since the value for a year is the value at midyear, this amount is the trend on July 1,

[1]This is an application of the principle stated on page 142 that period data are always considered as falling on the center of the period. If the trend of point data is being used, the same general principles of computing the monthly trend would be used; but each case would be a special problem depending on the date within the month for which trend values were wanted.

1946. In step 2 it was determined that the trend rose .452 per month; so between July 1 and July 15 it would rise one half this amount, or .226. The trend for July 15, 1946, is therefore 61.7, a figure reached by adding .226 to the trend for July 1, which was 61.5.

Step 4 consists of adding the monthly increase of .452 to each monthly trend value to secure the trend value for the next month.

In adjusting current data for secular trend it is necessary to project the trend into the current year without using that year for its computation. After complete data are available for a year, it is wise to check to see if these data make it necessary to revise the trend, although generally it is safe to project the monthly trend values for several years before recomputing the trend. The trend values in Table 15-4 were based on the years 1946 to 1956, but have been projected through 1957.

When computing the monthly trend values for the straight line fitted to the logarithms of the actual data, the steps performed above are carried out with the a and b values to compute the logarithms of the monthly trend values. Then the additional step of finding the antilogs must be carried out.

When computing the monthly trend values for the second-degree parabola, it is satisfactory to interpolate on a straight line between the annual trend values of the parabola.

FORECASTING CYCLICAL FLUCTUATIONS IN TIME SERIES

Forecasting the cyclical fluctuations in industry and trade is a major problem of management, for many decisions made in individual business concerns depend upon the course of activity in a particular industry or in business as a whole. This chapter has described in some detail the methods of measuring these cyclical fluctuations. Probably the chief reason for wanting to measure them is for the purpose of forecasting them. However, the contributions of statistical methods to the subject of forecasting have been confined chiefly to furnishing a more precise measurement of these changes. Purely statistical methods of forecasting cyclical fluctuations have so far not been discovered.

Nevertheless business plans must be made for the future, and forecasting is inevitable even though it is inaccurate. The only basis for the future is the history of what has happened in the past; so the more accurately we can measure the past, the better equipped we are to forecast. The causes of fluctuations are so complex that probably no simple method of forecasting will ever be found. Actually only the beginnings have been made at the

present time. But the information made available on the fluctuations in individual time series represents part of the factual basis for a forecast of changes in general business activity.

In the following chapter, "Construction of Index Numbers," the methods of measuring composite changes in a number of related series will be described. This technique makes possible the measurement of changes in industries and in total business activity, which furnishes additional information that is helpful in arriving at a forecast.

Following the discussion of correlation in Chapters 17 to 19, the amount of information available to the business forecaster is outlined in Chapter 20, "Business Barometers." Until that point it is not worth while attempting to discuss the subject of forecasting cyclical fluctuations because there are no methods by which the cyclical fluctuations in a given series of data can be predicted solely on the basis of the past performance of the series.

STUDY QUESTIONS

1. The fact that seasonal fluctuations occur over a fixed period of a year makes their measurement somewhat less difficult than the measurement of the cyclical fluctuations in business. Do you agree with this statement? Give reasons for your answer.

2. A very large percentage of the published measures of important elements of business activity that are classified on the basis of periods less than a year are adjusted for seasonal variation. What types of fluctuations may still be present in these adjusted series?

3. A very small proportion of published time series are adjusted for secular trend. Why is the adjustment for seasonal variation made so much more frequently than for secular trend?

4. Determining the cyclical fluctuations of a time series by removing the effects of the types of fluctuations that show a regular pattern is not considered ideal by those who use it. Since the method is not held in high regard, how do you explain its extensive use? What are the chief objections to this method of measuring the cyclical fluctuations in business?

5. Under what conditions would you recommend adjusting for secular trend by computing the *amount* of deviation from the trend rather than the *relative* deviation?

6. Are there any situations likely to arise where it would be better to adjust for seasonal variation by subtracting the amount of the seasonal influence rather than dividing by the seasonal adjustment factor?

7. Why would you expect that the adjustment of a series by means of a changing index of seasonal variation would give a more precise measure of the cyclical fluctuations than by means of a fixed index of seasonal variation?

8. Business activity during the summer months is below the average monthly volume in enough lines of business to cause the index of seasonal variation for total business activity to decline during the summer and to rise sharply in the fall. When business is in a period of cyclical decline, it is very common for reports on the business situation during the summer to predict that business will improve "as soon as the fall upturn" is felt.

(a) Does this "fall upturn" refer to a cyclical or seasonal movement?
(b) Is a cyclical upturn in business more likely to occur at the same time as a seasonal rise, or is the cyclical pattern independent of the seasonal pattern?
(c) Is it possible that business analysts are inclined to confuse the seasonal and the cyclical fluctuations in business? Discuss.

9. Assume that the values of the index of seasonal variation for sales of a certain company are 85 for August and 110 for September. Sales in August were $245,650 and in September, $316,800.

(a) How did the cyclical position of the business in September compare with August?
(b) If the manager of the business estimated correctly in August that the sales in September would be approximately $316,800, would he have been justified in predicting that an improvement in September business was expected? (Should he have explained further what he meant by "improvement"?)

10. A certain retailer makes, on the average, 12.5 per cent of his annual sales in December and only 6 per cent of his annual sales in January. If sales in January were $649,440 compared with $1,230,000 in the preceding December, does he have reason to worry about the underlying trend of business? Or, can he consider his business in January to be reasonably good in comparison with the volume in December?

11. Assume that during a period of twelve months the statistical measures of business activity declined 25 per cent but that the measures were still above the peak reached during the most prosperous period 15 years earlier. Does this mean that business is not in reality depressed, even though a decline of 25 per cent has been registered?

PROBLEMS

1. Table 13-2 (page 300) gives the annual consumption of anthracite coal in the United States from 1900 to 1956 and the trend values for the same years. Adjust the annual consumption of anthracite coal for this period and plot the adjusted series on an arithmetic chart.

2. Refer to the problem on page 294 of Chapter 12 in which are given nine time series to which a straight-line trend may be fitted. Adjust each of these series for secular trend and plot the adjusted data on an arithmetic chart.

3. The following table shows the consumption of newsprint in the United States from 1929 to 1956. Adjust the series for secular trend and plot the adjusted series on an arithmetic chart. (In order to adjust the series for secular

trend, it is necessary to compute the trend values. Plot the consumption of newsprint on an arithmetic chart and also on a semilogarithmic chart before deciding which trend to use.)

AVERAGE MONTHLY NEWSPRINT CONSUMPTION IN THE UNITED STATES: 1929 TO 1956

(THOUSANDS OF SHORT TONS)

Year	Thousand tons	Year	Thousand tons
1929.....	245	1943.....	227
1930.....	235	1944.....	196
		1945.....	205
1931.....	218		
1932.....	188	1946.....	261
1933.....	179	1947.....	297
1934.....	206	1948.....	334
1935.....	222	1949.....	355
		1950.....	379
1936.....	245		
1937.....	246	1951.....	376
1938.....	221	1952.....	379
1939.....	228	1953.....	389
1940.....	238	1954.....	390
		1955.....	420
1941.....	246		
1942.....	236	1956.....	434

SOURCE: *Business Statistics*, 1951, 1955, and 1957 (Supplements to the *Survey of Current Business*).

4. (a) From the trend line fitted in Problem 3 above, derive the monthly trend values for each month from January, 1947, to August, 1957.

(b) Use the monthly trend values and the index of seasonal variation computed in Problem 2, Chapter 14, to adjust the consumption of newsprint for trend and seasonal variation from January, 1947, to August, 1957.

(c) Plot the adjusted series on an arithmetic chart.

(d) How well does the chart measure the cyclical fluctuations in newsprint consumption?

5. The following table gives the production of evaporated milk by months and the stocks of evaporated milk held by manufacturers at the end of each month from January, 1954, to August, 1957.

(a) Adjust stocks and production for seasonal variation and plot the unadjusted and adjusted series on the same arithmetic charts. (Construct one chart for stocks and another for production.)

(b) Use the indexes of seasonal variation given in Problem 3, page 355, of Chapter 14.

(c) After making allowance for seasonal variation, summarize briefly the changes that have taken place in the stocks and production of evaporated milk.

EVAPORATED MILK

MONTHLY PRODUCTION AND STOCKS AT END OF MONTH: 1954 TO 1957

(Millions of pounds)

Year and month	Production	Stocks	Year and month	Production	Stocks
1954			**1956**		
Jan.	163.4	313.8	Jan.	164.5	156.8
Feb.	156.1	262.9	Feb.	170.9	110.6
Mar.	194.5	237.0	Mar.	209.0	111.4
Apr.	243.8	262.6	Apr.	240.1	124.9
May	314.6	365.2	May	300.7	168.5
June	308.0	475.3	June	304.1	312.0
July	265.8	511.7	July	266.7	401.9
Aug.	238.7	524.0	Aug.	232.6	434.5
Sept.	186.3	481.2	Sept.	189.1	425.5
Oct.	159.0	410.3	Oct.	164.2	383.7
Nov.	150.6	339.8	Nov.	146.5	310.4
Dec.	153 2	262.7	Dec.	152.2	224.0
1955			**1957**		
Jan.	164.3	192.8	Jan.	162.0	156.4
Feb.	174.2	127.7	Feb.	169.9	123.6
Mar.	229.5	102.6	Mar.	212.6	83.4
Apr.	256.1	127.7	Apr.	260.5	154.0
May	325.1	231.5	May	312.2	244.5
June	303.4	320.5	June	278.4	368.9
July	255.7	381.1	July	245.7	421.7
Aug.	226.2	410.4	Aug.	217.3	438.8
Sept.	184.6	410.2			
Oct.	164.8	355.5			
Nov.	143.3	290.6			
Dec.	152.5	207.0			

SOURCE: *Business Statistics*, 1957; *Survey of Current Business, October, 1957.*

16 CONSTRUCTION OF INDEX NUMBERS

The methods of analysis of time series described in preceding chapters have been applied to individual series, with the objective of isolating specific types of fluctuations, either secular trend, seasonal variation, or cyclical movements. The methods of index-number construction which will be discussed in this chapter deal with the techniques of combining time series that cannot be added because they are not in comparable units.

THE NATURE OF AN INDEX NUMBER

The annual production of wheat can be measured by simply totaling the output of the individual producers or the total amount moving through the markets. Even though there may be different grades of the product, as is generally true for commodities, all bushels of wheat are nearly enough alike to make the total production a significant amount. In general, items of the same kind, when classified according to differences of kind as discussed on pages 19 and 20, may be added if all measurements are expressed in the same units.

Many situations, however, arise when it is desired to measure the composite changes in the production of a number of commodities that are not expressed in the same units and that cannot, therefore, be added in order to secure a total. Changes in the production of manufactured goods cannot be measured by totaling the production of all kinds of goods, since there is no common physical unit of measurement that can be used for all. When the production of the different items cannot be added, it would also be improper to average the prices of the items. Accordingly, when it is desired to measure the change in the average level of prices, it is necessary to use the methods of index-number construction.

It should be emphasized that as a general rule it is preferable to reduce the different items to comparable units and then add them. This device should be used whenever possible. For example, in measuring the total production of fuel, it would not be satisfactory to add coal, petroleum, and natural gas output; but it would be possible to convert the production of each fuel into units of heat, which could be added. In this case the average price for each B.T.U. (British thermal unit) could also be computed and

this, over a period of time, would measure the change in fuel prices. Even this computation, however, can be criticized on the ground that the efficiency with which the different fuels are used varies. As a result, it would be necessary to reduce the various fuels to B.T.U. content after taking into consideration the effective energy that could be obtained from each fuel, using the equipment that is currently available. Although on a few occasions it might be possible to reduce the different items to comparable units and then add them, such a device has only limited applicability. When it cannot be used, the methods of index-number construction should be utilized.

The basic device used in all methods of index-number construction is to average the *relative change* in either quantities or prices, since relatives are comparable and can be added even though the data from which they were derived cannot themselves be added. Pounds of cotton and bushels of wheat cannot properly be added; but if wheat production was 110 per cent of the previous year's production and cotton production was 106 per cent, it is valid to average these two percentages and to say that the volume of these commodities produced was 108 per cent of the previous year. This assumes they are of equal importance, since each is given the same weight; but if cotton production is six times as important as wheat production, the percentages should be weighted 6 and 1. The average relative secured by this process is referred to as an *index number.*

For most purposes, measurements of the relative change in volume of production or prices would be satisfactory; the use of a fixed-base relative as a substitute for a time series is discussed on pages 117 to 119. The index number based on an average of relatives can be used in exactly the same way as the simple relative, except that it represents the average relative variation in a number of series instead of the variation in a single series. For instance, the relatives on page 118 show the changes in the production of steel ingots between 1926 and 1956. An index of industrial production that measured the composite change in all manufacturing industries would take the same form as the simple relatives measuring changes in steel ingot production.

When data can be added and the single series is reduced to a fixed-base series of relatives, the relatives are called index numbers by some statisticians; others reserve the term "index number" exclusively for an average of relatives derived from series that cannot properly be added. Since many relatives based on a single series are widely used as measures of business conditions, it seems simpler to use the term "index number" to describe both simple relatives and averages of relatives. The term will be used in the following chapters to refer to relatives derived in both ways.

If it is desired that the index number measure the cyclical fluctuations, it is necessary to adjust for trend and seasonal variation by the methods described in Chapter 15 when either of these fluctuations is present in the data. When a number of series are averaged to secure a composite index, the individual series may be adjusted before they are averaged or the composite index may be adjusted. The latter method requires less work and is ordinarily satisfactory.

SIMPLE RELATIVES

Index numbers computed as simple relatives may be fixed-base relatives such as discussed on pages 117 to 119, or chain relatives as discussed on pages 119 to 122. If these relatives are on a monthly basis, they may be adjusted for seasonal variation when necessary. It makes no difference whether the basic data were adjusted for seasonal variation before being reduced to relatives or whether the relatives were used in the computation of the index of seasonal variation. In Table 15-2 (page 365), bank debits were adjusted for seasonal variation and then reduced to fixed-base relatives. This series becomes an index of bank debits, according to the definition of an index number given on page 385. In Table 15-3 (page 370), it was demonstrated how the series could be reduced to fixed-base relatives and then the relatives adjusted for seasonal variation. The results will be identical for the two methods of performing the analysis.

When a series has been adjusted for both seasonal variation and secular trend, such as the demand for motor fuel in Table 15-4 (page 372), the series becomes an index number. The base in this case is not one year or a period of years, but is the computed normal or trend value for each period. The base in such a series of relatives may be described by stating "Normal = 100 per cent."

AVERAGES OF RELATIVES

Some of the well-known methods of computing composite index numbers will be illustrated by using the data on prices and production of four fuels given in Tables 16-1 and 16-2. In every case the index number will be an average of relatives rather than an average of the original data. For this reason the series representing the quantity of fuels produced and the prices for which they were sold have been reduced to relatives, with 1947 equal to 100 in each series. These relatives are given in Tables 16-3 and 16-4.

Table 16-1

PRODUCTION OF FOUR FUELS IN THE UNITED STATES: 1947 TO 1956

Year	Anthracite coal (Million tons)	Bituminous coal (Million tons)	Crude petroleum (Million barrels)	Natural gas (Billion cu. ft.)
1947...	57.2	631.6	1,857	4,582
1948...	57.1	599.5	2,020	5,148
1949...	42.7	437.9	1,842	5,420
1950...	44.1	516.3	1,974	6,262
1951...	42.7	533.7	2,248	7,457
1952...	40.6	466.8	2,290	8,013
1953...	30.9	457.3	2,357	8,397
1954...	27.1	392.0	2,315	8,743
1955...	24.9	469.4	2,484	9,405
1956...	28.6	500.5	2,617	10,275

SOURCE: Anthracite and bituminous coal — *Survey of Current Business*; crude petroleum and natural gas — *World Oil*, February 15, 1957.

Table 16-2

PRICES OF FOUR FUELS IN THE UNITED STATES: 1947 TO 1956

Year	Anthracite coal wholesale price f.o.b. mine (Dollars per ton)	Bituminous coal wholesale price f.o.b. mine (Dollars per ton)	Crude petroleum price at Oklahoma-Kansas wells (Dollars per bbl.)	Natural gas value at wells (Dollars per 1,000 cu. ft.)
1947...	14.11	6.87	1.90	060
1948...	15.67	8.10	2.57	.065
1949...	12.04	5.65	2.57	.063
1950...	12.58	5.58	2.57	.065
1951...	14.19	5.69	2.57	.073
1952...	14.30	5.75	2.57	.078
1953...	15.45	5.79	2.72	.092
1954...	14.01	5.54*	2.82	.101
1955...	12.98	4.82#	2.82	.108
1956...	13.53	5.31#	2.82	.110

*Average for the first four months of 1954.
#Weighted average of industrial and domestic grades.

SOURCE: Anthracite coal, bituminous coal, and crude petroleum — *Survey of Current Business*; natural gas — *World Oil*, February 15, 1957.

Table 16-3

PRODUCTION OF FOUR FUELS IN THE UNITED STATES: 1947 TO 1956

RELATIVES, 1947 = 100

Year	Anthracite coal	Bituminous coal	Crude petroleum	Natural gas
1947...	100.0	100.0	100.0	100.0
1948...	99.8	94.9	108.8	112.4
1949...	74.7	69.3	99.2	118.3
1950...	77.1	81.7	106.3	136.7
1951...	74.7	84.5	121.1	162.7
1952...	71.0	73.9	123.3	174.9
1953...	54.0	72.4	126.9	183.3
1954...	47.4	62.1	124.7	190.8
1955...	43.5	74.3	133.8	205.3
1956...	50.0	79.2	140.9	224.2

SOURCE: Table 16-1.

Table 16-4

PRICES OF FOUR FUELS IN THE UNITED STATES: 1947 TO 1956

RELATIVES, 1947 = 100

Year	Anthracite coal	Bituminous coal	Crude petroleum	Natural gas
1947...	100.0	100.0	100.0	100.0
1948...	111.1	117.9	135.3	108.3
1949...	85.3	82.2	135.3	105.0
1950...	89.2	81.2	135.3	108.3
1951...	100.6	82.8	135.3	121.7
1952...	101.3	83.7	135.3	130.0
1953...	109.5	84.3	143.2	153.3
1954...	99.3	80.6	148.4	168.3
1955...	92.0	70.2	148.4	180.0
1956...	95.9	77.3	148.4	183.3

SOURCE: Table 16-2.

The computation of indexes performed by averaging these relatives in different ways will be illustrated in the remainder of this section. Before proceeding with the discussion of the methods of averaging relatives, however, the following explanation of the symbols that will be used in writing formulas is presented.

P_0 Prices in the base year or period.
P_1 Prices in the month, year, or other period for which the index number is being computed.
Q_0 Quantities in the base year or period. The quantities may be production, consumption, marketings, or any other information that is expressed in physical units.
Q_1 Quantities in the month, year, or other period for which the index number is being computed.
N Number of different series in the index.

It is also possible to indicate the years by means of subscripts. For example, prices for 1947 might be represented by P_{47} and quantities for 1947, by Q_{47}.

Unweighted Arithmetic Mean of Relatives

Table 16-5 gives for each fuel the ratio of production in 1956 to production in 1947; it will be noted that there is considerable variation in these ratios for the different fuels. The problem is to average these relatives to secure a typical ratio of 1956 to 1947, which will be the index number for 1956. The index number for any year may be computed in the same manner.

Table 16-5

COMPUTATION OF INDEX OF FUEL PRODUCTION, 1956

1947 = 100

ARITHMETIC MEAN OF RELATIVES

Fuels	Relatives 1947 = 100
Anthracite coal	50.0
Bituminous coal	79.2
Crude petroleum	140.9
Natural gas	224.2
Total	494.3
Arithmetic mean	123.6

SOURCE: Table 16-3.

If we could assume that the products were of equal importance, an unweighted arithmetic mean of the relatives might appear to be correct.

This mean is computed by adding the relatives and dividing the sum by the number of relatives, which gives an average relative of 123.6. This means that the volume of fuels produced in 1956 was 123.6 per cent of 1947, the base year.

To find the index of prices of these fuels, an arithmetic mean of the price relatives is computed in the same manner in Table 16-6. The sum of the relatives is 504.9, which is divided by four to give the average relative of 126.2. This index of prices means that the level of fuel prices in 1956 was 126.2 per cent of the level of prices in 1947.

Table 16-6

COMPUTATION OF INDEX OF FUEL PRICES, 1956

1947 = 100

ARITHMETIC MEAN OF RELATIVES

Fuels	Relatives 1947 = 100
Anthracite coal	95.9
Bituminous coal	77.3
Crude petroleum	148.4
Natural gas	183.3
Total	504.9
Arithmetic mean	126.2

SOURCE: Table 16-4.

The computations performed above can be summarized by formulas based on the symbols described on page 389. The quantity relatives for the year 1956 are represented by $\frac{Q_1}{Q_0}100$, since Q_1 represents the quantities produced in 1956 and Q_0, the quantities produced in 1947, the base year. The index of quantities produced was computed by the following formula:

$$\textbf{Quantity Index} = \frac{\Sigma\left(\frac{Q_1}{Q_0}100\right)}{N} \quad \textbf{(41)}$$

$$\text{Quantity Index} = \frac{494.3}{4} = 123.6$$

The price relatives for 1956 are represented by $\frac{P_1}{P_0}100$, and the index of fuel prices is computed by the formula:

$$\text{Price Index} = \frac{\Sigma\left(\frac{P_1}{P_0}100\right)}{N} \quad \ldots\ldots\ldots \quad (42)$$

$$\text{Price Index} = \frac{504.9}{4} = 126.2$$

Unweighted Geometric Mean of Relatives

The index of volume of fuel production may be computed by using a geometric mean of the relatives instead of the arithmetic mean. This computation is made in Table 16-7.

$$\text{Log of Quantity Index} = \frac{\Sigma \log\left(\frac{Q_1}{Q_0}100\right)}{N} \quad \ldots\ldots\ldots \quad (43)$$

$$\text{Log of Quantity Index} = \frac{8.09725}{4} = 2.02431$$

$$\text{Quantity Index} = 105.8$$

Table 16-7

COMPUTATION OF INDEX OF FUEL PRODUCTION, 1956

1947 = 100

GEOMETRIC MEAN OF RELATIVES

Fuels	Relatives 1947 = 100	Logarithms of relatives
Anthracite coal	50.0	1.69897
Bituminous coal	79.2	1.89873
Crude petroleum	140.9	2.14891
Natural gas	224.2	2.35064
Total	...	8.09725
Arithmetic mean of logarithms	...	2.02431
Geometric mean	...	105.8

SOURCE: Table 16-3.

The geometric mean of the price relatives is computed by the following formula and found to be 119.2 for 1956.

$$\text{Log of Price Index} = \frac{\Sigma \log \left(\frac{P_1}{P_0} 100\right)}{N} \quad \ldots\ldots\ldots \quad (44)$$

$$\text{Log of Price Index} = \frac{8.30459}{4} = 2.07615$$

$$\text{Price Index} = 119.2$$

This formula is the same as (43) except that the price relatives are substituted for the quantity relatives. The computations are shown in Table 16-8.

Table 16-8

COMPUTATION OF INDEX OF FUEL PRICES, 1956

1947 = 100

GEOMETRIC MEAN OF RELATIVES

Fuels	Relatives 1947 = 100	Logarithms of relatives
Anthracite coal	95.9	1.98182
Bituminous coal	77.3	1.88818
Crude petroleum	148.4	2.17143
Natural gas	183.3	2.26316
Total	...	8.30459
Arithmetic mean of logarithms	...	2.07615
Geometric mean	...	119.2

SOURCE: Table 16-4.

The Problem of Averages

A great deal of attention has been given to the selection of the proper average to use in the computation of an index number. The arithmetic mean, the geometric mean, and the harmonic mean have all been suggested by statisticians; but the geometric mean has certain points of decided superiority over the other two. It was pointed out on page 174 that the

geometric mean is always less than the arithmetic mean of the same data; and in the construction of index numbers, it is generally agreed that the arithmetic mean has an upward bias to the extent that it is larger than the geometric mean. At one time this led to the conclusion that only the geometric mean should be used in index-number construction, since it alone is free from bias. However, the average to be used is related to the weighting system, which will be discussed before any conclusion is stated as to which average should be used.

The Problem of Weights

The unweighted arithmetic mean of the relatives was computed on page 390, assuming that each of the fuels was of the same importance. It is rather obvious that in this problem the fuels are not of equal importance, and it is reasonable to assume that they will seldom be equally important. For this reason a weighted average of the relatives will normally be computed. The problem of how the weights should be determined has been given much attention, and a number of weighting schemes have been developed. Any system of weights must meet two requirements:

1. The amounts used as weights must measure the comparative importance of the different relatives to be averaged.
2. They must be amounts that can properly be added, since the formula for a weighted average involves dividing by the sum of the weights.

The weights should be computed from actual data that measure the relative importance of the different series; although if no data were available, weights might be estimated. We might assign anthracite coal a weight of 10; bituminous coal, a weight of 15; and so on. But since the basic reason for using statistical methods is to replace opinions with objectively derived facts, computed weights should always be used if any kind of information can be obtained to measure the relative importance of the items included in the index.

Many different kinds of data may represent the basis for the computation of weights, but the most generally used is some measure of the *value* of the different items. For the index of fuel prices, the relative importance of each fuel might be determined from the value of each fuel produced in a typical year or in several years. Values of different products expressed in dollars are comparable and so they can be added — a condition that must be met by any system of weights. Table 16-9 gives the value of fuels produced for each year from 1947 to 1956.

The next question is that of the period for which the values should be computed, since the values of the fuels produced in any one year or combination of years might be used as the weights. The base year or period is

considered a particularly desirable basis for the weights, since a normal year or period is usually chosen as the base. On the other hand, it has been confirmed by a considerable amount of research that values for certain years are biased when used as weights either for price or for quantity indexes. In a price index, the use of the values of the base period *gives too much weight to the items with the smallest rise in price, or the largest drop if prices were falling.* This results in the index being too low; or, in other words, this scheme of weights has a *downward bias.* In a quantity index, the values of the base period have the same bias as when used for a price index.

Table 16-9

VALUE OF FUELS PRODUCED: 1947 TO 1956

(THOUSANDS OF DOLLARS)

Year	Anthracite coal	Bituminous coal	Crude petroleum	Natural gas	Total
1947...	807,092	4,339,092	3,528,300	274,920	8,949,404
1948...	894,757	4,855,950	5,191,400	334,620	11,276,727
1949...	514,108	2,474,135	4,733,940	341,460	8,063,643
1950...	554,778	2,880,954	5,073,180	407,030	8,915,942
1951...	605,913	3,036,753	5,777,360	544,361	9,964,387
1952...	580,580	2,684,100	5,885,300	625,014	9,774,994
1953...	477,405	2,647,767	6,411,040	772,524	10,308,736
1954...	379,671	2,171,680	6,528,300	883,043	9,962,694
1955...	323,202	2,262,508	7,004,880	1,015,740	10,606,330
1956...	386,958	2,657,655	7,379,940	1,130,250	11,554,803

SOURCE: Tables 16-1 and 16-2.

Total values for the whole period covered by the index will not suffer from such a bias as described above, but such a weighting scheme requires a great deal of computation. The scheme can be used for constructing an index over a past period of time, but the chief interest in most indexes is in current changes. This requires the use of a weighting scheme based on past data and one that does not need to be changed frequently. The ease with which they can be used is an important recommendation for the values for the base period.

It was stated on page 393 that the arithmetic mean of the relatives gives the index an *upward bias.* Base-period values used as weights result in an index that is too low; that is, they give the index a *downward bias.* While these two biases are not equal, they are always in opposite directions. Sometimes the upward bias of the arithmetic mean is larger and sometimes

the downward bias of the weight scheme is larger. The result of this approximate offsetting of the biases is that the arithmetic mean of relatives weighted with base-year values is not completely correct but it is not biased, since sometimes it is too large and sometimes it is too small. The amount of error is small enough that the formulas described below may be used satisfactorily. This method of weighting is about the simplest that can be used, and the arithmetic mean is the simplest average. The advantages of a method as simple as this are so great that few other methods of construction are now used in any well-known index numbers.

Arithmetic Mean of Relatives Weighted with Base-Year Values

In Table 16-10 the weighted arithmetic mean is computed for the production of fuels, using 1947 as the base year and the values of the base-year production as weights. All the necessary data, both the relatives and the data needed as weights, are given in Tables 16-3 and 16-9. The computations simply consist of finding a weighted average of the relatives in Column (2), using the values in Column (3) as weights. The computation of a weighted average was explained in Chapter 8 and demonstrated in Table 8-5 (page 170). The only difference in Table 16-10 is that a different set of symbols is used to identify the quantities entering into the computation. The relatives to be averaged in Column (2) are multiplied by the weights in Column (3). The products are entered in Column (4) and totaled.

Table 16-10

COMPUTATION OF INDEX OF FUEL PRODUCTION, 1956

1947 = 100

WEIGHTED ARITHMETIC MEAN OF RELATIVES

Base-Year Values as Weights

Fuels (1)	Relatives 1947 = 100 (2) $\frac{Q_1}{Q_0}$	Weights (Base-year values*) (3) $P_0 Q_0$	(4) $P_0 Q_0 \frac{Q_1}{Q_0} 100$
Anthracite coal...	50.0	807,092	40,354,600.0
Bituminous coal..	79.2	4,339,092	343,656,086.4
Crude petroleum..	140.9	3,528,300	497,137,470.0
Natural gas.....	224.2	274,920	61,637,064.0
Total.....	...	8,949,404	942,785,220.4
Index.....	...	...	105.3

*Thousands of dollars.

SOURCE: Tables 16-3 and 16-9.

This total is the numerator of the equation, and is divided by the sum of the weights in Column (3). The formula for a weighted arithmetic mean of quantity relatives using base-year values as weights is as follows:

$$\text{Quantity Index} = \frac{\Sigma\left(P_0Q_0\frac{Q_1}{Q_0}100\right)}{\Sigma(P_0Q_0)} \quad \text{(45)}$$

$$\text{Quantity Index} = \frac{942{,}785{,}220.4}{8{,}949{,}404} = 105.3$$

The computation of an index of fuel prices would be made in the same manner as the computation of the index of fuel production, except that the relatives averaged in this case would be computed from the *prices* of the individual fuels. The weights used should be the same as for the index of fuel production. The details of computation of the price index are given in Table 16-11, which differs from Table 16-10 only in that the price relatives instead of the quantity relatives are entered in Column (2). In the equation used for computing the index, the price relative $\frac{P_1}{P_0}$ is substituted for the quantity relative $\frac{Q_1}{Q_0}$, which results in the following formula:

$$\text{Price Index} = \frac{\Sigma\left(P_0Q_0\frac{P_1}{P_0}100\right)}{\Sigma(P_0Q_0)} \quad \text{(46)}$$

$$\text{Price Index} = \frac{986{,}804{,}490.4}{8{,}949{,}404} = 110.3$$

Weighted Geometric Mean

The computation of a weighted geometric mean raises the question of the bias in the base-year values. Since the geometric mean is not biased, there is no upward bias in the average to offset the downward bias in the weights, as in the case of the arithmetic mean weighted with base-year values. When the geometric mean is used to compute an index number, it is necessary that an unbiased system of weights be used.

One possible system of unbiased weights that can be used with the index for any given year is derived from the average values for the given year and the base year. The values of the current year will be biased upward to the same degree that the values for the base year are biased downward, with the result that the average of the two values will be unbiased. Table 16-12 gives the percentage weights derived from the 1947 and 1956 values, and

Table 16-11

COMPUTATION OF INDEX OF FUEL PRICES, 1956

1947 = 100

WEIGHTED ARITHMETIC MEAN OF RELATIVES
Base-Year Values as Weights

Fuels (1)	Relatives 1947 = 100 (2) $\frac{P_1}{P_0}$	Weights (Base-year values*) (3) P_0Q_0	(4) $P_0Q_0\frac{P_1}{P_0}100$
Anthracite coal...	95.9	807,092	77,400,122.8
Bituminous coal..	77.3	4,339,092	335,411,811.6
Crude petroleum..	148.4	3,528,300	523,599,720.0
Natural gas.....	183.3	274,920	50,392,836.0
Total.....	...	8,949,404	986,804,490.4
Index.....	...	...	110.3

*Thousands of dollars.

SOURCE: Tables 16-4 and 16-9.

Table 16-12

VALUE OF PRODUCTION OF SELECTED FUELS: 1947, 1956, AND AVERAGE OF 1947 AND 1956

USED AS WEIGHT SCHEMES FOR INDEX NUMBERS

Fuels	1947		1956		Average of 1947 and 1956	
	(Thousands of dollars)	Per cent of total	(Thousands of dollars)	Per cent of total	(Thousands of dollars)	Per cent of total
Anthracite coal...	807,092	9.0	386,958	3.3	597,025.0	5.8
Bituminous coal..	4,339,092	48.5	2,657,655	23.0	3,498,373.5	34.1
Crude petroleum..	3,528,300	39.4	7,379,940	63.9	5,454,120.0	53.2
Natural gas.....	274,920	3.1	1,130,250	9.8	702,585.0	6.9
Total.....	8,949,404	100.0	11,554,803	100.0	10,252,103.5	100.0

SOURCE: Table 16-9.

the average values for the two years, 1947 and 1956. It will be noted that there was a substantial change in the relative importance of the different fuels between 1947 and 1956, and neither distribution is a satisfactory measure of the relative importance of the individual fuels in the two years. Since the index number compares one year with the base year, 1947, the

Table 16-13

COMPUTATION OF INDEX OF FUEL PRODUCTION, 1956

1947 = 100

WEIGHTED GEOMETRIC MEAN OF RELATIVES

Fuels	Relatives 1947 = 100 $\frac{Q_1}{Q_0}(100)$	Logarithms of relatives $\log\left[\frac{Q_1}{Q_0}(100)\right]$	Weights (Per cent of average value 1947 and 1956) *wt.*	(*wt.*) $\log\left[\frac{Q_1}{Q_0}(100)\right]$
Anthracite coal..	50.0	1.69897	5.8	9.85403
Bituminous coal.	79.2	1.89873	34.1	64.74669
Crude petroleum	140.9	2.14891	53.2	114.32201
Natural gas....	224.2	2.35064	6.9	16.21942
Total.....	...	...	100.0	205.14215
Weighted mean of logarithms..	...	...	...	2.05142
Weighted geometric mean	...	...	...	112.6

SOURCE: Tables 16-3 and 16-12.

weight scheme should be a measure of the importance of each fuel in both years. In the present situation, with so much difference between the two years, an average of the two distributions is the most logical solution to the problem of finding an unbiased weight scheme for the 1956 index.

Indexes of quantities and prices are computed in Tables 16-13 and 16-14, using the geometric mean with the average values of 1947 and 1956 as weights. The amounts have been calculated by the following formulas:

$$\text{Log of Quantity Index} = \frac{\left[(wt.)\left(\log \frac{Q_1}{Q_0} 100\right)\right]}{\Sigma wt.} \quad . \; . \; . \quad (47)$$

$$\text{Log of Quantity Index} = \frac{205.14215}{100} = 2.05142$$

$$\text{Quantity Index} = 112.6$$

Table 16-14

COMPUTATION OF INDEX OF FUEL PRICES, 1956

1947 = 100

WEIGHTED GEOMETRIC MEAN OF RELATIVES

Fuels	Relatives 1947 = 100 $\frac{P_1}{P_0}(100)$	Logarithms of relatives $\log\left[\frac{P_1}{P_0}(100)\right]$	Weights (Per cent of average value 1947 and 1956) *wt.*	$(wt.)\ \log\left[\frac{P_1}{P_0}(100)\right]$
Anthracite coal..	95.9	1.98182	5.8	11.49456
Bituminous coal.	77.3	1.88818	34.1	64.38694
Crude petroleum	148.4	2.17143	53.2	115.52008
Natural gas....	183.3	2.26316	6.9	15.61580
Total.....	...	...	100.0	207.01738
Weighted mean of logarithms..	...	...	...	2.07017
Weighted geometric mean.....	...	...	...	117.5

SOURCE: Tables 16-4 and 16-12.

$$\text{Log of Price Index} = \frac{\left[(wt.)\left(\log\frac{P_1}{P_0}100\right)\right]}{\Sigma wt.} \quad \ldots\ldots \quad (48)$$

$$\text{Log of Price Index} = \frac{207.01738}{100} = 2.07017$$

$$\text{Price Index} = 117.5$$

FACTORS TEST

It has been stated that the geometric mean of relatives, weighted by values that do not have any bias, gives an index number that is free from bias. It was also stated that the arithmetic mean weighted by base-year values offsets an upward bias in the arithmetic mean against a downward bias in the weighting scheme. It was pointed out that this offsetting of errors does not occur in every case and that in some situations the weight bias will be greater than the bias in the mean, while in other situations the reverse will be true. This section will present a test that can be applied to a given situation to determine how well a formula measures the changes in prices and in quantities. This test is called the *factors test*, or *factor reversal test*.

The basis of the factors test is the fact that value is the product of price and quantity. The value of $7,379,940,000 for petroleum production in 1956 may be factored into the quantity produced, 2,617,000,000 barrels, and the price, $2.82 per barrel. It is also true that since the price of petroleum in 1956 was 148.4 per cent of the 1947 price, and the production in 1956 was 140.9 per cent of the 1947 production, the value of petroleum produced in 1956 would be 209 per cent of 1947. ($1.484 \times 1.409 = 2.09$, or 209 per cent). After calculating this percentage from the price and quantity relatives, it can be checked by computing it directly from the value of petroleum production in 1947 and 1956. The 1947 value was $3,528,300,000 and the 1956 value of $7,379,940,000 is 209 per cent of the 1947 value. This calculation checks the accuracy of the computation made from the price and quantity relatives. Such a relationship obviously holds for any commodity used in the construction of index numbers.

Since the product of the price relative and the quantity relative for each individual commodity entering into an index number equals the value relative, it follows that the product of a composite price index and a composite quantity index for the same commodities should also equal the value relative computed for the same commodities. If the product of a price index and a quantity index, each computed by a given formula (such as the arithmetic mean of relatives with base-year values as weights), equals the value relative computed directly from the original price and quantity data, the index number formula is said to have met the factors test. If the formula does not meet the factors test, it is subject to criticism on the grounds that if the relatives for individual commodities used in the computation give the correct value relative, the composite price index and the composite quantity index should give the same value relative as computed from the original data.

The factors test can be applied to the price and quantity indexes for 1956 computed on page 396. The price index for 1956 computed by the arithmetic mean of relatives weighted with base-year values is 110.3 and the quantity index computed by the same method is 105.3. The product of the two indexes is 116:

$$1.053 \times 1.103 = 1.161 \text{ or } 116 \text{ per cent}$$

while the value relative computed from the total value of fuels produced in 1956 and in 1947 was 129:

$$\$11,554,803 \div \$8,949,404 = 1.29 \text{ or } 129 \text{ per cent.}$$

The difference between 116 per cent and 129 per cent is a rather serious discrepancy and emphasizes the weakness of this formula. The changes that have been taking place in the relative positions of coal, oil, and natural gas as fuels introduce more bias than will usually be present in this formula. However, it should be recognized that the advantage of a good formula is the fact that it will give the correct answer under all circumstances. The arithmetic mean of relatives using base-year values as weights works very well most of the time; this fact, combined with the fact that it is simple to compute and easy to understand, accounts for its popularity.

When the factors test is applied to the geometric mean of relatives weighted by the average values of fuels in 1947 and 1956, a much better check is found. The product of the quantity and the price indexes gives a value relative of 132:

$$1.126 \times 117.5 = 1.323 \text{ or } 132 \text{ per cent.}$$

This is reasonably close to the true value relative of 129 per cent for 1956. This close agreement on the factors test is indicative of the accuracy of this formula, but the difficulty of using it for a current index prevents it from being a popular one.

Table 16-15 gives the index of prices and the index of quantities produced for each year from 1947 to 1956. It also shows the product of the price and quantity indexes for each year and the true value relative computed from the price and production data.

METHOD OF AGGREGATES

The aggregative method of computing index numbers is sometimes considered a different method from the method of averaging relatives, but it is in reality an alternative method of computing some of the indexes described above. Methods using the arithmetic mean of relatives may be expressed

Table 16-15

FACTORS TEST FOR INDEXES OF PRODUCTION AND PRICES OF FUELS: 1947 TO 1956

INDEXES COMPUTED BY ARITHMETIC MEAN WEIGHTED WITH BASE-YEAR VALUES

Year	Production index	Price index	Product of price index and production index	Value relative 1947 = 100
(1)	(2) $\frac{\Sigma\left(P_0Q_0\frac{Q_1}{Q_0}100\right)}{\Sigma P_0Q_0}$	(3) $\frac{\Sigma\left(P_0Q_0\frac{P_1}{P_0}100\right)}{\Sigma P_0Q_0}$	(4) Col. (2) x Col. (3)	(5) $\frac{\Sigma(P_1Q_1)}{\Sigma(P_0Q_0)}100$
1947...	100.0	100.0	100.0	100.0
1948...	101.4	123.8	125.5	126.0
1949...	83.1	104.1	86.5	90.1
1950...	92.7	104.1	96.5	99.6
1951...	100.4	106.3	106.7	111.3
1952...	96.2	107.0	102.9	109.2
1953...	95.6	111.9	107.0	115.2
1954...	89.4	111.7	99.9	111.3
1955...	99.0	106.4	105.3	118.5
1956...	105.3	110.3	116.1	129.1

SOURCE: Tables 16-3, 16-4, and 16-9.

as an aggregative formula, and this form of computation is usually used in preference to the average of relatives because the computations are easier. Methods using the geometric mean cannot be computed by the aggregative method.

On pages 169 to 173, in the discussion of averaging ratios, it was pointed out that whenever possible it is better to compute a weighted average of ratios by going back to the basic data from which the ratios were computed rather than to try to weight the ratios. Since index numbers are averages of ratios, this general principle applies to their computation; and it is better to use an aggregative formula if it can be applied. In some cases relatives and values may be available but the original data may not be. In such a case it is necessary to use an average of relatives.

The relationship between the aggregative method and the average of relatives can be illustrated by the computation of the index of quantities of fuels produced in Table 16-10 (page 395). The amounts in Column (4) were derived for each commodity by multiplying the 1947 value of fuel

produced (P_0Q_0) by the quantity relative $\frac{Q_1}{Q_0}$, or $P_0Q_0 \frac{Q_1}{Q_0}100$. For anthracite coal, the computation was

$$\$14.11 \times 57{,}200{,}000 \text{ tons} \times \frac{28{,}600{,}000}{57{,}200{,}000} 100 = \$40{,}354{,}600{,}000.$$

But instead of dividing 28,600,000 by 57,200,000 and then multiplying, it is easier to cancel the 57,200,000 from the numerator and the denominator as follows:

$$\cancel{57{,}200{,}000} \text{ tons} \times \$14.11 \times \frac{28{,}600{,}000}{\cancel{57{,}200{,}000}} 100 = \$40{,}354{,}600{,}000.$$

It may, therefore, be written that

$$(P_0Q_0)\left(\frac{Q_1}{Q_0}100\right) = (P_0Q_1)\ 100$$

and the formula for the quantity index becomes $\frac{\Sigma(P_0Q_1 \cdot 100)}{\Sigma(P_0Q_0)}$. Or, since it is simpler to multiply the final ratio by 100 than to multiply each of the individual products by 100, the formula is usually written

$$\text{Quantity Index} = \frac{\Sigma(P_0Q_1)}{\Sigma(P_0Q_0)}100. \quad \ldots\ldots\ldots \quad (49)$$

The simplification of the numerator of the equation gives the result with one multiplication for each commodity instead of a division to secure the relative $\frac{Q_1}{Q_0}$ and then the multiplication of the relative by the weight P_0Q_0.

The work sheet for the computation of the index by the method of aggregates must be set up differently, since it must contain both 1947 and 1956 quantities and 1947 prices. This is done in Table 16-16 where P_0Q_1 is computed for each commodity from the original prices and production data given in Tables 16-1 and 16-2 (page 387). The resulting index is the same as computed in Table 16-10 (page 395), except for a slight difference due to rounding.

The formula for the price index by the method of aggregates is derived from formula (46) by the same process of simplifying the numerator used in deriving formula (49). The quantities in Column (4) of Table 16-11 can be secured directly from the price and quantity data, since $P_0Q_0 \times \frac{P_1}{P_0} = P_1Q_0$. In Table 16-17 the quantities P_1Q_0 are computed from the original

Table 16-16

COMPUTATION OF INDEX OF FUEL PRODUCTION, 1956

1947 = 100

METHOD OF AGGREGATES

Fuel	Units	Price per unit 1947 P_0	Quantities produced		P_0Q_0	P_0Q_1
			1947 Q_0	1956 Q_1		
Anthracite coal..	Tons	$14.11	57,200,000	28,600,000	807,092,000	403,546,000
Bituminous coal.	Tons	6.87	631,600,000	500,500,000	4,339,092,000	3,438,435,000
Crude petroleum	Barrels	1.90	1,857,000,000	2,617,000,000	3,528,300,000	4,972,300,000
Natural gas....	Th. cu. ft.	.06	4,582,000,000	10,275,000,000	274,920,000	616,500,000
Total.....	...	...	...	...	8,949,404,000	9,430,781,000
Index.....	...	...	...	...	100.0	105.38

SOURCE: Tables 16-1 and 16-2.

Table 16-17

COMPUTATION OF INDEX OF FUEL PRICES, 1956

1947 = 100

METHOD OF AGGREGATES

Fuel	Units	Quantity produced 1947 Q_0	Price per unit		P_0Q_0	P_1Q_0
			1947 P_0	1956 P_1		
Anthracite coal..	Tons	57,200,000	$14.11	$13.53	807,092,000	773,916,000
Bituminous coal.	Tons	631,600,000	6.87	5.31	4,339,092,000	3,353,796,000
Crude petroleum	Barrels	1,857,000,000	1.90	2.82	3,528,300,000	5,236,740,000
Natural gas....	Th. cu. ft.	4,582,000,000	.06	.11	274,920,000	504,020,000
Total.....	...	...	...	...	8,949,404,000	9,868,472,000
Index.....	...	...	...	...	100.0	110.3

SOURCE: Tables 16-1 and 16-2.

data, and differ from $P_0Q_0 \frac{P_1}{P_0}$ computed in Column (4), Table 16-11, only because of the variations due to rounding. The formula for a price index by the method of aggregates is

$$\text{Price Index} = \frac{\Sigma(P_1Q_0)}{\Sigma(P_0Q_0)} 100 \quad \ldots\ldots\ldots \quad (50)$$

GENERAL COMMENTS ON INDEX NUMBER CONSTRUCTION

The discussion in this chapter has been devoted to the problems of combining statistical series that cannot properly be added because they are not in comparable units. While this is the central problem of index number construction, it should not be overlooked that the computation of an index number makes use of a wide variety of statistical techniques that have been discussed previously. Although these techniques are not used solely in the construction of index numbers, they will be mentioned to emphasize the fact that the business uses of statistics employ a wide variety of methods.

Sampling

Most index numbers are based on samples rather than on universe data, so the problems of sampling precision enter into the computation of an index number.

Comparability of Data

In comparing prices or volume of production over a period of time, it is essential that a series measure the same factors throughout the period. This gives rise to many difficult problems in constructing index numbers, since the units in which prices and production are measured change with technological improvements. When the price of this year's automobile is compared with the price of the same make in 1926, it is obvious that the units are not the same. Problems of this type, which are present in all types of statistical analysis involving time series, have been discussed in earlier chapters. Securing comparable data for use in constructing an index number is basically no different from the problem of securing comparable data in any other type of analysis, but it is a problem that must be given careful attention.

Selection of Commodities and Weights

In the construction of an index number, it is necessary to define the problem that is to be solved by the index before deciding which series to include. If the index is to measure the monthly changes in industrial production, for example, it is necessary to include in the index commodities that are produced in factories and that lend themselves to the measurement of the monthly volume of production. If an index is to measure the prices received by farmers for their products, it must be based on series that measure the prices paid to farmers. The farm price of cotton is not the same as the price on the cotton exchanges. Decisions of this kind are not unique to the construction of index numbers, for such decisions must be made as part of the solution to many business problems.

In selecting the weights for an index number, attention must be paid to the use that will be made of the result. If a price index is to measure the changes in the prices paid by consumers, it is necessary for the weights applied to the various commodities to reflect the importance of the individual items to consumers. Likewise, the weights for an index of farm prices should reflect the importance of the various commodities in the income of farmers. Frequently the commodities included in an index and the weights assigned the items do not give the information wanted for the solution of a particular problem. An index may be computed for the specific problem at hand, or the best available index may be used in spite of the fact that it does not measure exactly what the user wants the index to measure.

Choice of Base Period

If an index number is computed for a specific purpose, the base selected may be the period that is the most useful for comparative purposes. However, most indexes used by business are published by public agencies, and since these indexes are used for many purposes it is desirable that the base be a period with which frequent comparisons may be made. Sometimes the weights are computed from data for the base period, and in this situation it is desirable that the base period be one in which there was as normal a relationship among the individual items as possible.

Since index numbers are expressed as percentage relatives with the base period equal to 100, it is possible to shift the base to any period covered by the index by merely dividing the index for each period by the value for the new period that is to be used as the base. For example, if 1947 is the base with the index number for 1957 equal to 120 and 1958 equal to 132, the year 1957 may be made the base by dividing the 1958 value of 132 by 120, giving a new percentage relative of 110 for 1958 (1957 equal to 100). The index number for 1947 on the new base would be found by dividing 100 (the value of the index for 1947) by the value for the new base year, 120. The index for 1947 on the 1957 base would be 83.

$$\frac{100}{120}\,100 = 83$$

STUDY QUESTIONS

1. In what important respect are the index numbers in Chapter 16 different from the ratios discussed in Chapter 6?

2. Explain the logic that underlies the use of the index number technique to combine series that cannot be added.

3. Basset Jones in his book *Horses and Apples* argues that it is incorrect to average relatives from series that cannot be averaged. Can you support this

argument, or do you consider it proper to average relatives based on data that cannot themselves be properly averaged?

4. It was stated in Chapter 8 that averaging ratios offers serious problems that must be given careful attention. What is the importance of this principle in the construction of index numbers, since index numbers are averages of relatives?

5. The following data give the price relatives for commodities A and B for the years 1947 and 1957, and the arithmetic mean for the two years.

	1947	1957
Commodity A	100	50
Commodity B	100	200
Arithmetic mean	100	125

If 1957 were chosen as the base instead of 1947, the following relatives would show the relationship of the price of each commodity in 1947 to the price in 1957. The arithmetic mean of the relatives for each year is also given.

	1947	1957
Commodity A	200	100
Commodity B	50	100
Arithmetic mean	125	100

The relationship between the two sets of relatives for commodity A is the same, no matter which year is the base. In each case the 1947 relative is double the 1957 relative. Likewise the two sets of relatives for commodity B show the same relationship between prices for the two years. In each case the 1947 relative is one half that of 1957.

The averages of the two sets of relatives, however, give two completely different measures of relationship between the two years. When 1947 is taken as the base, the arithmetic mean indicates that the average price of the commodities in 1957 was 25 per cent higher than in 1947. But when 1957 is taken as the base, the average of the relatives indicates that *prices in 1947 were 25 per cent higher than in 1957*. Obviously both of these results cannot be correct. Do you consider one of them to be correct? Explain.

6. Compute the geometric mean of the relatives in Question 5 and explain what the results show.

7. It was stated in the text that the arithmetic mean is biased when used to average relatives. Does this mean that it is incorrect to use the arithmetic mean to average relatives in constructing index numbers? Give reasons for your answer.

8. The index of consumer prices compiled by the Bureau of Labor Statistics now includes items that were not for sale in the years before World War II. For example, television sets were not included in the index in 1939. Does this mean that a valid comparison cannot be made between the level of prices in 1957 and 1939, since all the commodities in the two years were not the same?

9. Explain the meaning of the "factors test" for an index number. Does it seem reasonable to require that an index number formula give consistent answers when the factors test is applied?

10. Formula (50) is sometimes considered as a significant formula in its own right. The aggregate for the base year represents the value of the goods sold in the base year, computed by multiplying the quantity sold by the price per unit. The aggregate for the current year represents the value of the goods sold in the base year if they were valued at current year prices. The ratio between these two aggregates is a weighted average of the change in prices.

(a) Why would it not be just as logical to construct an index number using quantities in the current year (Q_1) as the multiplier in computing the aggregates? Write the formula that would be used in such a calculation.

(b) What is your opinion of using an average of Q_0 and Q_1 as the multiplier in constructing an index by the aggregative method? Write the formula that would be used in such a calculation.

PROBLEMS

1. The data in the following tables will serve as the information to be used in Problems 1 to 7 to construct various types of index numbers of prices and volume of production.

PRICES OF FOUR BUILDING MATERIALS: 1950 TO 1956
(DOLLARS PER UNIT)

Material	Unit	1950	1951	1952	1953	1954	1955	1956
Southern pine, boards No. 2.....	1,000 bd. ft.	74.36	79.84	80.45	78.69	73.91	78.83	82.21
Portland cement.....	barrels	2.35	2.55	2.54	2.67	2.76	2.78	2.89
Brick, unglazed.....	1,000	25.67	27.33	27.34	27.85	28.22	29.11	30.61
Linseed oil	pound	.18	.21	.16	.15	.15	.13	.14

SOURCE: *Statistical Abstract of the United States*, 1957, and *Business Statistics*, 1957.

PRODUCTION OF FOUR BUILDING MATERIALS: 1950 TO 1956

Material	Unit	1950	1951	1952	1953	1954	1955	1956
Southern pine, boards No. 2.....	million bd. ft.	9,939	8,495	8,572	7,581	8,004	8,700	8,436
Portland cement.....	million barrels	226	246	249	264	271	297	316
Brick, unglazed.....	millions	6,333	6,625	5,889	5,874	6,153	7,148	7,320
Linseed oil.....	million pounds	750	759	545	503	651	618	620

SOURCE: *Business Statistics*, 1957.

(a) The first step in the construction of the index numbers is to reduce all the series to percentage relatives with the year 1950 equal to 100.

(b) Compute the value of production of each building material in each year. Express the value of production of each building material as a percentage of the total production for the year. Prepare this distribution of values for each year.

(c) Construct a line chart showing the four price relatives on the same chart for the years 1950 to 1956.

(d) Construct a line chart showing the four production relatives on the same chart for the years 1950 to 1956.

2. (a) Construct an index of prices for 1956 using an unweighted arithmetic mean of the relatives.

(b) Construct an index of production for 1956 using an unweighted arithmetic mean of the relatives.

(c) Apply the factors test to the two indexes just computed. What conclusion can you draw from this test?

3. (a) Construct an index of prices for 1956 using an unweighted geometric mean of the relatives.

(b) Construct an index of production for 1956 using an unweighted geometric mean of the relatives.

(c) Apply the factors test to the two indexes just computed. How do the results of this test compare with the same test applied to the arithmetic means computed in Problem 2?

4. (a) Construct an index of prices for 1956 using the arithmetic mean of relatives weighted by the values of the building materials produced in 1950.

(b) Construct an index of production for 1956 using the arithmetic mean of relatives weighted by the values of the building materials produced in 1950.

(c) Apply the factors test to the two indexes just computed.

5. (a) Construct an index of prices for 1956 using a weighted geometric mean of the relatives. Use as weights the average value of building materials produced in 1950 and 1956.

(b) Construct an index of production for 1956 using a weighted geometric mean of the relatives. Use as weights the average value of building materials produced in 1950 and 1956.

(c) Apply the factors test to the two indexes just computed. How does the formula used in this problem compare with the use of the factors test in Problem 4?

6. Construct an index of prices for 1956 using the aggregative formula (50), given on page 404. How does the result compare with the index computed in Problem 4 (a)?

7. Construct an index of production for 1956 using the aggregative formula (49), given on page 403. How does the result compare with the index computed in Problem 4 (b)?

8. The following data show the index of wholesale prices for the years 1947 to 1957. The base period is the three-year average 1947 to 1949.

(a) Shift the base of these index numbers to the year 1950, the year the Korean War started.

(b) Which base shows more clearly the changing prices following the start of the Korean War?

INDEX OF WHOLESALE PRICES: 1947 TO 1957

Year	Index	Year	Index	Year	Index	Year	Index
1947	96.4	1950	103.1	1953	110.1	1956	114.3
1948	104.4	1951	114.8	1954	110.3	1957	117.6
1949	99.2	1952	111.6	1955	110.7		

SOURCE: *Economic Indicators*, January, 1958.

9. A man made a will in 1941 providing that after his death his widow should receive a monthly income of $2,000 from a trust fund. Since he realized that the war was likely to cause a substantial rise in the price level, he provided that the payments should be adjusted each month to insure that the purchasing power of the monthly payment would remain the same as it was in 1941. The will specified that the purchasing power of the dollar should be measured by the cost of living index published by the Bureau of Labor Statistics, and if this index were revised or replaced by another index, the purchasing power should then be computed by the use of the revised or new index.

The following table shows the values of the Consumers' Price Index for the twelve months of 1947 (by 1947 the name of the index had been changed from "Cost of Living Index" to "Consumers' Price Index"). In 1953 the Consumers' Price Index was revised using the average for the three years 1947 to 1949 as the base. The values for the twelve months of 1957 are also given in the table.

(a) Compute the amount of income in current dollars to be paid to the widow in each month of 1947. The average value of the index for 1941 was 105.2, with 1935 to 1939 = 100.

(b) Compute the amount of income in current dollars to be paid to the widow in each month of 1957. The average value of the index for 1941 was 62.9, with 1947 to 1949 = 100.

CONSUMERS' PRICE INDEX

BY MONTHS, 1947 AND 1957

Month	1947 index 1935–39 = 100	1957 index 1947–49 = 100	Month	1947 index 1935–39 = 100	1957 index 1947–49 = 100
Jan.	153.6	118.2	July	158.8	120.8
Feb.	153.5	118.7	Aug.	160.7	121.0
Mar.	156.7	118.9	Sept.	164.3	121.1
Apr.	156.6	119.3	Oct.	164.3	121.1
May	156.4	119.6	Nov.	165.4	121.6
June	157.5	120.2	Dec.	167.5	121.6

SOURCE: 1947 data — *Business Statistics*, 1951 Edition.
1957 data — *Economic Indicators*, January, 1958.

Part

V

ANALYSIS OF RELATIONSHIP

chapter 17 ♦ Correlation

chapter 18 ♦ Measurement of Simple Linear Correlation

chapter 19 ♦ Measurement of Non-linear and Multiple Correlation

17 CORRELATION

The methods of analysis described in the preceding chapters represent methods that are used to summarize the characteristics of one statistical series, although reference has been made occasionally to the problem of comparing two or more series. The methods of correlation analysis, which are described in this chapter, are systematic procedures that are used to measure the degree of relationship between two or more series. It will be found that, when attempting to explain the fluctuations in one time series by comparing them with fluctuations in another series, correlation analysis has many applications in business.

LINE CHARTS

Chart 17-1 shows the relationship between the fluctuations in two time series by plotting them on the same grid. This method, which has been used in previous charts, is a standard method of comparing the cyclical and seasonal fluctuations and the secular trend of two or more time series. The comparison of seasonal fluctuations in different series was illustrated in Chart 14-6 (page 346), and the comparison of the secular trend of two series, in Chart 13-5 (page 309). The discussion in Chapter 11 of using two Y scales on arithmetic and semilogarithmic charts was related to the problem of comparing the fluctuations in two time series.

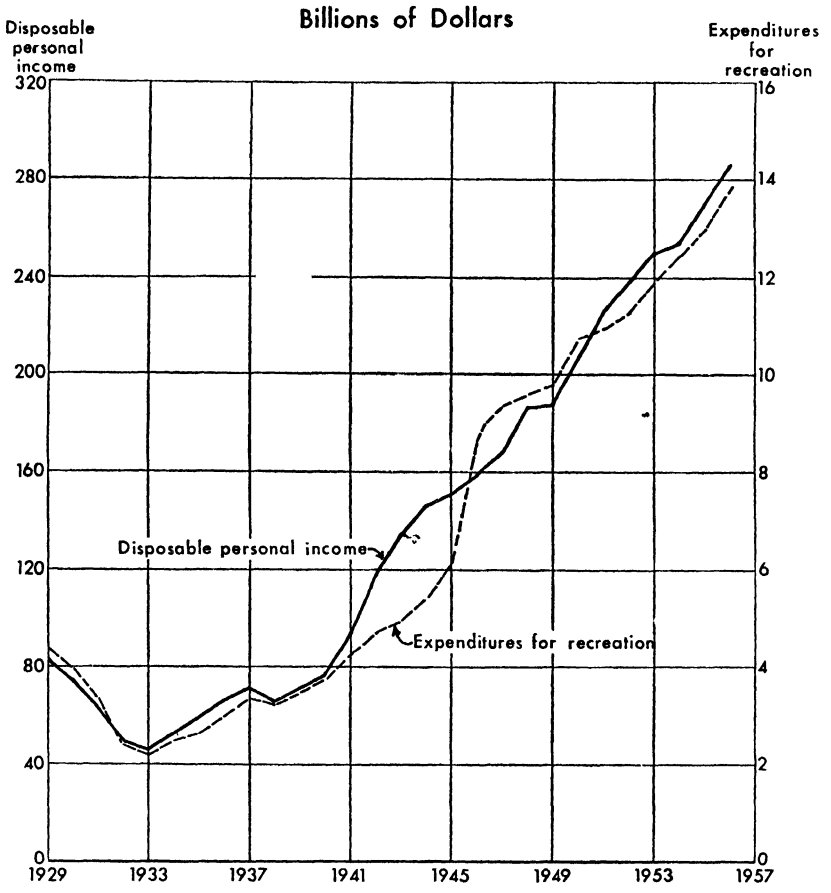
Comparing two time series by means of two line charts is made possible by the fact that the common time scale can be plotted on the horizontal axis of the chart; and since each series is related to the same time interval, they can be compared with each other. Types of classifications other than time series, however, require a different graphic method of showing relationships; the following section illustrates a method that can be used to compare related variables of any kind.

SCATTER DIAGRAMS

Chart 17-2 shows the relationship between prices and net earnings per share for 25 common stocks with above-average stability and long dividend records. These stocks were recommended by the Brookmire Investors Service for investors requiring a high degree of principal safety and assured income. It is to be expected that a substantial portion of the variation in

Chart 17-1

DISPOSABLE PERSONAL INCOME AND CONSUMER EXPENDITURES FOR RECREATION IN THE UNITED STATES, 1929-1956



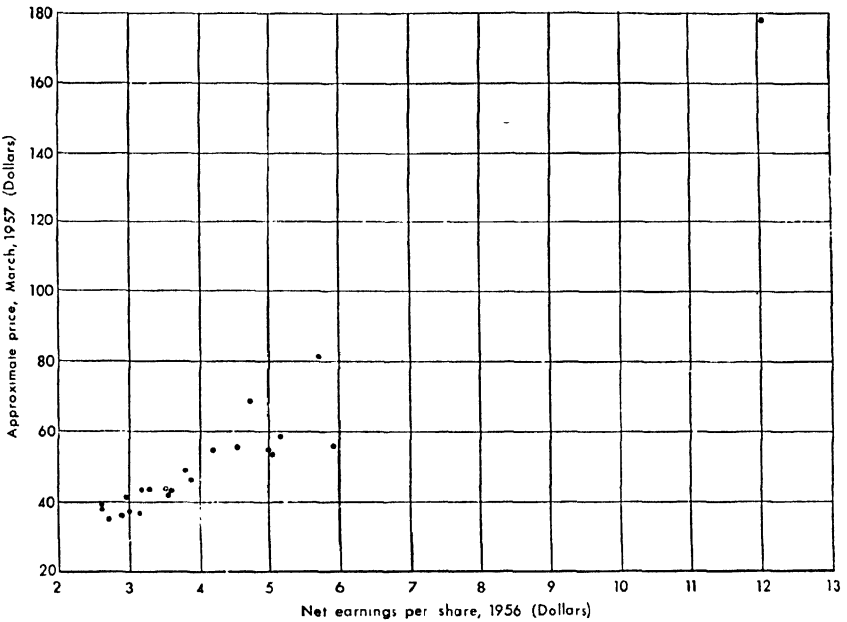
Source of data. *National Income, 1954*
Survey of Current Business, July 1957

the prices of these stocks would be related to the earnings of the companies, and Chart 17-2 was constructed to measure the extent of this relationship. The method of construction is typical of all *scatter diagrams*, which are used to show the amount of correlation existing between two series.

In the scatter diagram, the scale on the X axis represents the net earnings per share in 1956; and the scale on the Y axis represents the approximate price at which the common stock was selling early in 1957. For each company a dot was located on the X axis with respect to the amount of the 1956 earnings and on the Y axis according to the approximate price of

Chart 17-2

SCATTER DIAGRAM OF 1957 PRICES AND 1956 EARNINGS OF 25 CONSERVATIVE* COMMON STOCKS



*Stocks with an above average stability and long dividend records

Source of data *Brookmire Investment Reports*, Spring 1957

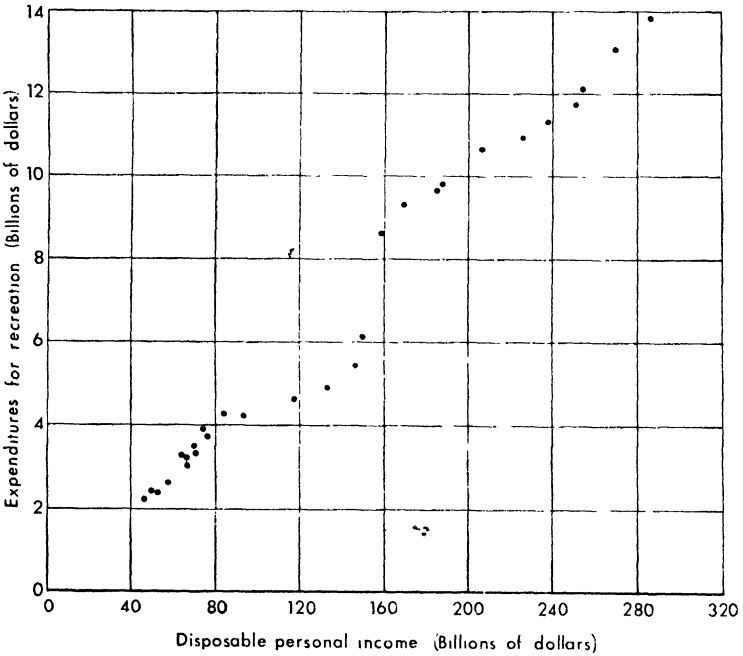
the common stock. A dot was located in this manner for each of the 25 companies, and the resulting scatter diagram summarizes the relationship between the prices of the common stocks and the earnings of the company in the previous year.

In general it appears that the larger the earnings per share, the higher the price of the common stock of the company, although the relationship is not rigidly fixed. The more the plotted points "scatter" over a chart, the less relationship there is between the two variables. The more nearly the points come to falling on a line, the higher the degree of relationship.

A scatter diagram may be used to show the relationship between two variables whenever these variables are both derived from a common situation. In Chart 17-2 the earnings and the price of each common stock were both derived from the same company. The two values of the variable might be the number of persons in a county and the sales of a given product in that county. Another example might be the score of an employee on an aptitude test and the performance of the same employee on a given machine. The two variables might relate to the same time unit, such as the

Chart 17-3

SCATTER DIAGRAM OF DISPOSABLE PERSONAL INCOME AND CONSUMER EXPENDITURES FOR RECREATION IN THE UNITED STATES, 1929-1956



Source of data *National Income, 1954*
Survey of Current Business, July, 1957

data in Chart 17-3, where the amount of disposable income in a given year represents the X variable and the amount of personal expenditures on recreation in the same year is the Y variable. Any two time series can be plotted on a scatter diagram in this manner to show the degree of correlation between the two series. Chart 17-3 shows disposable personal income and expenditures for recreation for 1929 to 1956 as a scatter diagram instead of two lines on a time series chart. The extent to which the points fall in a line indicates the degree of correlation between the two series. In Chart 17-1 the extent to which the two curves move parallel to each other measures the degree of relationship between the two series.

The scatter diagram may be regarded as a specialized chart which is used to show correlation analysis. Conventional line charts of time series, such as Chart 17-1, are more in the nature of general-purpose charts that may also be used to show the relationship between two series, particularly when the relationship is close. The scatter diagram may be used to show the degree of relationship between two variables under any circumstances and is a simple but effective tool for this type of analysis.

The use of graphs to indicate the degree of correlation is an effective first step in the analysis of relationships. Mathematical methods of analysis, which will be described in the next two chapters, can be used for more extensive analysis. However, no matter how elaborate the analysis that is to be performed, it is generally worth the time it requires to make a graphic analysis of the data before proceeding with further calculations. The scatter diagram will permit one to see approximately how close a relationship exists between the two variables, and will also indicate whether the relationship is direct or inverse. If large values of the Y variable are associated with large values of the X variable, the correlation is *direct* or *positive*. The plotted points will be grouped about a line that has a positive slope. If large values of the Y variable are associated with small values of the X variable, the slope of the line formed by the points will be *negative* and the correlation is *inverse*. Chart 17-6 (page 417), measuring the relationship between consumption of heating oil per burner and temperature, shows this inverse relationship, with the consumption of heating oil being smaller in the years with the higher average temperature.

RELATIONSHIPS INVOLVING MORE THAN TWO VARIABLES

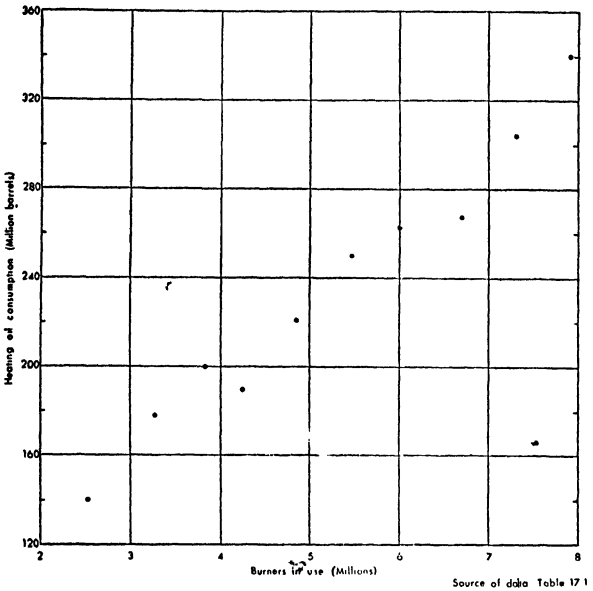
Since the factors affecting business problems are often extremely complex, it is not always possible to limit the analysis of a problem to the relationship between two variables. If it is necessary to measure the correlation between more than two variables, the simple scatter diagram described above will not be adequate. Of course, it is possible to show the relationship between all of the variables involved by taking them two at a time and plotting scatter diagrams, but this will not usually serve the purpose.

Chart 17-4 shows the relationship between the consumption of heating oil and the number of burners in use, plotted from the data given in Table 17-1. It seems logical to assume that the consumption of heating oil is related to the number of burners in use, and the extent of this correlation is shown by Chart 17-4. It also seems logical to assume that the consumption of heating oil would vary from year to year because of variations in temperature. Chart 17-5 is a scatter diagram showing the relationship between heating oil consumption and temperature for the same period used in Chart 17-4.

The scatter of the points on Chart 17-5 is so wide that it seems to indicate no relationship between temperature and heating oil consumption, in spite of the fact that there is good reason to think that some relationship should be visible. There is the possibility that the relationship between temperature and consumption of heating oil is obscured by the high correlation between the number of burners and consumption. It is possible to

Chart 17-4

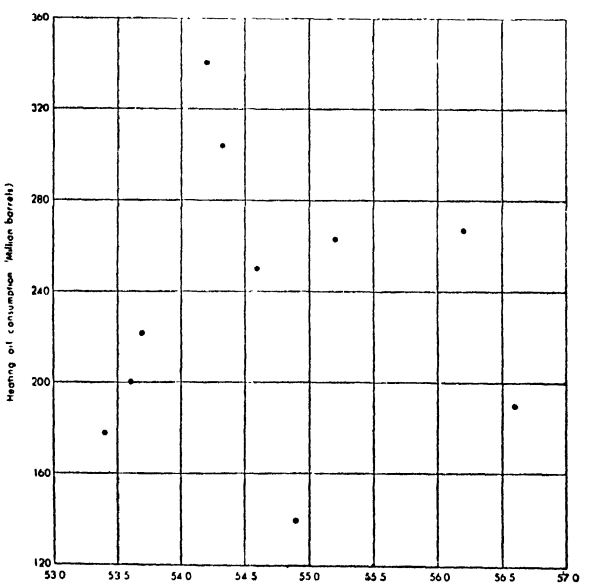
SCATTER DIAGRAM OF HEATING OIL CONSUMPTION
AND NUMBER OF DOMESTIC BURNERS IN USE, 1946-1955



Source of data Table 17 1

Chart 17-5

SCATTER DIAGRAM OF HEATING OIL CONSUMPTION AND AVERAGE
ANNUAL TEMPERATURE IN NEW YORK CITY, 1946-1955



Source of data Table 17 1

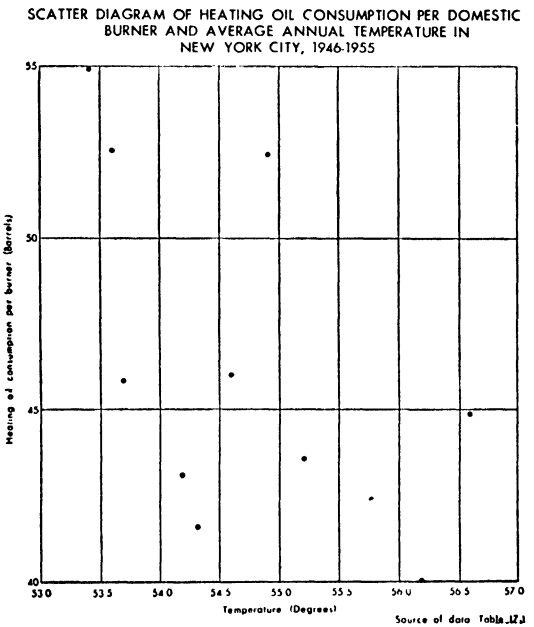
Table 17-1

HEATING OIL CONSUMPTION, DOMESTIC BURNERS IN USE, HEATING OIL CONSUMPTION PER BURNER, AND AVERAGE ANNUAL TEMPERATURE IN NEW YORK CITY: 1946 TO 1955

Year	Heating oil consumption (Million barrels)	Domestic burners in use (Millions)	Heating oil consumption per burner (Barrels)	Average temperature in New York City (Degrees)
1946...	140	2.67	52.4	54.9
1947...	178	3.24	54.9	53.4
1948...	200	3.81	52.5	53.6
1949...	190	4.23	44.9	56.6
1950...	221	4.83	45.8	53.7
1951...	250	5.44	46.0	54.6
1952...	263	6.03	43.6	55.2
1953...	267	6.68	40.0	56.2
1954...	304	7.31	41.6	54.3
1955...	340	7.90	43.1	54.2

SOURCE: Heating oil consumption and number of domestic burners in use — *Petroleum Facts and Figures*, 1956; average temperature in New York City — *World Almanac*, 1957.

Chart 17-6



remove the effect of the number of burners on consumption by computing heating oil consumption per burner. These values are given in Table 17-1, and Chart 17-6 shows graphically the relationship between heating oil per burner and temperature. It indicates that consumption per burner is less when the temperature is high, and greater when the temperature is low; in other words, the correlation is inverse, or negative.

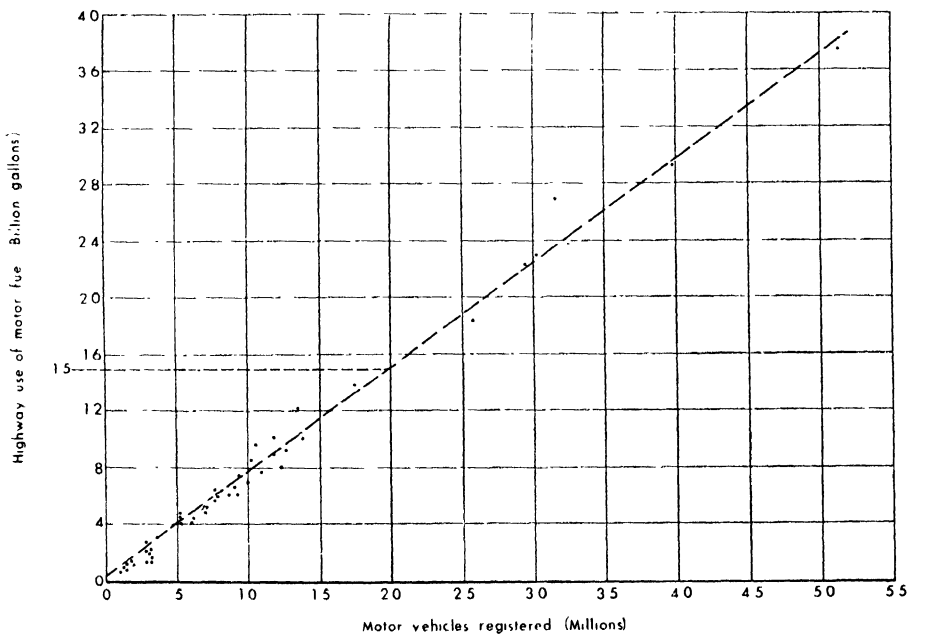
The preceding calculations represent a method that makes it possible to allow for the influence of one variable and then to measure the relationship between two other variables. In Chapter 19 the discussion of multiple and partial correlation presents a systematic approach to this problem.

ESTIMATES BASED ON A SCATTER DIAGRAM

The scatter diagrams shown in the preceding pages give information on how two factors are related, but the diagrams may also be used to estimate or predict the value of one variable when the value of the other is known. Chart 17-7 shows the relationship between highway consumption of motor fuel in each state and the number of motor vehicles registered. Using the same methods employed in computing trend lines in Chapter 12, a straight line was fitted to the two variables and plotted on Chart 17-7. This line represents the average relationship between the number of motor vehicles registered and the highway use of motor fuel. The small scatter about the line indicates a high degree of correlation between the two variables, which means that the consumption of motor fuel for use on the highways is very closely related to the number of motor vehicles registered in the state.

Chart 17-7

SCATTER DIAGRAM OF HIGHWAY USE OF MOTOR FUEL AND NUMBER OF MOTOR VEHICLES REGISTERED



Source of data Agricultural Statistics, 1955

The line of average relationship on Chart 17-7 can be used to estimate the amount of motor fuel that would be consumed on the highways of any area for which the number of motor vehicle registrations is known. For example, an area with 2 million vehicles registered would be estimated to consume approximately 1.5 billion gallons of motor fuel on the highways in a year. This is determined graphically by locating on the X axis the number of motor vehicles registered in the area and then drawing a vertical line to intersect with the line of average relationship. The point at which a line drawn parallel to the base line intersects the Y axis is then read as the estimated amount of motor fuel used on the highways. An estimate of motor fuel use on the highways can be made for any number of motor vehicles in the same manner.

In using the above estimate, it must be understood that the line of average relationship does not show the exact relationship between the two variables. It can be seen that the points scatter somewhat from this line, and the amount they scatter may be taken as an indication of the amount of error the estimate is likely to contain. The more the scatter, the less reliable the estimate would be considered; the more closely the points are concentrated around the line of average relationship, the more reliable the estimate.

In making estimates on the basis of relationship between two variables, it is important to keep in mind that there is a possibility of making a serious mistake if any basic change should occur in the factors affecting the relationship between the variables. The variations of the plotted points from the line of average relationship represent the effect, on the highway use of motor fuel, of forces other than the number of motor vehicles registered. It appears that the major portion of the variation in highway use of motor fuel is traceable to the variations in the number of motor vehicles registered, but this factor does not explain quite all of the variation.

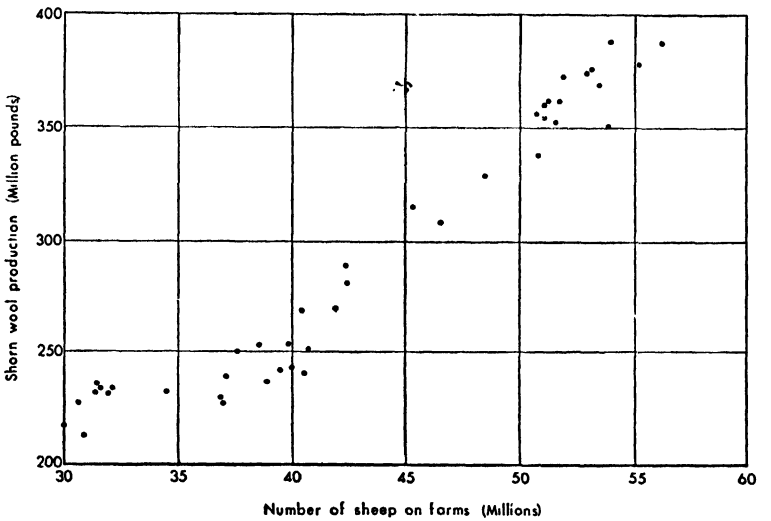
NONLINEAR RELATIONSHIPS

Up to this point the scatter diagrams have presented relationships that can be described by a straight line, but the situation can easily arise in which the line of average relationship is not best described by a straight line. The straight line relates the same change in the Y variable to a given change in the X variable over all of the values of the variables. When this is not the case, the line of average relationship should be represented by a curve instead of by a straight line.

Many studies have been made of the relationship of prices of farm products to the supply reaching the market, and the line of average relationship is often nonlinear. Chart 17-8 shows the relationship between the number of sheep on farms at the beginning of a year and the production of wool during the year. Since the number of sheep on farms is known early in the year, this relationship can be used to estimate the production of wool. However, the relationship between the number of sheep on farms and wool production is not represented by a straight line. When the number of sheep is less than 40 million, the relationship of wool production to the number of sheep generally tends to change. A more accurate description of the relationship between the two variables can be secured if a curve, rather than a straight line, is used to describe this relationship.

Chart 17-8

SCATTER DIAGRAM OF SHORN WOOL PRODUCTION AND NUMBER OF SHEEP ON FARMS AT BEGINNING OF YEAR 1915-1956



Source of data: *Agricultural Statistics, 1952*
Livestock and Meat Situation, May 1957

CROSS-CLASSIFICATION TABLES

When the number of items in a series is extremely large, the construction of a scatter diagram requires more work than is necessary. A *cross-classification table*, or *tally sheet*, can serve a similar purpose. Table 17-2 shows a cross classification of the consumption of natural gas per customer and the average price paid in 383 Texas cities.

Table 17-2

CROSS TABULATION OF CONSUMPTION OF NATURAL GAS PER DOMESTIC CUSTOMER AND AVERAGE PRICE PAID IN 383 TEXAS CITIES, 1952

PRICE (DOLLARS PER THOUSAND CUBIC FEET)

CONSUMPTION PER CUSTOMER (THOUSAND CUBIC FEET)	0 - .09	.10- .19	.20- .29	.30- .39	.40- .49	.50- .59	.60- .69	.70- .79	.80- .89	.90- .99	1.00- 1.09	1.10- 1.19	1.20- 1.29	1.30- 1.39	1.40- 1.49	1.50- 1.59
180-189					1											
170-179																
160-169						1										
150-159																
140-149																
130-139				2	2	2										
120-129						3										
110-119				1	1	6 1	1									
100-109					1	4		1								
90-99						6 1	3	2								
80-89					2	6 1	4	5								
70-79				1	1	6 1	7	9								
60-69					1	6 2	13	8	12		1					
50-59						4	8	12	14	4	1					
40-49						3	2	6	11	9	3			1		
30-39						2	3	1	7 1	2	6	1				
20-29						1				1						1
10-19									1							

SOURCE: Sixty-First Annual Report of the Railroad Commission of Texas, Gas Utilities Division, 1952.

The figures on price are divided into the different groups as shown in the captions, and the groups into which the consumption per customer is to be classified are shown in the stubs. The meaning of the tally marks in each square of the table is determined by reference to the stubs and captions of the table. Each tally mark represents a city, just as each plotted point in the scatter diagram in Chart 17-2 (page 413) represents a company. The total of the tally marks in a column represents the number of cities in which the average price paid for gas falls within the limits of that column. The total of the tally marks in a given square represents the number of cities

with consumption per customer as shown in the stub, and price as shown in the caption.

The construction of a cross-classification tally sheet follows the principles discussed in Chapters 3 and 5, except that the tally sheets described previously classified the data with respect to only one characteristic. The concentration of tally marks serves the same purpose as the plotted points in the scatter diagram; and at the same time the tally sheet can be used in computing measures of correlation, which are discussed in the following chapter. The more nearly the tally marks fall on a straight line, the higher the degree of correlation between the variables being compared. The plotted points in a scatter diagram are located precisely on the chart, but the tally marks are located within the proper square only. The particular location of a mark within its proper square is of no significance, and to that extent the tally sheet is an approximation. The chief advantage of the tally sheet is the saving of time in tabulating, instead of plotting each point accurately on the chart. The savings that can be made in calculations are discussed in Chapter 18.

CORRELATION AND CAUSE-AND-EFFECT RELATIONSHIPS

The construction of a scatter diagram that shows the relationship between two variables is a graphic method of determining the degree of correlation that exists, as well as the nature of correlation. In common with all graphic methods of analysis, the scatter diagram is not so definite as mathematical methods, although it is usually more simple. Graphic analysis is usually not only more simple to perform but also it is frequently easier to understand. Chapter 18 will be devoted to the simpler methods used in the measurement of correlation between two variables and Chapter 19 will describe the methods of measuring the correlation between three or more variables.

Before turning to the precise measurement of the degree of correlation, the importance of measures of relationship will be discussed. It should always be kept in mind that correlation analysis, both graphic and mathematical, merely shows the degree of relationship between the two variables. None of the measures of correlation give any clue as to the cause-and-effect relationships that are found in the series.

There are at least four possibilities with regard to cause-and-effect relationships.

1. Two series may show a high degree of correlation, yet this close relationship may result purely from chance. It may have just happened that the

two series fluctuate together. The fact that the price of gas and the consumption per customer in the 383 Texas cities showed an inverse relationship (Table 17-2) may be merely a coincidence.

2. A second possibility is that the fluctuations in one series may cause the fluctuations in the other. It may be true that the price of gas is lower in cities with large consumption because the lower price is made possible because of the larger volume.

3. Another possibility is the interchanging of the cause-and-effect relationship, as set up in No. 2 above. In other words, it may be that consumption is high in certain cities because the price is low, and is less in other cities because the price is high.

4. A fourth possibility is that the fluctuations in both series are common results of some other cause. In the study of the business cycle, it is commonly found that business conditions within two industries fluctuate together because they are both affected by the same underlying economic factors. In this situation the changes in one industry are not the cause of the change in the other.

In price analysis it is usually decided that price has an influence on the amount consumed, and, at the same time, the amount consumed is one of the factors affecting price. In the analysis of the demand for motor fuel, it is probably true that the number of motor vehicles registered is the cause of the variation in the amount of motor fuel consumed, rather than the reverse. But correlation analysis does not give any evidence as to which of the four possibilities stated above applies in a given case. The cause-and-effect relationships involved must be determined by a process of reasoning; both the correlation chart and the more elaborate methods of mathematical analysis are completely noncommital on the subject of cause-and-effect relationships.

In spite of the difficulty of determining which of the variables is the cause of change in the other, it is customary to call the X variable the *independent variable*, and the Y variable, the *dependent variable*. When it is not clear which is the cause, or when neither can be considered the causal factor, this terminology does not seem to be particularly appropriate; but it is usually used. When it seems likely that one variable is the cause of the variations in the other, the causal factor should be designated as the X variable. When one variable is to be forecast or estimated from the other, the variable to be estimated is designated as the dependent variable, even though both variables are known to be common causes of another factor. The decision to consider one variable as the dependent variable is important in interpreting the results of the analysis, since the designation of the variable is based on the type of causal relationship that is assumed to exist.

STUDY QUESTIONS

1. Outline the advantages and disadvantages of using the line chart and the scatter diagram as a measure of relationship.

2. How does the scatter diagram show the degree of relationship between the two variables plotted?

3. When more than two variables are believed to be related, it is sometimes possible to remove the effect of one variable by expressing the relationship between two of the variables as a ratio. This will permit measuring the relationship between the ratio and a third variable. Give an example of this type of analysis, in addition to those given in the text.

4. Does "measuring the relationship between two variables" mean that one of the variables *causes* the variations in the other? Explain exactly what is meant by this phrase.

5. How is the line of average relationship used to show the relationship between two variables?

6. How can the line of average relationship be used to estimate the values of one variable from the known values of the other?

7. What is meant by saying that the relationship between two variables is "nonlinear"? Give an example of nonlinear relationship that is not described in the text.

8. The cross-classification table is presented as a method of showing the relationship between two variables. What are its advantages and disadvantages in comparison with the scatter diagram?

9. How can a scatter diagram be used to compare the cyclical pattern in two series such as farm income and the sale of agricultural machinery? Would a scatter diagram based on these two series be of any help in forecasting the sales of a manufacturer of agricultural machinery?

10. A scatter diagram shows a high correlation between consumer income by counties and the sales of a given product. How can this information be used in planning the sales promotion of the product?

11. In showing the correlation between two variables on a scatter diagram, how is the decision made as to which variable should be plotted on the X axis if there is no clear indication of how they are related causally?

12. How could you use a scatter diagram to determine the closeness of the correspondence between the seasonal patterns of two time series? Could this degree of correlation be shown better by plotting the two series on a time series chart?

13. Many statements have been made to the effect that correlation analysis is not valid in comparing the fluctuations in two time series. Under what circumstances do you think it particularly undesirable to measure the degree of correlation between two time series?

PROBLEMS

1. The following table shows the condition of the corn crop on September 1 of each year, expressed as a percentage of normal, and the final yield for each year from 1944 to 1953. Plot this data on a scatter diagram and state how closely you think the September 1 report on condition of the crop is related to the final yield per acre. (*Retain this solution for use in Problem 1 of Chapter 18, and Problem 1 of Chapter 19.*)

YIELDS OF CORN PER ACRE IN MICHIGAN, AND THE CONDITION OF THE CROP REPORTED ON SEPTEMBER 1: 1944 TO 1953

Year	Yield per acre (Bushels) *Y*	Condition (Percentage of normal) *X*
1944.....	32.0	68
1945.....	35.0	77
1946.....	28.0	64
1947.....	27.5	60
1948.....	39.5	84
1949.....	49.0	99
1950.....	38.5	87
1951.....	41.5	85
1952.....	50.0	94
1953.....	45.5	89

SOURCE: *Graphic Analysis in Economic Research*, United States Department of Agriculture, June, 1955.

2. The following table gives the per capita consumption of meat in the United States since the end of World War II, and two other series that may be related to the changes in meat consumption. Deflated per capita disposable income was derived from data on disposable income (income received by persons less direct taxes paid), population, and the level of consumer prices. Deflated meat prices show the relationship of meat prices to the prices of all consumer goods. This problem is to show graphically the relationship between per capita meat consumption and the other two series: one representing income of consumers, and the other, the price of meat.

(a) Construct two scatter diagrams, one showing the relationship of per capita meat consumption to deflated per capita disposable income, and the other showing the relationship of per capita meat consumption to deflated meat prices.

(b) Summarize briefly what these two scatter diagrams show with respect to the fluctuations in meat consumption.

(c) Which series appears to show the higher correlation with per capita meat consumption?

(d) *Retain this solution for use in Problems 2 and 3 of Chapter 18 and Problem 3 of Chapter 19.*

PER CAPITA CONSUMPTION OF MEAT, DEFLATED PER CAPITA DISPOSABLE INCOME, AND DEFLATED MEAT PRICES: 1946 TO 1956

Year	Per capita meat consumption (Pounds)[1] X_1	Deflated per capita disposable income (1947 dollars)[2] X_2	Deflated meat prices (1947–49 = 100)[3] X_3
1946.....	152.1	1,247	83.2
1947.....	153.1	1,173	97.9
1948.....	143.5	1,211	103.2
1949.....	142.7	1,203	98.7
1950.....	142.6	1,280	102.0
1951.....	136.2	1,290	105.6
1952.....	144.3	1,306	102.4
1953.....	153.7	1,343	96.1
1954.....	153.3	1,332	94 1
1955.....	162.8	1,384	88.7
1956.....	166.8	1,423	83.6

SOURCE: [1]*Agricultural Statistics.*
[2]*Survey of Current Business*, July, 1957, and *Statistical Abstract of the United States, 1957.*
[3]*Business Statistics, 1957.*

3. The following table, giving the results of a study by the Automobile Manufacturers Association, shows the percentage of persons in selected cities that use a motor vehicle as a method of travel. Two additional series are given, which will be used in an attempt to explain the variation in the percentage of persons using an automobile. Variable X_2 represents the percentage of incomes that were over $2,000 in the urbanized area of each city. This series may be used to test the hypothesis that families with very small incomes cannot afford an automobile. The concentration of population may also be a factor entering into the use of an automobile, so variable X_3, the population per square mile for each urbanized area, is also given.

(a) Plot a scatter diagram showing the relationship between the percentage of persons using a motor vehicle as a method of travel and the percentage of incomes over $2,000. What conclusions can you draw from this scatter diagram concerning the relationship between these two variables?

(b) Plot a scatter diagram showing the relationship between the percentage of persons using a motor vehicle as a method of travel and the population per square mile. Does this diagram indicate that there is more or less correlation than in (a)?

(c) *Retain this solution for use in Problems 4 and 5 of Chapter 18 and Problem 4 of Chapter 19.*

PERCENTAGE OF PERSONS USING A MOTOR VEHICLE AS METHOD OF TRAVEL, PERCENTAGE OF INCOMES OVER $2,000, AND POPULATION PER SQUARE MILE FOR 26 CITIES IN THE UNITED STATES

City	Percentage of persons using a motor vehicle[1] X_1	Percentage of incomes over $2,000 in urbanized area[2] X_2	Population per square mile in urbanized area[2] X_3
Allentown, Pa.	67.9	73.0	4,583
Baltimore, Md.	46.2	69.0	7,654
Charleston, S. C.	59.8	47.2	6,537
Columbus, Ga.	59.5	56.8	3,376
Duluth, Minn.	71.4	68.9	1,362
Erie, Pa.	77.1	74.5	5,091
Grand Rapids, Mich.	87.1	73.1	4,857
Indianapolis, Ind.	65.1	72.9	5,545
Kalamazoo, Mich.	82.3	71.7*	3,931
Lansing, Mich.	86.7	70.2	4,352
Macon, Ga.	54.7	54.5*	4,250
Newark, N. J.	50.5	73.2**	9,810*
Phoenix, Ariz.	83.8	73.3*	3,931
Portland, Ore.	72.1	68.4*	4,517
Reading, Pa.	55.0	71.5	5,869
Sacramento, Cal.	84.4	72.3	5,091
Saginaw, Mich.	91.4	73.8	4,414
Salt Lake City, Utah	74.7	73.6	2,988
San Francisco, Cal.	66.4	70.7	7,038
Seattle, Wash.	65.4	70.0	5,057
South Bend, Ind.	82.1	74.2	4,521
Spokane, Wash.	69.7	69.1	3,444
Tacoma, Wash.	74.9	69.4	2,696
Washington, D. C.	60.7	74.3	7,216
Wilmington, Del.	65.5	72.1	4,012
Worcester, Mass.	65.7	66.5	5,031

*Metropolitan area
**City
SOURCE: [1]*Automobile Facts and Figures*, 1951.
[2]*Census of Population, 1950.*

4. The administrative officers of a governmental agency employing a large number of field workers have been giving serious consideration to the question of the influence of a worker's age on his efficiency. At least three views are held by different individuals responsible for the administration of the agency. One person holds that the young field workers are the best, while another states that the older men are more satisfactory. A third viewpoint is that there are good men among both groups, and that age is not a significant factor in the performance of the field force.

Assume that you are instructed to present to the proper officials an analysis that will give information on the subject of the relationship of age and efficiency. You discover that previously each supervisor of the field workers had been asked to rate each individual under his supervision. The ratings were made as carefully as it was possible to make them, and you believe that they represent rather accurately the performance of the employees. The rating and the age for each of the 237 employees in the field force are given in the following table.

(a) Construct a cross-classification table from the data on ages and efficiency ratings.

(b) What does the table indicate is the relationship between efficiency and age in the field force of this agency?

(c) *Retain this solution for use in Problem 6 of Chapter 18 and Problem 2 of Chapter 19.*

AGE AND RATING OF 237 EMPLOYEES

Empl. no.	Age (Yrs.)	Rating (%)	Empl. no.	Age (Yrs.)	Rating (%)	Empl. no.	Age (Yrs.)	Rating (%)
1	67	75	81	62	71	159	60	77
3	34	79	83	69	80	160	34	82
6	36	78	84	50	73	161	30	84
7	48	62	85	28	83	162	42	74
12	49	77	86	44	80	163	51	53
13	31	72	87	43	60	165	43	83
14	61	89	90	60	72	167	43	75
15	34	63	91	39	67	168	27	91
16	43	67	92	47	66	169	33	75
21	45	78	95	42	72	170	56	79
22	38	76	99	49	58	176	33	69
23	32	69	101	44	90	181	50	71
24	27	79	102	37	92	183	44	88
28	61	81	103	30	72	184	29	91
34	29	69	107	42	76	189	39	78
35	21	85	108	42	78	191	34	91
38	36	76	109	31	73	193	39	82
39	50	59	111	34	84	195	33	90
40	46	81	113	24	82	196	35	89
41	30	80	114	38	68	200	33	86
42	46	47	123	40	79	201	43	78
44	32	59	124	41	73	204	34	82
45	30	88	125	46	70	206	30	66
47	33	77	126	43	40	207	44	62
48	45	90	127	48	65	211	40	68
50	49	81	128	34	78	212	32	74
52	48	63	131	37	60	213	51	82
53	41	81	132	26	90	215	35	73
56	53	95	133	59	69	220	34	71
57	36	80	135	53	84	221	33	41
66	46	93	140	36	70	222	39	66
67	50	83	144	61	58	223	37	66
70	28	75	147	46	77	224	29	47
71	35	86	148	41	76	225	37	56
73	35	79	150	29	88	232	26	88
74	38	82	151	27	86	235	51	66
75	36	72	152	58	61	236	45	69
76	46	73	154	41	85	237	37	83
79	43	69	157	50	82	239	31	70
80	34	81	158	55	55	241	42	75

AGE AND RATING OF 237 EMPLOYEES *(Continued)*

Empl. no.	Age (Yrs.)	Rating (%)	Empl. no.	Age (Yrs.)	Rating (%)	Empl. no.	Age (Yrs.)	Rating (%)
243	39	83	362	30	59	422	39	37
250	44	89	363	53	86	424	32	60
251	37	69	365	33	70	426	24	51
254	55	68	366	31	72	432	34	76
257	44	70	367	40	56	446	23	91
260	47	74	368	47	73	447	46	77
261	31	78	369	35	67	448	27	82
264	32	76	370	42	80	449	32	81
265	37	81	372	37	82	450	56	42
266	30	80	373	35	55	451	30	85
267	32	75	375	41	80	452	24	90
268	30	85	376	30	84	454	61	41
272	42	96	378	40	88	455	55	43
273	56	71	380	41	77	456	60	29
275	32	72	382	30	77	458	24	88
278	56	75	383	39	68	459	29	80
280	35	88	384	26	75	460	35	72
284	30	80	286	26	79	461	42	71
287	27	55	387	33	81	463	55	80
301	38	83	389	28	85	464	21	90
303	29	67	390	30	70	465	36	72
305	34	87	392	31	75	467	29	80
308	42	85	393	43	72	468	48	51
313	35	80	395	43	81	470	51	62
317	42	63	396	26	62	472	37	49
322	42	85	397	28	76	473	36	80
324	41	88	401	44	72	474	24	75
327	64	67	402	26	81	475	42	61
340	26	75	403	28	79	476	51	59
342	31	78	411	38	77	477	38	80
343	50	61	412	25	72	478	31	83
344	29	62	414	34	79	479	35	77
345	56	87	415	26	45	480	42	60
351	41	71	416	33	69	481	38	90
354	31	74	417	37	59	484	25	42
355	27	74	418	33	72	486	42	39
356	29	83	419	37	89	487	39	63
358	30	71	420	26	81	489	44	62
359	37	56	421	39	70	491	52	59

SOURCE: Records of the agency.

18 MEASUREMENT OF SIMPLE LINEAR CORRELATION

Measures of correlation have as their objective the expression in definite, quantitative terms of the relationship existing between variables. It can be seen from Chart 17-4 (page 416) that the consumption of heating oil in the United States is related to the number of domestic burners in use; the next step is measuring as precisely as possible how the two are related. In other words, for a given increase in the number of burners in use how much increase can be expected in the amount of heating oil consumed? The scatter diagram in Chart 17-4 gives an approximate picture of the relationship between these variables. This chapter will develop the mathematical measurement of the relationship between two variables when the relationship is represented by a straight line. Chapter 19 will describe methods of measuring the relationship between more than two variables, and between two variables when the relationship is best described by a curve (nonlinear).

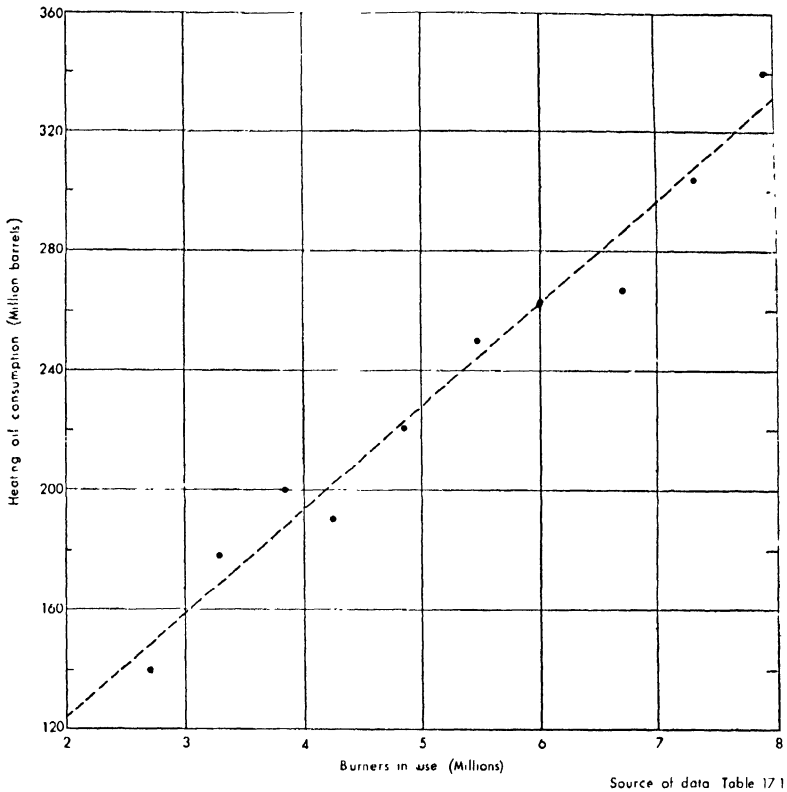
THE LINEAR REGRESSION EQUATION

In Chart 17-7 (page 418) a line was used to show the average relationship and the nature or shape of the relationship between the two variables; the equation of this line is referred to as the *regression equation*. The principles underlying this measure of relationship are the same as those used in the measurement of secular trend. A trend line measures the average change in the series over a period of time, while a regression line measures the changes in one variable associated with changes in the other. In Chart 17-7 the regression line measures the changes in the highway use of motor fuel that are associated with variations in the number of motor vehicles registered in the different states. Chart 18-1 shows the relation of heating oil consumption to the number of domestic burners in use between 1946 and 1955. In both of these charts the relationship can be described by a straight line.

The line showing the average relationship between two variables may be located graphically, or may be computed by the method of least squares. When a straight line is desired to be used to describe the relationship between two variables, it is common practice to fit the line by the method of least squares. When the relationship is nonlinear, a curve may be

Chart 18-1

SCATTER DIAGRAM OF HEATING OIL CONSUMPTION AND NUMBER OF DOMESTIC BURNERS IN USE, 1946-1955



Source of data Table 17.1

fitted by the method of least squares, or the graphic method of locating the line may be used. The fitting of a least squares straight line is illustrated below, and the use of a nonlinear regression line is shown on page 453 of the next chapter.

The formula for a straight line was given in Chapter 12 as: $Y_c = a + bX$. Fitting the straight line to the figures on which the scatter diagram (Chart 18-1) is based consists of finding the a and b values for the line that best fits the data. When these values are determined, the equation expresses the relationship between heating oil consumption and the number of domestic burners in use. This is done by substituting in the following equations for a and b.

$$a = \frac{\Sigma X^2 \Sigma Y - \Sigma X \cdot \Sigma XY}{N \cdot \Sigma X^2 - (\Sigma X)^2} \quad \text{(51)}$$

$$b = \frac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{N \cdot \Sigma X^2 - (\Sigma X)^2} \quad \text{(52)}$$

The summations in Table 18-1 are as follows:

$$\Sigma X = 52.14 \qquad \Sigma X^2 = 299.7874$$

$$\Sigma Y = 2{,}353 \qquad \Sigma Y^2 = 587{,}199$$

$$\Sigma XY = 13{,}221.34 \qquad N = 10$$

When the above values are substituted in the equations, the values of a and b are found.

$$a = \frac{(299.7874)\,(2{,}353) - (52.14)\,(13{,}221.34)}{(10)\,(299.7874) - (52.14)^2} = \frac{16{,}039.0846}{279.2944} = 57.42716$$

$$b = \frac{(10)\,(13{,}221.34) - (52.14)\,(2{,}353)}{(10)\,(299.7874) - (52.14)^2} = \frac{9{,}527.98}{279.2944} = 34.11447$$

The regression for the data plotted in Chart 18-1 is, therefore,

$$Y_c = 57.43 + 34.11X$$

Table 18-1

COMPUTATION OF THE LINEAR REGRESSION EQUATION

HEATING OIL CONSUMPTION AND DOMESTIC BURNERS IN USE: 1946 TO 1955

Year	Heating oil consumption (Million barrels) Y	Domestic burners in use (Millions) X	Y^2	X^2	XY
1946...	140	2.67	19,600	7.1289	373.80
1947...	178	3.24	31,684	10.4976	576.72
1948...	200	3.81	40,000	14.5161	762.00
1949...	190	4.23	36,100	17.8929	803.70
1950...	221	4.83	48,841	23.3289	1,067.43
1951...	250	5.44	62,500	29.5936	1,360.00
1952...	263	6.03	69,169	36.3609	1,585.89
1953...	267	6.68	71,289	44.6224	1,783.56
1954...	304	7.31	92,416	53.4361	2,222.24
1955...	340	7.90	115,600	62.4100	2,686.00
Total...	2,353	52.14	587,199	299.7874	13,221.34

SOURCE: Table 17-1.

ESTIMATES

The chief reason for expressing the relationship between two variables as an equation is to be able to estimate one variable when the other is known. When the relationship between heating oil consumption and domestic burners in use is known, it is possible to estimate the probable consumption of heating oil in a given year on the basis of the number of burners in use.

In the regression equation, b equals 34.11. The consumption of heating oil increases 34.11 million barrels for each increase in one million domestic burners. The b coefficient is the more significant of the two coefficients in the regression equation, since b measures the change in X and the change in Y. It shows the change in the Y variable that, on the average, is associated with changes in the X variable. The a coefficient, which merely serves to locate the level of the regression line, is called the Y intercept, and is the Y value when X equals zero.

Assume that the sales of domestic oil burners indicate that the average number in use during the coming year will be 9.00 million. Using the regression equation, it is possible to compute the probable consumption of heating oil for that year. In computing the equation in Table 18-1, the number of domestic burners is the X variable, in millions of burners, and heating oil consumption is the Y variable, in millions of barrels. The estimated consumption of heating oil for the coming year is, therefore, 364 million barrels, computed as follows:

$$Y_c = 57.43 + 34.11X$$

$$\text{When } X = 9.00,\ Y_c = 57.43 + (34.11)\ (9.00)$$

$$Y_c = 57.43 + 306.99$$

$$Y_c = 364.4 \text{ million barrels}$$

This equation may be used to estimate heating oil consumption as long as the equation continues to describe the relationship between the number of burners and consumption. If something should change the average rate of consumption of oil by domestic burners, it would be necessary to modify the equation to take this into account. It is always important to avoid using such an estimating equation uncritically or mechanically, since obviously it is a valid estimating device only so long as it continues to measure the relationship between the variables.

The chart on which the regression line is plotted might be used to compute the consumption of heating oil, as illustrated on page 418 of Chapter 17. The degree of accuracy with which graphic estimates may be read from the chart depends upon the size of the graph, but such estimates are usually accurate enough for most uses.

RELIABILITY OF ESTIMATES

When the relationship between two variables is known, it is always possible to compute one if the value of the other is known. The question that immediately arises, however, is: How accurate is the computed value? It can be seen from Chart 18-1 that there is a tendency for the plotted points to fall near the regression line, but very few fall exactly on the line. This means that any estimate based on the regression line will be only approximate. The more closely the X and Y variables are related, the more accurate the estimates are likely to be; but as long as there is scatter about the line, an estimate is not completely accurate.

If the regression equation is used to make estimates, it is desirable to have a measure of the accuracy of these estimates. It can be seen from looking at Charts 17-1 to 17-8 that there is a wide difference in the degree of relationship between different pairs of variables. The more closely the plotted points are grouped around the regression line, the more accurate the estimates based on the regression line are likely to be. Likewise, the more the points scatter, the less reliable an estimate based on the regression line is likely to be. What is needed is a measure of the degree of scatter of the individual items about the regression line.

Table 18-2 shows the Y_c values, computed by using the regression equation. If the actual Y values were not known, these computed values, made from the regression equation, would be the best estimates. It can be seen, however, that the actual values differ in varying degrees from the estimates based on the regression equation. Column (5) shows the amount by which each of the estimates differs, or deviates, from the actual Y values.

Table 18-2 shows that in 1946, when there were 2.67 million domestic burners in use, the actual consumption of heating oil was 140 million barrels. If the actual consumption of heating oil were not known, the estimate of 148.5 million barrels, given in Column (4), would be the most probable value. The difference between the actual consumption and the best estimated consumption is -8.5 million barrels. The amounts in Column (5) represent the deviation of the actual consumption from the estimated consumption of each year. These amounts are referred to as *residuals*, since the variations of the actual consumption from the estimated consumption are in effect what is left over after taking into account the variations in consumption related to the number of burners in use. Presumably these residuals represent the variations in consumption related to factors other than the number of burners in use.

Finding a measure of the scatter of the actual values about the regression line is a problem of averaging the residuals. In Chapter 9 the averaging of deviations from the mean was discussed. Finding an average of the

Table 18-2

COMPUTATION OF THE STANDARD ERROR OF ESTIMATE

HEATING OIL CONSUMPTION AND DOMESTIC BURNERS IN USE: 1946 TO 1955

Year (1)	Domestic burners in use (Millions) (2) X	Heating oil consumption (Million barrels) (3) Y	(4) Y_c	(5) $Y - Y_c$	(6) $(Y - Y_c)^2$
1946...	2.67	140	148.5	− 8.5	72.25
1947...	3.24	178	168.0	10.0	100.00
1948...	3.81	200	187.4	12.6	158.76
1949...	4.23	190	201.7	−11.7	136.89
1950...	4.83	221	222.2	− 1.2	1.44
1951...	5.44	250	243.0	7.0	49.00
1952...	6.03	263	263.1	− .1	.01
1953...	6.68	267	285.3	−18.3	334.89
1954...	7.31	304	306.8	− 2.8	7.84
1955...	7.96	340	327.0	13.0	169.00
Total...	52.14	2,353	2,353.0	0	1,030.08

SOURCE: Table 17-1.

residuals is essentially the same problem. The sum of the residuals is zero, just as the deviations of the individual items from their mean total zero. The residuals are squared, summed, divided by the number of items, and the square root extracted, just as in the computation of the standard deviation. In other words, the arithmetic mean of the residuals squared is computed, and the square root extracted. This measure is called the *standard error of estimate*, and measures the average amount that the computed values vary from the actual values.

$$S_{y} = \sqrt{\frac{\Sigma (Y - Y_c)^2}{N}} \quad \ldots\ldots\ldots\ldots \quad (53)$$

This measure may be used as an indication of the amount of error to expect in any individual estimate. It can be seen from Column (5) in Table 18-2 that the errors range from .1 to 18.3. The standard error of estimate, which is the average error in the estimate, is 10.2 million barrels.

$$S_{y} = \sqrt{\frac{1,030.08}{10}} = \sqrt{103.008} = 10.2$$

Sometimes the error in the individual estimates will be more than 10.2 and sometimes less. If the errors form a normal distribution, 68 per cent

of the errors in estimates based on the regression equation will be less than the standard error of estimate. In other words, the assumption can be made that 68 per cent (or more precisely, 68.27 per cent) of the estimates based on the regression equation will come within 10.2 million barrels of the actual consumption.

Using the regression equation, the estimate was previously made that the consumption of heating oil for the coming year would be 364.4 million barrels. The question "How good is this estimate?" may very well be asked at this point. Since the standard error of estimate has been calculated as 10.2, the chances are about two out of three (or 68.27 out of 100) that the error in the consumption estimate was not more than 10.2 million barrels. In other words, the chances are about two out of three that the amount of heating oil consumed in the coming year would not vary more than 10.2 million barrels above or below 364.4 million barrels. This means that the chances are about two out of three that the actual consumption would fall between 354.2 and 374.6 barrels. These probabilities are based on the areas under the normal curve, discussed in Chapter 9.

Whenever the regression equation is used to estimate one variable from another, it is important that the reliability of the estimate be determined. The standard error of estimate just described is the measure commonly used to indicate the amount of error likely to be found. In reality it measures the average amount of error in the estimates computed for the values already known. If the relationship between the two variables remains the same, it can be used to measure the reliability of estimates based on the measure of relationship. It should be observed that this last assumption is vital to the whole procedure described above.

SHORTER METHOD OF COMPUTING S_y

It is possible to use the summations already made to compute the standard error of estimate, instead of going through the long process shown in Table 18-2. The following formula may be used. It will be observed that this formula uses values already available.

$$S_y = \sqrt{\frac{\Sigma Y^2 - a\,\Sigma Y - b\,\Sigma XY}{N}} \quad \ldots\ldots\ldots \quad (54)$$

The following values were computed in Table 18-1 (page 432).

$$\Sigma Y = 2{,}353$$
$$\Sigma Y^2 = 587{,}199$$
$$\Sigma XY = 13{,}221.34$$

The values of the constants in the regression equation were found to be:

$$a = 57.4271615$$
$$b = 34.114468$$

Substituting the above values, S_y is found to be 10.2, the same amount secured by formula (53).

$$S_y = \sqrt{\frac{587{,}199 - (57.4271615)\,(2{,}353) - (34.114468)\,(13{,}221.34)}{10}}$$

$$S_y = \sqrt{\frac{1{,}033.90864}{10}} = \sqrt{103.3909} = 10.2$$

It will be noted that the values of a and b are carried out to a large number of decimals for use in the calculation of S_y. This is necessary to avoid introducing any rounding error into the calculation. The final figures are relatively small in comparison with the figures used in the calculations; and if the latter figures were rounded as much as is common in most calculations, the resulting error would be unduly large.

ABSTRACT MEASURES OF RELATIONSHIP

It has been shown on the preceding pages that the regression equation makes it possible to estimate the value of one variable when the other is known, and that the standard error of estimate serves as a measure of the reliability of these estimates. However, there are occasions when it is desirable to have a measure of the extent to which the fluctuations in two variables are related and the extent to which they are not related.

The scatter diagram shown in Chart 18-1 (page 431) shows how closely the consumption of heating oil parallels the number of domestic burners in use. The previous sections of this chapter have described a method of estimating the consumption of fuel oil when the number of burners in use was known, but this section will present a method of computing an abstract measure of the degree of relationship between the two variables.

The size of the standard error of estimate is a measure of the degree of relationship between two variables, since the larger the scatter of the individual values about the regression line, the lower the correlation. The difficulty with using the standard error of estimate as a measure of correlation is the fact that it is expressed in concrete units that vary with each situation. The standard error of estimate for the data plotted in Chart 18-1 is 10.2 million barrels. Whether this is a large or a small scatter depends upon its relative size.

The standard error of estimate serves as the basis for several measures of correlation, all of which indicate the degree of relationship between two variables. Since the standard error of estimate computed in Table 18-2 measures the variation in heating oil consumption that is not related to the number of burners in use, a satisfactory measure of correlation can be computed if it is possible to determine what proportion this represents of the total variation in heating oil consumption. Obviously if a determination can be made of the proportion of the total variation that is *not* related to the independent variable, it will also show the proportion that *is* related.

A measure of correlation is computed by using the square of the standard deviation of the Y variable as a measure of the total variation in that variable. In Chapter 9 the standard deviation was described as the best-known measure of dispersion, but the square of the standard deviation, known as the *variance*, may also be used as a measure of dispersion, particularly in further calculations The variance, which is the average of the squared deviations from the arithmetic mean, is as valid a method of measuring dispersion as an average of the actual deviations. The standard deviation has the practical advantage of giving a measure of dispersion in original units which reflects the average size of the deviation of the items. The variance, being the average of the *squared deviations from the mean*, does not have the same practical significance as the standard deviation, but in some ways its mathematical characteristics make it more useful. The importance of this feature is pointed out in the following discussion.

By relating the square of the standard error of estimate to the square of the standard deviation, it is possible to determine the proportion of the total variance that is *not* related to the changes in the X variable. Chart 18-1 shows that an increase in heating oil consumption is associated with an increase in the number of burners in use, but in none of the years is the variation in consumption related exactly to the number of burners. From Table 18-2 (page 435) it can be seen that in 1952, consumption was only .1 million barrels less than would have been expected on the basis of the number of burners in use. However, in 1953, consumption was 18.3 million barrels less than the number of burners in use would indicate would be consumed. This variation of 18.3 million barrels in heating oil consumption is not explained by the number of burners in use. It appears that some other factor must be causing some variations in consumption.

The standard deviation of heating oil consumption represents the variation in consumption due to all causes that might affect the amount of heating oil used. The square of the standard deviation, or the variance, also represents the variation in heating oil consumption. Using equations (14) to (16) from Chapter 9, (substituting Y for X), the square of the standard deviation of heating oil consumption (σ_y^2) is computed as follows:

$$\sigma_y^2 = s_y^2 - c_y^2$$

$$s_y^2 = \frac{\Sigma Y^2}{N} = \frac{587{,}199}{10} = 58{,}719.9$$

$$c_y^2 = \left(\frac{\Sigma Y}{N}\right)^2 = \left(\frac{2{,}353}{10}\right)^2 = 55{,}366.09$$

$$\sigma_y^2 = 58{,}719.9 - 55{,}366.09 = 3{,}353.81$$

The square of the standard error of estimate, computed on page 435, is 103.008. This amount measures the variance about the regression line, or the variance that is not related to the number of burners in use. The *coefficient of nondetermination*, represented by k^2, can be computed from the unexplained variance by expressing the unexplained variance as a fraction of the total variance of 3,353.81 computed above.

$$k^2 = \frac{S_y^2}{\sigma_y^2} \quad \text{(55)}$$

$$k^2 = \frac{103.008}{3{,}353.81} = .031$$

The *coefficient of determination*, defined as $r^2 = 1 - k^2$, represents the fraction of the total variance that is explained by the number of burners in use.

$$r^2 = 1 - .031 = .969$$

The value of r^2 may be computed directly from the values of S_y^2 and σ_y^2 by the following equation:

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad \text{(56)}$$

This coefficient may be interpreted as meaning that 96.9 per cent of the variance in heating oil consumption is related to the number of burners in use.

Two other measures of correlation may be derived by taking the square roots of the coefficient of determination and the coefficient of nondetermination. These measures are known as the *coefficient of correlation* (r) and the *coefficient of alienation* (k).

$$r = \sqrt{.969} = .984$$

$$k = \sqrt{.031} = .176$$

The sum of r and k never equals 1.00 unless one of the two is 1.00 and the other zero. In any other circumstances $r + k > 1.00$.

The nearer r is to 1.00, the higher the degree of correlation; the smaller r is, the less the correlation. When r equals zero (which means k equals 1.00) there is no correlation.

Since r is the square root of the coefficient of determination, its sign is either plus or minus. When the b value in the regression equation is plus, the correlation is said to be positive. In other words, small X values are associated with small Y values, and large X and large Y values occur together. When the b value is negative, the correlation is said to be negative, and large values of one variable are associated with small values of the other. The coefficient of correlation is given the sign of the b value in the regression equation and in this manner shows whether the correlation is positive or negative.

AN ALTERNATIVE METHOD OF COMPUTATION (PRODUCT-MOMENT)

Other formulas than the ones just given have been derived for the computation of the various measures of correlation. It will be observed that the method used in the preceding pages secured the values of the coefficients, or abstract measures of relationship, as the last step in the analysis. There may be occasions when the coefficient of determination is wanted first, or is the only measure wanted. If this is the case, the following formula[1] gives the value of r directly from the values computed in Table 18-1 (page 432). If the value of r^2 is wanted, it can be found by squaring r or by removing the radical from the denominator and squaring the numerator before making the final division.

$$r = \frac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{[N \cdot \Sigma X^2 - (\Sigma X)^2]\,[N \cdot \Sigma Y^2 - (\Sigma Y)^2]}} \quad . \quad . \quad . \quad (57)$$

The summations from Table 18-1 are substituted in this formula, and the value of r computed. These computations are shown on the next page.

[1]This formula is particularly desirable for use on a calculator. Only the figures shown in the calculation above need be written down; all other operations can be performed on the machine. The multiplication of N and ΣXY is performed and the product is left in the dials. ΣX and ΣY are then multiplied with negative multiplication, which will subtract the product of ΣX and ΣY from the product of N and ΣXY that is already in the dials. The remaining number in the dials is the difference between the products and is the numerator of the equation

The two parts of the denominator should be computed in the same manner. N is multiplied by ΣX^2 and the product is left in the dials. ΣX is then squared using negative multiplication, which subtracts the square of ΣX from the product already in the dials. The result is copied, the machine cleared, and the same calculation made for the second term in the denominator. Care needs to be taken to be certain that the decimals are lined up correctly; but if this precaution is taken, the method is easy to carry out.

$$N = 10 \qquad \Sigma X^2 = 299.7874$$
$$\Sigma X = 52.14 \qquad \Sigma Y^2 = 587{,}199$$
$$\Sigma Y = 2{,}353 \qquad \Sigma XY = 13{,}221.34$$

$$r = \frac{(10)(13{,}221.34) - (52.14)(2{,}353)}{\sqrt{[(10)(299.7874) - (52.14)^2][(10)(587{,}199) - (2{,}353)^2]}}$$

$$r = \frac{9{,}527.98}{\sqrt{(279.2944)(335{,}381)}} = \frac{9{,}527.98}{9{,}678.33} = .98447$$

$$r^2 = .969$$

The calculation of the value of r from formula (57) provides all of the data needed for the calculation of b as performed on page 432. The numerator of the equation for b is the same as for r, and the denominator of the equation for b is the first term of the denominator of the equation for r.

The computation of the value of b consists of one division, as follows:

$$b = \frac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{N \cdot \Sigma X^2 - (\Sigma X)^2} = \frac{9{,}527.98}{279.2944} = 34.11447$$

The calculation of a can be performed by the following equation:

$$a = \overline{Y} - b\overline{X} \quad \text{. . . . . (58)}$$

The values of the two arithmetic means can be computed from the summations in Table 18-1 and the resulting values substituted in the equation for a.

$$\overline{X} = \frac{\Sigma X}{N} = \frac{52.14}{10} = 5.214$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{2{,}353}{10} = 235.3$$

$$a = 235.3 - (34.11447)(5.214) = 57.42716$$

The value of the standard error of estimate (S_y) can also be calculated from the data already assembled. The equation for the standard error of estimate is as follows:

$$S_y = \sigma_y k \quad \text{. . . . . (59)}$$

$$\text{or} \quad S_y = \sigma_y \sqrt{1 - r^2} \quad \text{. . . . . (60)}$$

The required value of σ_y can be derived from the values already computed, by writing the equation for σ_y in the following form:[2]

$$\sigma_y = \frac{\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}{N} = \frac{\sqrt{335{,}381}}{10} = \frac{579.1}{10} = 57.9$$

Using the value of k, calculated as .176 on page 439, the computation of the value of S_y is as follows:

$$S_y = \sigma_y k = (57.9)\,(.176) = 10.2$$

GROUPED DATA (PRODUCT-MOMENT METHOD)

When a large number of items is included in a correlation analysis, the work of making the summations is extremely heavy. Methods have been developed by which these summations can be made by electronic computers and tabulating machines, but these machines are not available in many businesses. The need for a method of computation that reduces the amount of clerical work has led to an extensive use of the product-moment method based on group data.

Table 17-2 (page 421) illustrates how a cross-classification table shows the relationship between two variables in the same manner as a scatter diagram. The method to be described here uses the cross-classification table instead of the original data for the computation of the various measures of correlation.

Table 18-3 is based on Table 17-2; the number of frequencies in each square of the table has been determined from Table 17-2. The total number of frequencies in each row and the total number in each column are next determined. The number of frequencies in each row is entered in the column with the caption f_y. There is one frequency in the row 10–19, three frequencies in the row 20–29, and so on. The total number of frequencies in each column is entered in the row with the stub f_x. There are four frequencies in the column .30–.39, nine frequencies in the column .40–49, and so on. The total of the f_x row is 383, and is entered at the right of the individual frequencies. The total of the f_y column is also 383. The totals of the frequencies in the rows and columns form two frequency distributions, each with 383 frequencies.

[2]It should be noted in this connection that the value of σ_x can be computed from the other term in the denominator by the following equation:

$$\sigma_x = \frac{\sqrt{N\Sigma X^2 - (\Sigma X)^2}}{N} \quad \text{(61)}$$

Table 18-3

A CORRELATION TABLE

X — AVERAGE PRICE PAID FOR GAS (DOLLARS PER THOUSAND CUBIC FEET)

Y — CONSUMPTION OF GAS PER CUSTOMER (THOUSANDS OF CUBIC FEET)

	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-.99	1.00-1.09	1.10-1.19	1.20-1.29	1.30-1.39	1.40-1.49	1.50-1.59	f_y	d'_y	$f_y d'_y$	$f_y(d'_y)^2$
180-189														1	12	12	144
170-179														0	11	0	0
160-169														1	10	10	100
150-159														0	9	0	0
140-149														0	8	0	0
130-139														6	7	42	294
120-199														3	6	18	108
110-119														10	5	50	250
100-109					1									6	4	24	96
90-99					2									11	3	33	99
80-89					4									16	2	32	64
70-79					24									49	1	49	49
60-69		1	7	38	46	35		1						128	0	0	0
50-59					27									83	-1	-83	83
40-49					6									17	-2	-88	176
30-39					1									21	-3	-63	189
20-29														3	-4	-12	48
10-19														1	-5	-5	25
f_x	4	9	52	76	111	92	24	10	1	0	1	0	1	383	—	19	1,725
d'_x	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	—	$\Sigma f d'_x d'_y = -742$		
$f_x d'_x$	-16	-27	-104	-76	0	92	52	30	4	0	6	0	8	-31			
$f_x(d'_x)^2$	64	81	208	76	0	92	104	90	16	0	36	0	64	831			

SOURCE: Table 17-2.

The two frequency distributions thus formed are analyzed in exactly the same manner as discussed on page 203 of Chapter 9 in the section dealing with the computation of the standard deviation by the short method. The column to the right of the f_y column is headed d'_y, the next column to the right is $f_y d'_y$, and the last column $f_y(d'_y)^2$. The only way in

which the captions of these columns are different from the columns in Table 9-5, page 204, is in the use of the subscript y. This symbol is necessary since there is another frequency distribution in the same table, designated by the subscript x. The subscript distinguishes the frequencies of the X distribution from those of the Y distribution.

The amounts in the $f_x d_x'$ column are computed by multiplying the amounts in the f_x column by the amounts in the d_x' column. The amounts in the $f_x(d_x')^2$ column are found by multiplying the amounts in the d_x' column by the amounts in the $f_x d_x'$ column. A similar procedure is followed in computing the amounts in the $f_y d_y'$ and $f_y(d_y')^2$ columns. The totals of these four columns are given below. N is equal to 383, since there are 383 frequencies.

$$\Sigma f_x d_x' = -31 \qquad \Sigma f_x (d_x')^2 = 831$$
$$\Sigma f_y d_y' = 19 \qquad \Sigma f_y (d_y')^2 = 1{,}725$$

The formula for the coefficient of correlation is as follows:

$$r = \frac{N \cdot \Sigma f d_x' d_y' - (\Sigma f_x d_x')(\Sigma f_y d_y')}{\sqrt{[N \cdot \Sigma f_x (d_x')^2 - (\Sigma f_x d_x')^2]\,[N \cdot \Sigma f_y (d_y')^2 - (\Sigma f_y d_y')^2]}} \quad \ldots . \quad (62)$$

It will be observed that all the amounts needed for substitution in the formula are available except $\Sigma f d_x' d_y'$. The computation of this amount is carried out on the cross-classification table. The symbol f refers to the frequencies in the individual rectangles in Table 18-3.

The number in the upper right-hand corner of the individual rectangles is the product of the d_x' and the d_y' values. The second number is the number of tallies recorded in that rectangle in Table 17-2. The last number is the product of the two preceding numbers and is the amount represented by $f d_x' d_y'$.

The heavy lines mark off the column for which d_x' is zero and the row for which d_y' is zero. The only numbers in the zero column and the zero row are the frequencies, since if either d_x' or d_y' is zero, the product of d_x' and d_y' is zero in each case. In each of the other rectangles that has a frequency, the computation of $f d_x' d_y'$ is made. No entry is made in a rectangle with no frequency, since the product is zero. $\Sigma f d_x' d_y'$ is the sum of all the individual products computed in the various rectangles. It is important to watch the signs of the individual products, since $\Sigma f d_x' d_y'$ is the algebraic sum of these amounts. For Table 18-3, $\Sigma f d_x' d_y'$ is -742. When all the above amounts are substituted in formula (62) for the coefficient of correlation, the following computations give the value of r and r^2.

$$r = \frac{(383)(-742) - (-31)(19)}{\sqrt{[(383)(831) - (-31)^2][(383)(1{,}725) - (19)^2]}}$$

$$r = \frac{-283{,}597}{\sqrt{(317{,}312)(660{,}314)}} = \frac{-283{,}597}{\sqrt{209{,}525{,}555{,}968}} = \frac{-283{,}597}{457{,}739}$$

$$r = -.620$$

$$r^2 = .384$$

The regression equation and the standard error of estimate are computed by the following equations:

$$b = \frac{N \cdot \Sigma f d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{N \cdot \Sigma f_x (d'_x)^2 - (\Sigma f_x d'_x)^2} \cdot \frac{i_y}{i_x} \quad \text{(63)}$$

$$a = \bar{Y} - b\bar{X} \text{ [equation (58)]}$$

The numerator and the denominator of the equation for b can be secured from the computation of the value of r, to give the following value:

$$b = \frac{-283{,}597}{317{,}312} \cdot \frac{10}{.10} = -89.4 \qquad (i_x = .10)$$

$$(i_y = 10)$$

The values of $\bar{X}$ and $\bar{Y}$ may be computed by using formulas (5) and (6) from Chapter 8:

$$\bar{X} = A + c \text{ [equation (5)] and}$$

$$c = \frac{\Sigma(fd')}{N} i \text{ [equation (6)]}$$

The arbitrary origin for the X distribution (A_x) is .745, so the computation of the mean of the X variable ($\bar{X}$) is as follows:

$$c = \frac{-31}{383} .10 = -.008 \qquad (i_x = .10)$$

$$\bar{X} = .745 - .008 = .737$$

The arbitrary origin for the Y distribution (A_y) is 64.5, so the computation of the mean of the Y variable ($\bar{Y}$) is as follows:

$$c = \frac{19}{383} 10 = .50 \qquad (i_y = 10)$$

$$\bar{Y} = 64.5 + .5 = 65.0$$

With the above values, the computation of the value of a is as follows:

$$a = 65.0 - (-89.4)(.737) = 130.9$$

The regression equation for the relationship between the average price paid for gas and the consumption of gas per customer is:

$$Y_c = 130.9 - 89.4X$$

The equation for the standard error of estimate (S_y) requires the standard deviation of the Y variable (σ_y), which may be computed from the following formula, making use of computations already made in finding r.

$$\sigma_y = i_y \frac{\sqrt{N \cdot \Sigma f_y (d_y')^2 - (\Sigma f_y d_y')^2}}{N} \quad \ldots\ldots\ldots \quad (64)$$

The numerator of the equation is $\sqrt{660,314}$, the second term in the denominator of the computation of r on page 445, which gives the following value for σ_y.

$$\sigma_y = 10 \frac{\sqrt{660,314}}{383} = 10 \frac{812.6}{383} = 21.2$$

Substituting in equation (60), the standard error of estimate is found as follows:

$$S_y = \sigma_y \sqrt{1 - r^2} = (21.2)\,(.785) = 16.6$$

SOME EXAMPLES OF THE USE OF CORRELATION ANALYSIS

The Crop Reporting Board in the United States Department of Agriculture, issues periodic reports during the growing months, giving estimates of the condition of important crops. The existence of a definite relationship between the estimates of the condition on different dates and the actual yield makes it possible to estimate the yield when the crop report on condition becomes available.

The relationship between the condition of the corn crop on September 1 in Kansas and the final yield per acre is shown by the following regression equation:

$$Y_c = .674 + .3155X$$

The X variable represents the condition of the crop on September 1, expressed as a percentage, comparing the condition of a growing crop at a given time with a normal condition taken as equal to 100 per cent. The Y variable is the actual yield per acre.

The regression equation expresses mathematically the relationship between the condition of the crop on September 1 and the final yield per acre. An increase of one per cent in the condition of the crop is associated with an increase of .3155 bushel. If on September 1 of a given year the condition were reported as 53 per cent of normal, the estimate based on the regression equation would be made as follows:

$$Y_c = .674 + .3155(53)$$

$$Y_c = .674 + 16.7215 = 17.4 \text{ bushels per acre}$$

The standard error of estimate, computed by using formula (53) on page 435, is 1.31 bushels. This is interpreted as indicating that 68 per cent of the estimates will fall within $\pm$ 1.31 bushels.

A study was made by the Bureau of Business Research at The University of Texas to determine the accuracy of the figures on bank debits in the Eleventh Federal Reserve District as a measure of changes in business conditions in Texas. The correlation between a composite index of business activity in Texas and an index of bank debits, adjusted for changes in the price level, was measured and the coefficient of determination was found to be .9965. Both of the series were adjusted for seasonal variation. The high degree of correlation between the two series meant that their fluctuations were almost parallel; and when certain data included in the composite index of business activity were no longer available, the decision was made to discontinue the computation of the index of business activity and to substitute the deflated index of bank debits in the Eleventh Federal Reserve District. The value of r^2 was high enough to justify this decision, and using this measure of correlation gave a more definite indication of the degree of relationship between the two series than could have been obtained from charts of the two time series.

One of the problems of scheduling advertising is to secure a distribution of the advertising that parallels fairly closely the potential market that is to be sold. Assume that a manufacturer of equipment that is normally bought by persons building new homes is planning an advertising campaign in magazines. One of the specifications he would probably make is that the total circulation of all the magazines used parallel fairly closely the potential market as represented by the volume of home building. A scatter diagram would graphically answer the question of the adequacy of coverage that the advertising campaign would secure, while the coefficient of determination would summarize mathematically the extent to which the advertising covered the potential market.

The correlation between the sales of nearly all types of consumer goods and disposable personal income is so high that an analysis of the market for nearly all kinds of goods includes a correlation analysis of sales with

personal income. The correlation can be made between the two variables related as to time units or geographical areas. Estimates of sales can be made from projections of personal income to future time periods by substituting these estimates in a regression equation describing the relationship between income and sales. Or, estimates of sales by regions can be made from regression equations measuring the relationship between sales and income in different areas. This use of correlation analysis is a standard technique in the analysis of markets and will be discussed further in Chapter 22.

Forecasting the prices of farm commodities relies heavily on correlation analysis of the relationship between supply and price and between demand and price. Relationships between price and both supply and demand frequently are nonlinear, and more than two variables may need to be studied simultaneously. Methods of analysis to be used in these situations are described in Chapter 19.

STUDY QUESTIONS

1. (a) In what way is the use of a regression line to measure the relationship between two variables like the use of a least squares line to measure secular trend?

(b) The arithmetic computations are usually easier for fitting a trend line than for fitting a regression line. Why is this the case?

2. Demonstrate that under certain conditions formulas (33) and (34), pages 287 and 288, are the same as formulas (51) and (52), (page 431).

3. (a) Could a measure of the reliability of estimates from a regression line be computed by ignoring the signs of the deviations from the line and computing the arithmetic mean of these deviations? Refer to the deviations in Column (6) of Table 18-2, page 435.

(b) What measure of dispersion computed in Chapter 9 would the above value resemble?

(c) S_y is preferred as a measure of dispersion over the measure described above. Why is this the case?

4. What assurance do you have that an estimate from a regression equation is a reasonably accurate value?

5. Why is it difficult to compare the degree of correlation in two problems by comparing the values of the standard error of estimate (S_y)?

6. What advantage does an abstract measure of correlation, such as the coefficient of correlation, have over S_y in comparing the degree of correlation present in two variables?

7. Explain what is meant by saying that the coefficient of determination (r^2) represents the fraction of the total variance that is explained by the X variable. Does this mean that it represents the fraction of the variance that is *caused* by the X variable?

8. Why is it impossible for r^2 to have a value of more than 1.00?

9. What indicates whether the relationship in a linear correlation is direct or inverse?

10. What are the advantages of using a cross-classification table in making a correlation analysis for a large number of items? Does the cross-classification table have any serious disadvantages?

PROBLEMS

1. Refer to the data in Problem 1, page 425, of Chapter 17.

(a) Compute the regression equation, standard error of estimate, and coefficient of determination.
(b) Plot the regression equation on the chart drawn in Problem 1, Chapter 17. (If you have not solved Problem 1, it will first be necessary to plot the data in the table on a scatter diagram.)
(c) Forecast the corn yield per acre in Michigan for a year in which the September 1 condition was reported to be 90 per cent of normal. How reliable would you consider this forecast?
(d) *Retain this solution for use in Problem 1 of Chapter 19.*

2. Refer to the data in Problem 2, page 425, of Chapter 17.

(a) For the variables X_1 and X_2, compute the following measures of correlation: regression equation, standard error of estimate, and coefficient of determination. The formulas in Chapter 18 designate the two variables in simple correlation as X and Y, so the two series given in Problem 2, Chapter 17, may be designated by these letters. Per capita meat consumption should be considered as Y, and deflated per capita disposable income, as X. (It would be equally valid to transform the formulas by substituting X_1 for Y and X_2 for X.)
(b) Plot the regression equation on the chart drawn in Problem 2, Chapter 17, to show the relationship between per capita meat consumption and deflated per capita disposable income. (If you have not solved Problem 2, it will first be necessary to construct a scatter diagram showing the relationship of per capita meat consumption to deflated per capita disposable income.)
(c) How accurate would you consider the regression equation as a forecaster of per capita meat consumption?
(d) *Retain this solution for use in Problem 3 of Chapter 19.*

3. Refer to the data in Problem 2, page 425, of Chapter 17.

(a) Compute the following measures of correlation for the variables X_1 and X_3: regression equation, standard error of estimate, and coefficient of determination. Per capita meat consumption (X_1) may be designated as the Y variable, and deflated meat prices (X_3) may be designated as the X variable.
(b) Plot the regression equation on the chart drawn in Problem 2, Chapter 17, to show the relationship between per capita meat consumption and deflated meat prices. (If you have not solved Problem 2, it will first be necessary to construct a scatter diagram showing the relationship of per capita meat consumption to deflated meat prices.)
(c) How accurate would you consider the estimate made of per capita meat consumption from deflated meat prices? Would the mean of the estimate in Problem 2 and this estimate be better than either estimate alone?
(d) *Retain this solution for use in Problem 3 of Chapter 19.*

4. Refer to the data in Problem 3, page 426, of Chapter 17.

(a) Compute the following measures of correlation for the variables X_1 and X_2: regression equation, standard error of estimate, and coefficient of determination. The percentage of persons using a motor vehicle (X_1) may be designated as the Y variable, and the percentage of incomes over \$2,000 (X_2) may be designated as the X variable.

(b) Plot the regression equation on the chart drawn in Problem 3, Chapter 17, to show the relationship between percentage of persons using a motor vehicle and percentage of incomes over \$2,000. (If you have not solved Problem 3, it will first be necessary to plot a scatter diagram showing the relationship between the percentage of persons using a motor vehicle and the percentage of incomes over \$2,000.)

(c) How accurately can you estimate the percentage of persons using a motor vehicle as a method of transportation from the percentage of incomes over \$2,000?

(d) *Retain this solution for use in Problem 4 of Chapter 19.*

5. Refer to the data in Problem 3, page 426, of Chapter 17.

(a) Compute the following measures of correlation for the variables X_1 and X_3: regression equation, standard error of estimate, and coefficient of determination. The percentage of persons using a motor vehicle (X_1) may be designated as the Y variable, and the population per square mile (X_3) may be designated as the X variable.

(b) Plot the regression equation on the chart drawn in Problem 3, Chapter 17, to show the relationship between percentage of persons using a motor vehicle and population per square mile. (If you have not solved Problem 3, it will first be necessary to plot a scatter diagram showing the relationship between the percentage of persons using a motor vehicle and the population per square mile.)

(c) How accurate do you consider the estimate of percentage of persons using a motor vehicle to be when made from the population per square mile for a given city? Would this estimate be more or less accurate than one using the relationship to percentage of incomes over \$2,000? Would an average of the two estimates be more accurate than either of the individual estimates?

(d) *Retain this solution for use in Problem 4 of Chapter 19.*

6. The analysis of the relationship between age and efficiency of the field workers of the state agency described in Problem 4, page 427, Chapter 17, will be continued in this problem.

(a) If you have not solved Problem 4, it will first be necessary to construct a cross-classification table from the data on ages and efficiency ratings.

(b) Compute the regression equation, standard error of estimate, and coefficient of determination.

(c) From the regression equation, estimate the efficiency rating you would expect a man 50 years old to receive. How accurate would you consider this rating to be?

(d) Using the results of the analysis of the correlation made in this problem and in Problem 4, Chapter 17, write a brief memorandum that summarizes what you have found out about the relationship between age and efficiency rating in the field force of the agency.

19 MEASUREMENT OF NONLINEAR AND MULTIPLE CORRELATION

This chapter will describe: (1) methods of measuring correlation when the relationship is best described by a nonlinear regression line, and (2) methods of measuring the relationship between more than two variables. When measuring the relationship between more than two variables, it may be found that nonlinear regressions describe the relationships better than straight lines. However, the measurement of nonlinear multiple correlation is an involved problem of an advanced nature and will not be discussed in this text. Mathematical methods of measuring nonlinear correlation between several variables are particularly burdensome, and the practical solution to this type of problem is usually found by employing a graphic method.

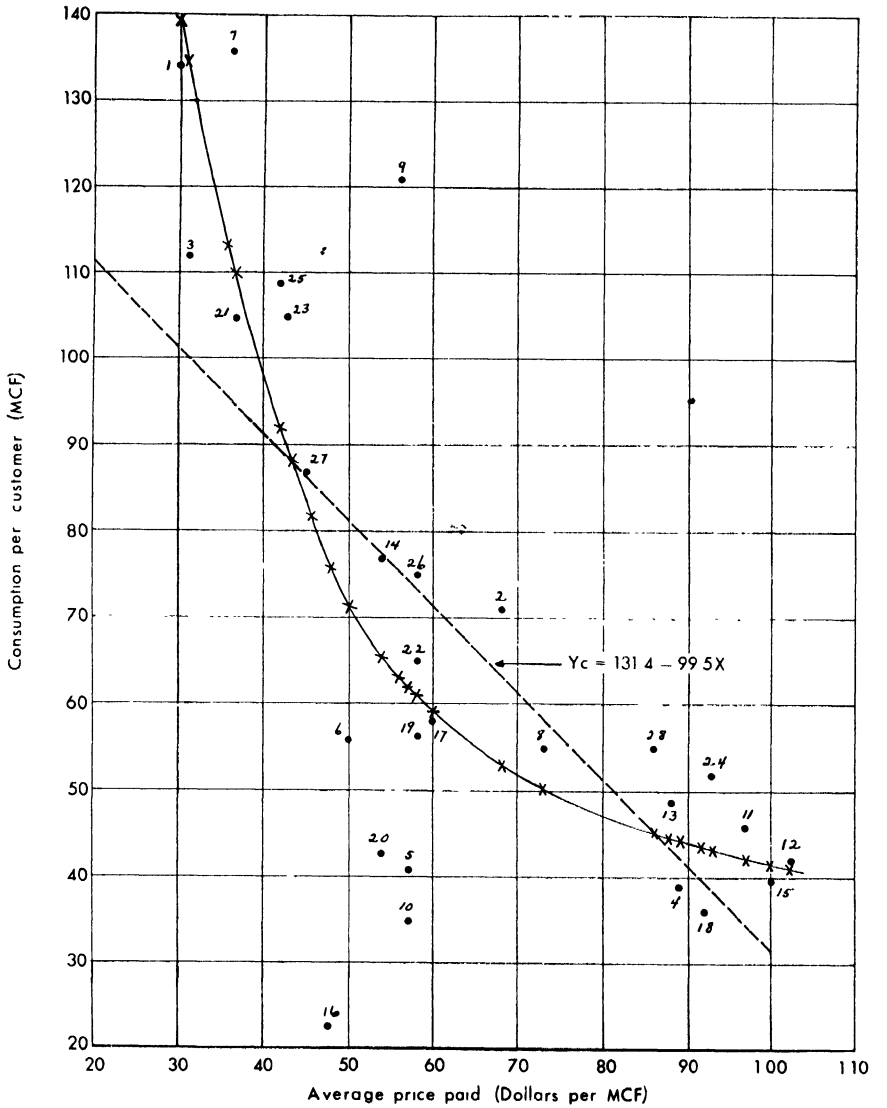
NONLINEAR CORRELATION

When a straight line is used to describe the relationship between two variables that are not related linearly, the straight line does not give so good an estimate of the Y variable as can be obtained from a curve that fits the data. The coefficient of determination (r^2) likewise will reflect this weakness when it is used as a measure of correlation. Nonlinear correlation involves deriving a measure of average relationship that is a good fit, computing a measure of reliability of estimates from this regression line, and deriving an abstract measure of the correlation represented by this regression line. The only difference in the nature of nonlinear and linear correlation is the fact that the regression line is a curve rather than a straight line.

Chart 19-1 is a scatter diagram showing the relationship between the price of gas in 28 Texas cities and the consumption per customer in the same cities. The number shown for each plotted point makes it possible to identify each city by reference to the numbers given in Table 19-1. The 28 cities in Table 19-1 are a sample from all of the cities in Texas with gas systems. This sample is from the same universe as the sample of 383 cities classified in Table 17-2 (page 421) by price and per capita consumption. The index of correlation will be computed for the sample of 28 cities; and the correlation ratio, another measure of nonlinear correlation, will be

Chart 19-1

SCATTER DIAGRAM OF CONSUMPTION OF NATURAL GAS PER CUSTOMER AND AVERAGE PRICE PAID IN 28 TEXAS CITIES, 1952



Source of data Table 19-1

computed for the sample of 383 cities in a later section. The two samples will be used because the correlation ratio is appropriate for a large number of items, since it is computed from a cross classification of the data. On the other hand, it is easier to compute the index of correlation from a small number of items.

Table 19-1

COMPUTATION OF INDEX OF CORRELATION

CONSUMPTION OF NATURAL GAS PER CUSTOMER AND AVERAGE PRICE PAID

City	Average price (Dollars per thousand cubic feet) X	Consumption per customer (Thousand cubic feet) Y	Y_c^*	$Y - Y_c$	$(Y - Y_c)^2$
1 Amarillo	.30	134	140	− 6	36
2 Arlington	.68	71	53	18	324
3 Borger	.31	112	135	−23	529
4 Brownsville	.89	39	45	− 6	36
5 Chillicothe	.57	41	62	−21	441
6 Corpus Christi	.50	56	72	−16	256
7 Dalhart	.37	136	113	23	529
8 Granger	.73	55	50	5	25
9 Hereford	.56	121	62	59	3,481
10 Iowa Park	.57	35	62	−27	729
11 Karnes City	.97	46	42	4	16
12 La Pryor	1.02	42	41	1	1
13 Llano	.88	49	45	4	16
14 Marshall	.54	77	66	11	121
15 Mathis	1.00	40	41	− 1	1
16 May	.48	23	76	−53	2,809
17 Memphis	.60	58	59	− 1	1
18 Mercedes	.92	36	43	− 7	49
19 Millsap	.58	56	61	− 5	25
20 Palestine	.54	43	66	−23	529
21 Pampa	.37	105	110	− 5	25
22 Palo Pinto	.58	65	61	4	16
23 Royalty	.43	105	89	16	256
24 San Marcos	.93	52	43	9	81
25 Shamrock	.42	109	92	17	289
26 Southland	.58	75	61	14	196
27 Texarkana	.45	87	82	5	25
28 Victoria	.86	55	45	10	100
Total	17.63	1,923	...	...	10,942

*Values read from the curve representing the regression line on Chart 19-1.

SOURCE: Sixty-First Annual Report of the Railroad Commission of Texas — Gas Utilities Division, 1952.

Regression Line

The straight line plotted on Chart 19-1 was fitted to the data by the method of least squares, but the curve was drawn by inspection as the best description of the relationship between the two variables. The computation of the measures of linear correlation are given on the next page. The calculation of the summations is not shown since they are made in the same manner as in Table 18-1 (page 432).

$$r = \frac{(28)\,(1{,}080.07) - (17.63)\,(1{,}923)}{\sqrt{[(28)\,(12.4139) - (17.63)^2]\;[(28)\,(160{,}329) - (1{,}923)^2]}}$$

$$r = \frac{-3{,}660.53}{\sqrt{(36.7723)\,(791{,}283)}}$$

$$r = -.6786$$

$$r^2 = .4605$$

$$\sigma_y^2 = \frac{791{,}283}{(28)^2} = 1{,}009.2895$$

$$b = \frac{-3{,}660.53}{36.7723} = -99.54585$$

$$\bar{X} = \frac{17.63}{28} = .62964$$

$$a = 68.6786 - (.62964)\,(-99.54585)$$

$$\bar{Y} = \frac{1{,}923}{28} = 68.6786$$

$$a = 131.36$$

$$Y_c = 131.36 - 99.546X$$

The drawing of the regression line by inspection is open to the same objections raised with respect to the graphic method of locating the trend line for a time series. The subjective element introduced by the person drawing the line injects a larger element of judgment into the solution than is to be desired; but the higher the correlation, the less the discretion that enters into the location of the line. The value of Y_c for each of the 28 cities was read from the curve and entered in Table 19-1.

Standard Error of Estimate

Using the values of Y_c read from the curve, the standard error of estimate was computed in exactly the same manner as in Table 18-2 (page 435). The deviation of the actual Y value from Y_c was computed for each of the 28 cities. These deviations were squared, averaged, and the square root extracted, with the result shown below.

$$S_y = \sqrt{\frac{10{,}942}{28}}$$

$$S_y = \sqrt{390.7857}$$

$$S_y = 19.8$$

The meaning of the standard error of estimate is the same for nonlinear as for linear correlation; it is an average of the deviations of the actual values from the values computed from the regression equation. If these deviations are normally distributed, approximately 68 per cent of the estimates from the regression equation will fall within the range $\pm 1S_y$. This means that the standard error of estimate serves as a measure of reliability

of the estimating equation. Stated in another way, the standard error of estimate measure the goodness of fit of the regression line, whether this line is a curve or a straight line.

Index of Correlation (rho)

The same need for an abstract measure of correlation exists for nonlinear correlation as for linear. Since the standard error of estimate is computed in the same basic manner for both types of relationship, the abstract measure of relationship can be computed from S_y^2 by formula (56) on page 439, which was used for the computation of r^2. The coefficient of correlation (r) is in reality a special case of the index of correlation. In order to call attention to the fact that the regression equation is not a straight line, the Greek letter *rho* (ρ) is used as the symbol for the measure of correlation, called the *index of correlation*. The term *coefficient of determination* is applied to ρ_{xy}^2 and is interpreted as the percentage of the variance in the Y variable that is explained by the variations in the X variable. The subscripts are always used with the index of correlation, since the value of ρ_{yx} is not necessarily the same as ρ_{xy}. The subscripts are not necessary for linear correlation, since r_{xy} is the same as r_{yx}.

When using the index of correlation, one must always have information on the regression line, since the value of *rho* has no significance except as related to the measure of average relationship. The coefficient of correlation is simpler to use since it is known that it is always the measure of abstract relationship when the relationship is measured by a straight line. The sign of the slope of the linear regression line is given to the value of r, with a plus slope representing a positive relationship between the variables, and a minus, an inverse relationship.

When the regression line is not a straight line, the slope may be positive at some points on the line and negative at other points. For example, the correlation between yield per acre (the dependent variable) and the amount of irrigation water applied might show a positive correlation at some points on the line and a negative correlation at other points. When the amount of water applied is relatively small, the yield per acre is higher for larger amounts of water than for smaller amounts. This means that the correlation is positive. However, farther to the right on the correlation chart, where relatively large amounts of water are applied, yields associated with very large amounts of water are less than for the smaller amounts. This means that the correlation is negative. In other words, up to a given amount of water, adding more water increases the yield and the correlation is positive. But after this optimum amount of water has been applied, a further increase results in a decrease in yield.

Since it is always possible that the relationship in a nonlinear correlation may be positive over part of the range of the variables and negative for other values, it is customary not to attach any sign to the value of a measure of correlation. The simplest way to determine whether the relationship is direct or inverse is to plot the regression line on a scatter diagram. Such a chart will indicate when the relationship is negative and when it is positive.

The calculation of the value of ρ is shown below, using the values of S_y^2 and σ_y^2 already computed.

$$\rho_{yx}^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad \text{(65)}$$

$$\rho_{yx}^2 = 1 - \frac{390.7857}{1{,}009.2895}$$

$$\rho_{yx}^2 = .6128$$

$$\rho_{yx} = .7828$$

The value of r previously computed on page 454 was $-.6786$. The higher value of ρ indicates that the nonlinear regression is a better description of the relationship between the consumption of gas per customer and the price.

In this connection it is important to remember that the significance of this measure of relationship depends upon the validity of the regression line as a measure of average relationship. One caution to be observed is to avoid using a regression line that follows the plotted points too closely. The extreme case would be for the regression line to pass through all of the points, with the result that there would be no scatter about the line, and, therefore, perfect correlation would exist. If this close fit were made by a very irregular line, it would be obviously incorrect to consider this line a measure of average relationship. This introduces the condition that the regression line, if not a straight line, must be a smooth curve that changes direction only slowly.

The Correlation Ratio (eta)

The use of the graphic method of locating the regression line described above is open to criticism because its use permits the person drawing the line to decide where he thinks it should go. Although it is possible to fit a nonlinear regression line using the least squares method described for the fitting of nonlinear trend lines, this method is extremely burdensome and has been omitted in this text in favor of the graphic method.

An alternative method of measuring nonlinear correlation is available by means of the *correlation ratio*, represented by the Greek letter *eta* (η). This method provides an abstract measure of correlation without the necessity of computing the regression equation. The means of the columns

Table 19-2

CROSS TABULATION OF CONSUMPTION OF NATURAL GAS PER DOMESTIC CUSTOMER AND AVERAGE PRICE PAID, 1952

PRICE (DOLLARS PER THOUSAND CUBIC FEET)

CONSUMPTION PER CUSTOMER (THOUSAND CUBIC FEET)

	0 - .09	.10- .19	.20- .29	.30- .39	.40- .49	.50- .59	.60- .69	.70- .79	.80- .89	.90- .99	1.00- 1.09	1.10- 1.19	1.20- 1.29	1.30- 1.39	1.40- 1.49	1.50- 1.59
180-189					1											
170-179																
160-169						1										
150-159																
140-149																
130-139				2	2	2										
120-129						3										
110-119				1	1	7	1									
100-109					1	4		1								
90-99						6	3	2								
80-89					2	6	4	4								
70-79				1	1	6	17	24								
60-69					1	7	38	46	35		1					
50-59						4	8	27	34	9	1					
40-49						3	2	6	15	14	3			1		
30-39						2	3	1	7	2	5	1				
20-29						1				1						1
10-19									1							
f				4	9	52	76	111	92	26	10	1	0	1	0	1
$\bar{Y}_k$				114.5	108.9	85.5	66.9	64.5	54.7	46.4	42.5	34.5	...	44.5	...	24.5

SOURCE: Table 17-2.

in a cross-classification table are used as the Y_c values, and the coefficient of determination (η^2) is computed by using the basic formula by which r^2 and ρ^2 have been computed.

Table 19-2 is the cross-classification table first given in Table 17-2 (page 421) and used in Table 18-3 (page 443) to compute the different measures of linear correlation. The relationship between the two variables appears to be somewhat nonlinear, and in this section the correlation ratio will be computed to describe the relationship.

In Table 19-2 the row with the stub f shows the number of frequencies in the columns. Each column is a frequency distribution and the mean of each frequency distribution is given in the last row of the table, with the stub $\overline{Y}_k$. The mean of each of these frequency distributions can be computed by the method given in Table 8-3 (page 166) or the method given in Table 8-4 (page 168). Each of the means is plotted on the table, using the cross-classification table as the grid of an arithmetic chart. A dotted line connects each of the points and serves as the regression line.

Table 19-3 is set up from the information in Table 19-2. Column (1) of Table 19-3 shows the class intervals on the X scale for each of the columns. In Column (2) the frequency for each column of the cross-

Table 19-3

COMPUTATION OF CORRELATION RATIO

CONSUMPTION OF NATURAL GAS PER CUSTOMER AND AVERAGE PRICE PAID IN 383 TEXAS CITIES, 1952

Class interval of columns (Price in dollars per thousand cubic feet) (1)	(2) f	Mean of each column (Thousand cubic feet) (3) $\overline{Y}_k$	Deviation of column mean from mean of all items (4) $(\overline{Y}=65.0)$ $\overline{Y}_k-\overline{Y}$	(5) $(\overline{Y}_k-\overline{Y})^2$	(6) $f(\overline{Y}_k-\overline{Y})^2$
0— .09....	0	...	...	...	...
.10— .19....	0	...	...	...	...
.20— .29....	0	...	...	...	...
.30— .39....	4	114.5	49.5	2,450.25	9,801.00
.40— .49....	9	108.9	43.9	1,927.21	17,344.89
.50— .59....	52	85.5	20.5	420.25	21,853.00
.60— .69....	76	66.9	1.9	3.61	274.36
.70— .79....	111	64.5	— .5	.25	27.75
.80— .89....	92	54.7	—10.3	106.09	9,760.28
.90— .99....	26	46.4	—18.6	346.05	8,997.30
1.00—1.09....	10	42.5	—22.5	506.25	5,062.50
1.10—1.19....	1	34.5	—30.5	930.25	930.25
1.20—1.29....	0	...	...	...	...
1.30—1.39....	1	44.5	—20.5	420.25	420.25
1.40—1.49....	0	...	...	...	...
1.50—1.59....	1	24.5	—40.5	1,640.25	1,640.25
Total.....	383	...	...	...	76,111.83

SOURCE: Table 19-2.

classification table is listed, and in Column (3) the value of the mean of each column in the cross-classification table is given. At this point a variation in the method of computing the coefficient of determination is employed in order to save time in calculation. In the computation of the value of r^2 and ρ^2, the deviation of each Y value from the related Y_c value was computed in finding the standard error of estimate (S_y). This method can be used in computing *eta* by finding the deviation of each of the frequencies in a given column from the mean of the column, but the same answer can also be secured by a formula in which *the deviation of the mean of each column from the mean of the entire Y distribution is computed and used in computing eta.* These deviations ($\bar{Y}_k - \bar{Y}$) are squared, and since the number of items from which the mean is computed is different for each column, each squared deviation must be weighted by the number of frequencies in the column from which the column mean ($\bar{Y}_k$) was computed. Column (4) in Table 19-3 gives the values of $\bar{Y}_k - \bar{Y}$, and Column (5) shows these values squared $(\bar{Y}_k - \bar{Y})^2$. In Column (6) each of the squared deviations was multiplied by the related frequency. This is represented by the term $f(\bar{Y}_k - \bar{Y})^2$; the explained variance is represented by σ_m^2 and is measured as follows:

$$\sigma_m^2 = \frac{\Sigma f(\bar{Y}_k - \bar{Y})^2}{N} \quad \text{(66)}$$

$$\sigma_m^2 = \frac{76{,}111.83}{383} = 198.73$$

The equation for η^2 gives the proportion of the total variance in Y that is explained by the X variable. σ_y was computed on page 446 and is 21.2.

$$\eta^2 = \frac{\sigma_m^2}{\sigma_y^2} \quad \text{(67)}$$

$$\eta^2 = \frac{198.73}{449.44} = .442$$

$$\eta = .665$$

The correlation ratio is the ratio between the dispersion in the column means and the total dispersion in the Y variable. If the column means show the same dispersion about the mean of the Y variable as the individual items, the classification of the data by the X variable adds nothing to our knowledge of the variations in the Y variable. In the above example, the 383 cities were classified into 16 price groups, with a uniform class interval

of $.10 per thousand cubic feet of gas. The cities in each price group were classified by the amount of gas consumed per customer, and Table 19-2 shows the average consumption per customer in the cities in each price group to be consistently lower for the higher prices. In other words, a considerable proportion of the variation in consumption per customer is related to the price of gas.

It probably should be pointed out again that measures of relationship do not indicate anything about cause-and-effect relationships. The consumption of gas may be high where the price is low because the lower prices stimulate consumption. On the other hand, consumption may be higher in some cities for other reasons, such as colder winter weather. The larger consumption in some cities may make lower prices possible there. All that the correlation ratio shows is the extent to which the variations of the two variables parallel; it does not give any indication of which factor is the cause, which is the effect, or whether either is the cause or effect.

If the variance of the means of the columns equaled the total variance in the Y variable, the individual amounts of consumption per customer in the columns would all coincide with the means of the columns. In other words, there would be no scatter about the column means, which is another way of saying that the correlation is perfect.

The correlation ratio measures the relationship between the two variables, whether the relationship is linear or nonlinear; but if the column means fall on a straight line, the value of *eta* is the equivalent of r. The extent to which *eta* and r diverge is a measure of nonlinearity in the relationship between the two variables.

MULTIPLE CORRELATION

The discussions of correlation up to this point have involved measuring the relationship between two variables only, although it has been indicated that at times more than two variables could, or should, be considered. In the discussion of the correlation between the number of burners in use and the consumption of heating oil, the scatter diagram in Chart 18-1 (page 431) gave some indication that factors other than the number of burners in use might have an influence on the consumption of heating oil. Table 18-2 (page 435) shows the consumption of heating oil to be 18.3 million barrels below the amount computed from the regression equation in 1953 and 13.0 million barrels above the computed consumption in 1955. These were the largest deviations from the computed values, although in 1947, 1948, and 1949 also the variation was substantial. The problem of multiple correlation is to relate these deviations to one or more variables that might have an influence on the consumption of heating oil.

Since it appears to be obvious that the consumption of heating oil is influenced by the temperature, it is worthwhile to make a study of the variations in the consumption of heating oil as related to the variations in temperature from year to year. The data on consumption of heating oil, number of burners in use, and the average temperature in New York City are given in Table 19-4 for the years 1946 to 1955.

When only two variables were included in the study, the dependent variable was designated Y, and the independent variable, X; but when more than two variables are studied, it is more convenient to use X to designate all the variables and to distinguish between them by means of subscripts. In this chapter, heating oil consumption will be indicated as X_1; number of burners in use, X_2; and temperature in New York City, X_3. This notation makes it possible to handle systematically any number of variables.

The simplest type of analysis that could be made would be to measure the correlation between heating oil consumption and temperature, ignoring any influence the number of burners might have on consumption. The scatter diagram in Chart 17-5 (page 416) shows that there is very little correlation between these two variables; the coefficient of determination (r^2) is .022. This seems to refute the assumption stated above that temperature and consumption of heating oil are obviously related; but, as was pointed out on page 415, the possibility exists that the high correlation between consumption and number of burners obscures the relationship between consumption and temperature. Multiple correlation analysis is a method of taking into account simultaneously the relationship between the three variables. When both temperature and the number of burners are correlated at the same time with consumption, it will give a clearer measure of the relationships existing between the variables.

The Regression Equation

The measurement of the relationship between three or more variables is an extension of the basic principles used in two-variable linear correlation. The first step is to determine an equation of the average relationship between the variables. From this equation, values of the dependent variable may be computed ($X_{1\,c}$); the higher the correlation between the variables, the smaller will be the deviation of the actual values from the computed values. This deviation will represent the variance in the dependent variable that is not explained by the independent variables, and will make possible the computation of a coefficient of determination similar to r^2 computed in Chapter 18.

Table 19-4

COMPUTATION OF REGRESSION EQUATION

HEATING OIL CONSUMPTION, DOMESTIC BURNERS IN USE, AND AVERAGE TEMPERATURE IN NEW YORK CITY: 1946 TO 1955

Year	Heating oil consumption (Million barrels) X_1	Domestic burners in use (Millions) X_2	Average temperature in New York City (Degrees) X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
1946...	140	2.67	54.9	19,600	7.1289	3,014.01	373.80	7,686.0	146.583
1947...	178	3.24	53.4	31,684	10.4976	2,851.56	576.72	9,505.2	173.016
1948...	200	3.81	53.6	40,000	14.5161	2,872.96	762.00	10,720.0	204.216
1949...	190	4.23	56.6	36,100	17.8929	3,203.56	803.70	10,754.0	239.418
1950...	221	4.83	53.7	48,841	23.3289	2,883.69	1,067.43	11,867.7	259.371
1951...	250	5.44	54.6	62,500	29.5936	2,981.16	1,360.00	13,650.0	297.024
1952...	263	6.03	55.2	69,169	36.3609	3,047.04	1,585.89	14,517.6	332.856
1953...	267	6.68	56.2	71,289	44.6224	3,158.44	1,783.56	15,005.4	375.416
1954...	304	7.31	54.3	92,416	53.4361	2,948.49	2,222.24	16,507.2	396.933
1955...	340	7.90	54.2	115,600	62.4100	2,937.64	2,686.00	18,428.0	428.180
Total...	2,353	52.14	546.7	587,199	299.7874	29,898.55	13,221.34	128,641.1	2,853.013

SOURCE: Table 17-1.

The regression equation for the relationship between heating oil consumption and number of burners was computed in Chapter 18 to be

$$Y_c = 57.43 + 34.11X$$

Using X with subscripts for the different values of the variables, the equation can be written in the following form:

$$X_{1c} = 57.43 + 34.11X_2$$

The reliability of estimates based on the above equation is shown by the value of the standard error of estimate (S_{12}), which was found to be 10.2. In Chapter 18 the standard error of estimate was represented by S_y, since it measures the reliability of estimates of the Y variable. In multiple correlation analysis the subscript y is replaced by the numerical subscripts representing the variable being estimated and the variables used in making the estimate. The symbol S_{12} is interpreted as the standard error of estimate of variable X_1 when the estimates are based on variable X_2. If heating oil consumption were estimated from temperature (X_3), the symbol for the standard error of estimate would be S_{13}. When the estimate of X_1 is made from values of X_2 and X_3, the standard error of estimate is represented by $S_{1.23}$.

The estimating equation for a correlation analysis with two independent variables is written in the form:[1]

$$X_{1c} = a + b_{12.3}X_2 + b_{13.2}X_3$$

The coefficient $b_{12.3}$ represents the net relationship between variables X_1 and X_2, when the influence of X_3 is taken into account. In making an estimate of X_1, the product of $b_{12.3}$ and X_2 constitutes the contribution of variable X_2 to the estimate. Likewise the coefficient $b_{13.2}$ represents the net relationship between variables X_1 and X_3, taking into account the influence of X_2. The product of $b_{13.2}$ and X_3 constitutes the contribution of variable X_3 to the estimate.

The b coefficient in simple correlation measures the relationship between the two variables included in the study, but if additional variables have an important influence on the dependent variable, their effect is not isolated but is simply mixed with the influence of the one variable that is being

[1]The subscripts of the b coefficients are designed to indicate their meaning. The two numbers to the left of the decimal are the *primary subscripts* and show which two variables are related by this coefficient. The subscript to the right of the decimal is the *secondary subscript*, and the variable represented by it is included in the study; but its effect is eliminated in determining the relationship of the two variables represented by the primary subscripts. For example, $b_{12.3}$ shows the relationship between variables X_1 and X_2 with the influence of variable X_3 eliminated. The coefficient $b_{13.2}$ shows the relationship between variables X_1 and X_3 with the influence of X_2 removed. If there were four independent variables instead of two, the subscript showing the relationship between variables X_1 and X_2 would be written $b_{12.345}$. The influence of variables X_3, X_4, and X_5 would be eliminated in determining the relationship between variables X_1 and X_2.

measured. When the relation between just two variables is measured, there is no way of knowing how much effect other factors may have on the dependent variable. It is possible for the influence of two independent variables to counteract each other to the extent that neither appears to show any correlation with the dependent variable, yet when the multiple relationship is measured the correlation may be high.

The following illustration will give all of the steps in computing the regression equation for estimating heating oil consumption from the number of burners in use and the temperature in New York City. It should be noted in this connection that more accurate values of the temperature could be secured by using the average temperature in many cities, but it is assumed that the variations of the temperature in New York City will be representative of an area that consumes a large proportion of the oil used for heating.

The equations below may be solved simultaneously to secure the values of the coefficients of the estimating equation given on page 463. It will be seen that the three coefficients are the unknowns in the three equations; the values of the summations may be obtained from the actual data for the three variables, X_1, X_2, and X_3. The computation of the sums of the squares and cross products is shown in Table 19-4.

$$\textbf{I.} \quad \Sigma X_1 = Na + b_{12\,3}\,\Sigma X_2 + b_{13\,2}\,\Sigma X_3$$

$$\textbf{II.} \quad \Sigma X_1 X_2 = a\Sigma X_2 + b_{12\,3}\,\Sigma X_2^2 + b_{13\,2}\,\Sigma X_2 X_3$$

$$\textbf{III.} \quad \Sigma X_1 X_3 = a\Sigma X_3 + b_{12\,3}\Sigma X_2 X_3 + b_{13\,2}\,\Sigma X_3^2$$

When the values of the summations from Table 19-4 are substituted in the three equations, they appear in the following form:

$$\textbf{I.} \quad 2{,}353 = 10a + 52.14b_{12\,3} + 546.7b_{13\,2}$$

$$\textbf{II.} \quad 13{,}221.34 = 52.14a + 299.7874b_{12\,3} + 2{,}853.013b_{13\,2}$$

$$\textbf{III.} \quad 128{,}641.1 = 546.7a + 2{,}853.013b_{12\,3} + 29{,}898.55b_{13\,2}$$

Since there are three unknowns and three equations, they may be solved by any of the standard methods. The values of the three coefficients obtained from the solution of the equations are as follows:

$$a = 498.8675$$

$$b_{12.3} = 34.8491$$

$$b_{13\,2} = -8.1447$$

The regression equation for estimating the consumption of heating oil is as follows:

$$X_{1\,c} = 498.8675 + 34.8491X_2 - 8.1447X_3$$

Table 19-5

COMPUTATION OF COEFFICIENT OF MULTIPLE CORRELATION

HEATING OIL CONSUMPTION, DOMESTIC BURNERS IN USE, AND AVERAGE TEMPERATURE IN NEW YORK CITY: 1946 TO 1955

Year (1)	Domestic burners in use (Millions) (2) X_2	Average temperature in New York City (Degrees) (3) X_3	Heating oil consumption (Million barrels) (4) X_1	(5) X_{1c}	(6) $X_1 - X_{1c}$	(7) $(X_1 - X_{1c})^2$
1946...	2.67	54.9	140	144.8	−4.8	23.04
1947...	3.24	53.4	178	176.9	1.1	1.21
1948...	3.81	53.6	200	195.1	4.9	24.01
1949...	4.23	56.6	190	185.3	4.7	22.09
1950...	4.83	53.7	221	229.8	−8.8	77.44
1951...	5.44	54.6	250	243.7	6.3	39.69
1952...	6.03	55.2	263	259.4	3.6	12.96
1953...	6.68	56.2	267	273.9	−6.9	47.61
1954...	7.31	54.3	304	311.4	−7.4	54.76
1955...	7.90	54.2	340	332.7	7.3	53.29
Total..	...	...	2,353	2,353.0	0	356.10

SOURCE: Table 17-1.

Reliability of Estimates

The problem of determining the accuracy of estimates made from the regression equation is the same for multiple correlation as for simple correlation. Estimates of X_1 are made from the regression equation, and the deviations of the values of X_1 from X_{1c} are the basis for the standard error of estimate. Employing the same rules for subscripts that are used in describing the coefficients of the regression equation, the symbol $S_{1.23}$ represents the standard error of estimate. The standard error of estimate represents the average of the difference between the actual values of the dependent variable (X_1) and the computed value (X_{1c}). The computed values were based on the multiple relationship between X_1 and the two independent variables, X_2 and X_3. The subscripts of $S_{1.23}$ show this fact, since the subscript to the left of the decimal is the dependent variable and the subscripts to the right of the decimal represent the independent variables on which the estimate is based.

The computation of the standard error of estimate ($S_{1\,23}$) is shown in Table 19-5. The values of $X_{1\,c}$ were calculated and entered in Column (5). The deviations from $X_{1\,c}$ were computed and entered in Column (6), and these deviations were squared in Column (7). The sum of Column (7) was used to compute the value of $S_{1.23}$ as follows:

$$S_{1.23} = \sqrt{\frac{\Sigma(X_1 - X_{1\,c})^2}{N}} \quad \ldots\ldots\ldots\ldots \quad (68)$$

$$S_{1.23} = \sqrt{\frac{356.10}{10}} = \sqrt{35.61}$$

$$S_{1.23} = 5.97$$

It is interesting to compare the values of $X_1 - X_{1\,c}$ in Table 19-5 with the values of $Y - Y_c$ in Table 18-2 (page 435). Estimates for seven years are substantially closer to the actual consumption in Table 19-5 than in Table 18-2; while for three of the years, Table 18-2 gives estimates closer to the actual consumption. The total deviation of actual from the estimated consumption is much greater in Table 18-2 than in Table 19-5; the sum of the squared deviations for the two-variable correlation is 1,032.69; while for the multiple regression equation, the sum of the squared deviations is 356.10.

The difference in accuracy between the estimating equation using one independent variable and the one using two independent variables reflects the influence of the second independent variable. The standard error of estimate for the multiple regression equation is 5.97, compared with the standard error of 10.2 when the estimates were made from the one variable, the number of burners in use. This difference indicates that more than half of the variance that was still unexplained by variations in the number of burners is explained by the fluctuations in temperature. If the net influence of the additional variable were zero, it would make no difference in the estimate whether it was included or not; but as long as there is any net correlation between an independent variable and the dependent variable, adding the independent variable will improve the estimate. This subject will be developed further in the section dealing with partial correlation.

Abstract Measures of Multiple Correlation

The coefficient of determination may be computed for multiple correlation by the same formula used for simple correlation, with the exception that the value of $S^2_{1.23}$ is substituted for the value of S^2_y.

$$R^2_{1.23} = 1 - \frac{S^2_{1.23}}{\sigma_1^2} \quad \text{(69)}$$

$$R^2_{1.23} = 1 - \frac{35.61}{3{,}353.81} = 1 - .011 = .989$$

The subscripts indicate that the coefficient measures the percentage of the total variance in variable X_1 that is explained by the combined variations in variables X_2 and X_3. The subscripts are written in the same manner as in the case of the standard error of estimate, with the subscript at the left of the decimal representing the dependent variable, and those at the right, the independent variables.

The significance of the coefficient of multiple determination is the same as that of the coefficient of determination for a two-variable correlation. It represents the proportion of the total variance in the dependent variable that is explained by variations in the independent variables. The only difference is that more than one variable is used instead of a single independent variable to explain the variance in the dependent variable.

Partial Correlation

It was stated earlier in the chapter that the *b* coefficients of the regression equation measure the net influence of each independent variable on the estimate of the variable X_1. This section will describe a method of measuring the *net correlation* or *partial correlation* between one independent variable and the dependent variable, eliminating the relationship with the other independent variables in the study. This measure of relationship is known as the *coefficient of partial correlation* and is represented by the symbol $r_{13.2}$. The primary and secondary subscripts are employed with the same meanings given them previously in this chapter. The subscripts to the left of the decimal point represent the variables for which the net correlation is being computed; the subscript to the right of the decimal point represents the variable that has been eliminated.

The simple linear correlation between heating oil consumption and number of burners gave the following regression equation:

$$X_{1c} = 57.43 + 34.11X_2$$

The reliability of estimates from this equation is given by the value of the standard error of estimate. For estimates of heating oil consumption made from the above equation, $S^2_{12} = 103.3$. The value of S^2_{12} represents the variance of variable X_1 that is *not explained* by variable X_2. The standard error of estimate for the multiple regression equation used to estimate heating oil consumption (X_1) from variables X_2 and X_3 was computed earlier in the chapter and it was found that $S^2_{1.23} = 35.61$.

The value of $S^2_{12.3}$ represents the variance in variable X_1 unexplained after variables X_2 and X_3 have both been used in making the estimate of X_1. The reduction in the unexplained variance accomplished by adding variable X_3 measures the net correlation of variables X_1 and X_3 when the influence of X_2 on both variables X_1 and X_3 is taken into account. The unexplained variance was 103.3 when only independent variable X_2 was considered. When X_3 was added, the unexplained variance dropped to 35.61. The coefficient of partial determination is computed from these values as follows:

$$r^2_{13.2} = 1 - \frac{S^2_{1.23}}{S^2_{12}} \quad \text{(70)}$$

$$r^2_{13.2} = 1 - \frac{35.61}{103.3}$$

$$r^2_{13.2} = .655$$

$$r_{13.2} = -.809$$

The coefficient of net correlation ($r_{13.2}$) takes the sign of the b coefficient equation, which is minus for variable X_3.

STUDY QUESTIONS

1. What is meant by the statement that the relationship between two variables is "nonlinear"?

2. How can you determine whether the relationship between two variables is direct or inverse when it is nonlinear?

3. Why is the value of ρ, which was computed on page 456, larger than the value of r computed from the same data? Is it to be expected that the value of ρ will generally be larger than r?

4. Why is it incorrect, for use as a measure of average relationship in a correlation problem, to draw a very irregular line that passes through most of the plotted points?

5. Do the correlation ratio (η) and the index of correlation (ρ) measure the same characteristic or two different characteristics? Explain.

6. Under what circumstances does eta (η) have the same value as r?

7. Why is simple correlation not adequate for all situations in which correlation analysis techniques may be used?

8. In measuring the relationship between heating oil consumption and the number of burners in use, it was found that $b = 34.1$ (page 432). In the analysis of multiple correlation, it was found that $b_{12.3} = 34.8$ (page 464). Explain the significance of these two coefficients, pointing out how they differ.

9. What is the difference between r_{12}, $R_{1.23}$, and $r_{12.3}$?

10. What does the difference between S_{12} (or S_y) and $S_{1.23}$ measure?

11. Give an example (other than given in the text) to illustrate how a measure of the relationship between three or more variables may be used to estimate or forecast one of the variables.

PROBLEMS

1. This problem is based upon the solutions to Problem 1 of Chapter 17, page 425, and Problem 1 of Chapter 18, page 449.

(a) Draw a nonlinear regression line on the chart prepared in Problem 1, Chapter 17. A curve should fit the data better than the straight line computed by the method of least squares in Problem 1, Chapter 18.
(b) Read the values of Y_c from the regression curve, and from these values compute the standard error of estimate.
(c) Compute the value of the index of correlation (ρ). How does this value compare with the value of r computed in Problem 1, Chapter 18?
(d) Do you consider the linear or the nonlinear regression the better device for forecasting the yield per acre from the condition on September 1?
(e) Forecast the corn yield in Michigan for a year in which the September 1 condition was reported to be 90 per cent of normal. How does this compare with the estimate made in Problem 1, Chapter 18?

2. This problem is based upon the solution to Problem 4, Chapter 17, page 427.

(a) If you have not completed Problem 4, it will first be necessary to construct a cross-classification table from the data.
(b) From the cross-classification table, compute the correlation ratio (η).
(c) Explain the significance of the correlation ratio and compare it with the significance of the coefficient of correlation.

3. This problem is based upon the solutions to Problem 2 of Chapter 17, page 425, and Problems 2 and 3 of Chapter 18, page 449.

(a) Make a complete multiple correlation analysis of the three series given in Problem 2, Chapter 17, computing the following measures: regression equation, $S_{1.23}$, and $R^2_{1.23}$.
(b) What is the significance of the value of $R^2_{1.23}$, compared to the values of the coefficient of determination computed in Problems 2 and 3, Chapter 18?
(c) Using the multiple regression equation, estimate the per capita consumption of meat in a year with deflated per capita disposable income of \$1,500 (1947 dollars) and deflated meat prices of 110 (index, 1947–49 = 100).
(d) Using the simple regression equations computed in Problems 2 and 3, Chapter 18, make two estimates of per capita meat consumption from the data given in (c) above. Are these estimates as accurate as that made from the multiple regression equation?

4. This problem is based upon the solutions to Problem 3 of Chapter 17, page 426, and Problems 4 and 5 of Chapter 18, page 450.

(a) Make a complete multiple correlation analysis of the three series given in Problem 3, Chapter 17, computing the following measures: regression equation, $S_{1.23}$, and $R^2_{1.23}$.

(b) What is the significance of the value of $R^2_{1.23}$, compared to the values of the coefficient of determination computed in Problems 4 and 5, Chapter 18?
(c) Using the multiple regression equation, estimate the percentage of persons using a motor vehicle in a city with 75 per cent of the incomes over $2,000 and a density of 8,000 persons per square mile.
(d) Using the simple regression equations computed in Problems 4 and 5, Chapter 18, make two estimates of the percentage of persons using a motor vehicle from the data given in (c) above. Are these estimates as accurate as that made from the multiple regression equation?

5. Compute the coefficients of partial correlation, $r_{12.3}$ and $r_{13.2}$, from the calculations made in Problem 3 above. Explain what these coefficients mean.

6. Compute the coefficients of partial correlation, $r_{12.3}$ and $r_{13.2}$, from the calculations made in Problem 4 above. Explain what these coefficients mean.

Part

VI APPLICATIONS IN BUSINESS

chapter 20 ♦ Business Barometers

chapter 21 ♦ Business Forecasting

chapter 22 ♦ Marketing Research

chapter 23 ♦ Statistical Quality Control

chapter 24 ♦ Community Economic Surveys

20 BUSINESS BAROMETERS

Up to this point emphasis has been placed upon the methods of analyzing quantitative data. The methods of statistical analysis and their usage have been described by specific examples which were limited to the individual method under discussion. In each case an attempt has been made to illustrate the types of problems to which the method of analysis being described could be adapted, selecting the examples from various fields of application. Each of the following chapters, on the other hand, is devoted to a discussion of a specialized professional field of activity and points out some of the most widely used methods of statistical analysis in each field. The subjects discussed are those in which substantial use of statistical analysis has been made in reaching solutions to business problems; however, no attempt has been made to cover all possible applications of statistical methods, for some of the topics included in this final part are recognized as specialized fields in which many books have been written.

THE NEED FOR BUSINESS BAROMETERS

There are various reasons why businessmen might be interested in the past fluctuations in business conditions, but probably the most important reason for using such information is its value in forecasting future changes in business activity. A businessman trying to forecast changes in the business situation is dependent upon two things: (1) he needs to understand the causal factors that tend to bring about change in the business situation, and (2) he needs an accurate and prompt measurement of the changes that have taken place in the different phases of business. This is true whether he is concerned with an individual business or with over-all business activity.

Economists have made progress in explaining why economic activity fluctuates, but the factors that cause these fluctuations are so complex that forecasting is still a very uncertain undertaking. One of the most serious problems in forecasting changes in business conditions at all levels is the lack of understanding of the forces responsible for change.

Reasonably satisfactory progress has been made in measuring changes in business activity and an imposing amount of statistical data on business conditions is now available, most of it being disseminated as frequently as monthly. This chapter is devoted to a description of those statistical time series that are of the most value to the businessman in measuring and forecasting changes in business conditions.

Statistical data on business conditions are frequently referred to as *business barometers*. Like the ordinary barometer that measures changes in atmospheric pressure, a business barometer measures changes in certain aspects of the business situation. On the basis of past observation, it is possible to forecast changes in the weather that are usually associated with certain changes in atmospheric pressure, but the barometer itself does not forecast. Similarly, if a business forecast is made, it must be made by an individual who knows that certain changes in business conditions follow a given combination of factors, as revealed by the business barometers. The basis of intelligent forecasting of business conditions is a complete and accurate measurement of the various elements in the economic system.

Business barometers are usually index numbers and employ one or more of the methods of adjusting, analyzing, and combining statistical series described in the preceding chapters. In fact, the executive is a consumer of statistical series after they have been subjected to a certain amount of processing. All the problems of statistical methods that have been discussed to this point may be used in the preparation of a series that is used to measure changes in some aspect of business conditions. Most business barometers are based on samples, with all the series collected and tabulated by the methods already described. Adjustments for trend and seasonal variation are made when needed, and the methods of combining unlike series into composite indexes are frequently used. In fact, the barometers described in the following pages represent the end product of a whole series of statistical operations. In most cases the businessman uses this end product without knowing what steps were taken in its preparation; one of the chief purposes of this book is to explain the methods by which this end product is prepared, so that it may be used with a better appreciation of its meaning.

In using published indexes of business activity, it is profitable to put the internal data of the business on a comparable basis with the external published series. This means that the important internal series should be reduced to indexes, and adjusted for seasonal variation and secular trend when necessary. In spite of the fact that the methods of analyzing time series have been widely used for thirty-five years, few businesses use them to analyze their internal data. Elaborate systems of records have been developed to provide detailed data on sales, production, shipments,

inventories, and costs, but it is still exceptional for a concern to measure trend, seasonal variation, and cycles in these series. Yet, if the internal series are to be related to external data now available, some such analysis should be performed.

For example, too many businessmen still adjust for seasonal variation by merely comparing one month with the same month a year ago, or the total to date with the total to the same date a year ago. Thirty-five years ago no other method was generally known, but that is not true now.

In Chapter 4 the statement was made that too many business executives do not know what statistical data are available in published form, and important decisions are often made without considering facts that could have been acquired with little difficulty. This statement is as true for business barometers as for any other type of statistical data.

MEASURES OF TOTAL ACTIVITY

Statistical series measuring total economic activity have become available only relatively recently, appearing much later than data measuring changes in the level of activity of individual industries and specific phases of business. A comprehensive measure of total business activity in the United States was not compiled until the latter part of the 1930's. Before that time businessmen wanting information on changes in total business activity relied on certain statistical series or combinations of series which were believed to *indicate* the change in total activity rather than to *measure* it. Some of these series are still very useful as indicators of business activity, even though more comprehensive measures are now available. The most important of these series will be described, not because they are of historical interest, but because they are still useful to businessmen.

Data on the number of freight cars loaded each week are published by the Car Service Division of the Association of American Railroads. Since movement of freight obviously is related to the level of business activity, changes in the number of cars loaded have been used as an indication of the changes in total business activity. In the same manner, changes in steel production indicate changes in total business activity, since the demand for steel varies directly with the level of industrial activity. It has been recognized for a long time that a major portion of the changes in total business results from changes in the industrial sector of the economy, so a statistical series that reflected changes in total industrial activity would be expected to be a good indicator of changes in the over-all economy. For this reason the figures on steel and pig iron production are still highly regarded as business barometers.

Among other series that reflect changes in business activity are department store sales, electric power consumption, postal receipts, the volume of checks going through clearing houses, and the volume of checks debited to individual accounts by banks. The acceptance of the fluctuations in these series as good indicators of changes in total business activity has been based on observation of the fact that, in general, their changes have coincided approximately with periods of expansion and contraction. Department stores sell a wide variety of goods, and changes in their sales would be expected to follow changes in the general level of consumer prosperity. Electric power is widely used by consumers and businesses, so it is not unexpected that the amount consumed should fluctuate with the volume of business and the level of consumer income. Since a very large portion of the mail transported is business mail, the volume of receipts by the postal system tends to vary with the level of business. A very high proportion of business is transacted by checks, so the variations in the total volume of checks charged to individual accounts reflect the changes in business activity.

All of the statistical series mentioned above reflect in one way or another the changes in business activity; but if the problem is to measure directly the total volume of business and economic activity, the factors included in the measurement should be rather comprehensive. It would be possible to define total economic activity in a number of ways, but three definitions that have been widely accepted are:

1. The total value of business transactions.
2. The total number of persons employed in all economic activity.
3. The total value of final goods and services produced.

Applying any of the three definitions on a monthly or quarterly basis is a large undertaking in the United States, due to the size and diversity of economic activities. However, the need for comprehensive measures of activity has been urgent enough to stimulate the collection of such data.

Bank Debits and Clearings

For a long time economists and business statisticians have used data on bank debits and clearings to measure changes in business transactions. It is possible to obtain data on bank clearings from as far back as the time that local clearing houses first began operation. Thus, this series is one of the first that was available for the study of business fluctuations. A major weakness of bank clearings as a measure of business transactions, however, is the fact that only checks going through a clearing house are recorded; checks deposited in the bank against which they are drawn, as well as checks cashed over the counter, do not appear in the record. Bank debits

are superior to clearings as a reflection of business transactions since the debits include all withdrawals, not just those going through a clearing house.

In 1918 the Federal Reserve System began to collect from a large number of cities weekly reports on debits to individual accounts and initiated the release of weekly and monthly series on national and individual city bases. Although some debits represented charges against savings accounts, the major portion represented charges against checking accounts. Since a large proportion of business payments is made by checks, total bank debits serve as a satisfactory measure of business transactions. Included in this total value of business transactions are transfers of property and speculative transactions, in addition to the production and distribution of goods. However, the same goods may be included several times, since payment may be made each time the goods change hands.

The weaknesses of measuring over-all economic activity by totaling the value of all business transactions paid by check have not been great enough to keep economists and statisticians from using this value as a measure of the business situation. However, as better measures of total business activity have been developed, bank debits have become less important in measuring changes in the level of business for the country as a whole.

When measures of regional business activity are desired, bank debits still represent one of the best statistical series readily available. The fact that they are accurately compiled and available soon after the end of the period covered, makes bank debits an extremely valuable measure of changes in local business conditions for areas as small as a city. Bank debits are used for measuring changes in regional business activity for exactly the same reason that they were used earlier in measuring business activity in the United States; namely, they are the best data available.

Employment

The statistical series showing employment represents basic information concerning the behavior of the economy, since labor is the most important economic resource. The extent to which labor is employed is a significant over-all measure of economic activity. The ability of the economic system to provide jobs is a criterion that everyone can understand, and the fluctuations in the number of unemployed can be considered a basic measure of the health of the economy.

There are some serious weaknesses in employment data as a measure of total business activity. Since the number of hours that employees work per week can be varied with changes in business conditions, the number of persons employed tends to be more stable than the volume of output

of goods and services. The number of man-hours worked is somewhat better than the number of persons employed as a measure of business activity, at least for short-term fluctuations. However, the secular trend in productivity per man-hour affects the validity of the data on man-hours as a measure of over-all production, especially in comparisons over a considerable period of time.

Statistical data on employment in the United States are available from two major sources: the Bureau of the Census and the Bureau of Labor Statistics. In addition to these two sources, the Department of Agriculture is a primary source of data on agricultural employment. The basic methods of compilation used by the Bureau of the Census and the Bureau of Labor Statistics are different, with the result that the data from the two sources give somewhat different information.

The Bureau of the Census, using the latest sampling techniques, collects data on the labor force from individual households. The results are published monthly in *Current Population Reports: Labor Force*. The data give the number in the labor force, the number employed, and the number unemployed. The number employed is classified as agricultural employment and nonagricultural employment, and all the categories are shown separately for men and women.

The series on employment and unemployment are major barometers of total business activity and are given wide publicity in current business publications. Important secondary sources that carry the series are the *Survey of Current Business*, the *Federal Reserve Bulletin*, and the *Monthly Labor Review*. Some of the series show strong seasonal fluctuations, and all of them reflect cyclical changes in business activity.

The Bureau of Labor Statistics collects information from a large sample of business firms on the number of persons employed in their establishments. Although there is a possibility that a person may be counted more than once by this method, it becomes possible to secure estimates of total nonagricultural employment that are grouped into broad industrial classifications, and into a very detailed manufacturing-industry classification. The estimates, also classified by states, give valuable data on the regional variations in business activity. Comparable data on hourly earnings, hours worked per week, and weekly wages are also compiled.

The same secondary sources that carry the Census data on employment also carry the data on employment compiled by the Bureau of Labor Statistics. The latter data are more important as sources of information for individual industries and regions than as over-all measures of economic activity, but the two sets of employment data tend to supplement each other; and together they give a rather complete picture of total business activity.

Gross National Product

The most complete measure of total economic activity in the United States is the gross national product estimate compiled by the Department of Commerce. The economic system is a highly complex organism, made up of many interdependent parts. The system is organized to produce goods and services that can be used by consumers; at the same time the economic system distributes incomes as payment to the factors of production employed in the production of the goods and services. The *gross national product* is defined as the total market value of all final goods and services produced during a given period, such as a quarter of a year or a year. It is important to observe the condition that the goods be in final form; otherwise it would result in counting more than once the value of raw materials or component parts to be used in further production. For example, the value of the tires sold to an automobile manufacturer to be included as original equipment on an automobile should not be counted, since the sale of the automobile to the consumer will include the value of the tires. However, a tire bought by an automobile owner to replace a worn-out tire would be considered part of the gross national product, since the sale was to the final consumer.

The output of the economic system consists of a great stream of goods and services, a large percentage of which are being sold for further processing before finally being purchased by the consumer. If the value of the goods and services were counted each time they were sold, the same duplication would occur that is present in the total for bank debits. It has been impossible to eliminate the duplication from the figures on bank debits, but the figures for gross national product do not have any of this type of duplication.

The gross national product is considered one of the best measures of total business activity because the total production of the goods and services in a given period is in reality the total business activity of that period. The gross national product represents the *value* of goods and services rather than the *physical volume* because of the great difficulty that would be encountered in compiling a comprehensive measure of physical volume. It is much easier to secure the data on total value; and by using price indexes to deflate the value aggregates, it is possible to compute a gross national product in constant dollars that is a measure of volume changes. Gross national product estimates in current dollars are compiled by quarters, but only annual figures are available in deflated form.

Even though the gross national product is the most comprehensive measure of total business activity, it suffers from some disadvantages. The fact that it is available quarterly but not monthly means that other

Table 20-1

GROSS NATIONAL PRODUCT OR EXPENDITURE: 1946 TO 1956

(BILLIONS OF DOLLARS)

	1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956
GROSS NATIONAL PRODUCT	209.2	232.2	257.3	257.3	285.1	328.2	345.4	363.2	361.2	391.7	414.7
1. Personal consumption expenditures	146.6	165.0	177.6	180.6	194.0	208.3	218.3	230.5	236.6	254.4	267.2
Durable goods	15.9	20.6	22.2	23.6	28.6	27.1	26.6	29.8	29.4	35.6	33.9
Nondurable goods	84.5	93.1	98.7	96.9	100.4	111.1	116.1	119.1	120.6	126.0	133.3
Services	46.2	51.3	56.7	60.1	65.0	70.1	75.6	81.7	86.6	92.8	99.9
2. Gross private domestic investment	27.1	29.7	41.2	32.5	51.2	56.9	49.8	50.3	48.4	60.6	65.9
New construction	10.3	14.0	17.9	17.5	22.7	23.3	23.7	25.8	27.8	32.7	33.3
Producers' durable equipment	10.7	16.7	19.1	17.8	21.1	23.2	23.1	24.3	22.5	23.7	28.1
Change in business inventories	6.1	−1.0	4.2	−2.7	7.4	10.4	3.0	.3	−1.9	4.2	4.6
3. Net foreign investment	4.6	8.9	2.0	.5	−2.2	.2	−.2	−2.0	−.4	−.4	1.4
4. Government purchases of goods and services	30.9	28.6	36.6	43.6	42.0	62.8	77.5	84.4	76.6	77.1	80.2
Federal	20.9	15.8	21.0	25.4	22.1	41.0	54.3	59.5	48.9	46.8	47.2
State and local	10.0	12.8	15.6	18.2	19.9	21.8	23.2	24.9	27.7	30.3	33.0

Items may not check exactly with totals because of rounding.

SOURCE: *Survey of Current Business*, July, 1957.

more current measures are needed as a preliminary measure of business activity. Some of the basic information used in compiling the gross national product is estimated and probably is not very accurate. But the quality of the data used continues to improve, and there is the possibility that estimates eventually will be made more frequently. In general, gross national product serves its purpose well in giving an over-all measure of business activity. It is well to remember that any measure of total business activity will fail to give detailed information on the functioning of specific portions of the economy. The purpose of a measure of total activity is to summarize many divergent trends. If a detailed analysis of separate components is desired, additional measures of the changes in these elements are needed.

COMPONENTS OF GROSS NATIONAL PRODUCT

The amount of gross national product is broken into component series that give detailed information on the various segments of the economy. Since any measure of total activity is an average of diverse fluctuations in the different parts of the economy, the individual component measures are in many ways more significant than the over-all measure. There is good reason to believe that the fluctuations in total business activity are the result of instability in certain sectors of the economy, while other parts of the economy are relatively passive. This means that an analysis of the fluctuations in business necessitates measuring the changes that occur in the various individual segments, particularly those that appear to be strategic in the sense that movements tend to originate there.

As stated before, the total gross national product represents the final output of goods and services, which is bought by many different purchasers. The most significant classification divides the total output among the following four groups of purchasers:[1]

1. Consumers (Personal consumption expenditures).
2. Business concerns (Gross private domestic investment).
3. Foreigners (Net foreign investment).
4. Governments (Government purchases of goods and services).

Table 20-1 shows this classification of the gross national product from 1946 to 1956, and a brief description of each category follows.

[1]The four categories of gross national product probably represent the most useful classification of the total, but there are many other subdivisions that are extremely valuable for many purposes. The complete presentation of all the classifications of the national income and product are published by the Office of Business Economics, United States Department of Commerce. A recent edition of *National Income Supplement to the Survey of Current Business* contains an authoritative explanation of the methods and fundamental concepts underlying the national income and product accounts, and gives detailed statistics dating back to 1929. The July issue of the *Survey of Current Business* each year brings the information in the latest supplement up to date. The February issue of the *Survey of Current Business* each year presents a preliminary tabulation of the national income and product for the previous year, although much of the detail is still not available at that time.

Personal Consumption Expenditures

The largest category of purchases is that taken by individual consumers. In 1956 personal consumption expenditures totaled $267.2 billion and represented 64.4 per cent of the total gross national product. This total expenditure included money payments by consumers, estimated rental value of owner-occupied housing, and a few other items that did not involve monetary payments, such as goods taken out of a business by owners for their own use.

Total consumer expenditures are classified into three groups:

1. Durable goods.
2. Nondurable goods.
3. Services.

This classification is significant because of the differences in stability of the groups. The greatest instability appears to be in the expenditures for durable goods. The reasons for this seem to be related to the fact that purchases of durable goods can be postponed in many cases if income is declining or shows an indication of declining. In other words, a reduction in income might more seriously affect the purchase of new cars than the purchase of gasoline. Since the business cycle is apparently generated by decisions of the different purchasers of goods and services, this inherent instability is important; and the statistical data on business conditions should be classified to show consumer behavior in purchasing the different kinds of consumer goods.

Gross Private Domestic Investment

The portion of the output of goods and services that is purchased by business concerns for use in their business is called *private investment.* Goods that are resold during the period are not included in this category, since the gross national product is the value of the goods bought by the final purchaser. The three major kinds of private investment are:

1. New construction.
2. Producers' durable equipment.
3. Changes in business inventories (either increases or decreases).

If a concern buys a new building or equipment to be used in the business, this represents a capital expenditure. A change in business inventories is an investment if the inventory increased, or a reduction in investment (negative investment) if the amount of inventory decreased. The money spent for goods placed in inventory is a business investment in exactly the same way that money spent for a new plant or equipment is an investment. Moreover, the volume of business inventories shows wide variations from period to period, and is therefore a dynamic factor in the business

situation. This means that the net change in business inventories is a very volatile figure and can exert an important influence on total investment.

Private investment in capital goods may be either gross or net, depending on whether or not allowance is made for the capital that wore out, or depreciated, during the period considered. Gross investment represents total capital expenditures and shows the volume of business in the capital goods industries. This information is generally more valuable in the analysis of business conditions than net investment, for the latter figure is subject to serious errors in estimating depreciation, causing it generally to be less accurate than gross investment.

Gross private domestic investment was $65.9 billion in 1956, or 15.9 per cent of the total gross national product. The importance of this category in the study of business fluctuations does not depend upon its size, but rather on the dynamic nature of investment. The importance of the investment segment of the economy will be discussed in more detail in the following chapter.

Net Foreign Investment

The smallest category in the gross national product is the net change in the international assets and liabilities of this country arising out of current transactions with foreign countries. Net foreign investment, which amounted to $1.4 billion in 1956, or 0.3 per cent of the total gross national product, is the sum of:

1. Domestic output sold abroad minus purchases by this country of foreign output.
2. Cash gifts and contributions received from abroad minus those sent abroad.
3. Production abroad by United States labor and property minus production in the United States by foreign-owned labor and property.

Exports that are financed by the government are counted as government expenditures, and those financed by individuals are considered consumption expenditures.

Government Purchases of Goods and Services

Government expenditures consist of purchases of goods and services by the:

1. Federal government.
2. State and local governments.

The fluctuations in expenditures of the federal government differ significantly from those of state and local governments, so it is important that this classification be made. Since federal government expenditures for national defense are subject to such wide variations, it is sometimes desirable to show this item separately from the total federal expenditures.

NATIONAL INCOME AND PERSONAL INCOME

The gross national product and its components are of major importance in the analysis of business conditions, but certain other series reflecting changes in national income are also useful. *National income* is the total income earned by the factors of production in producing current output. National income is somewhat less than the total gross national product, since the value of some items such as capital consumption allowances and indirect business taxes does not represent income to factors of production. The largest payment made to any of the factors of production is the compensation of employees. Other major factor costs are income of unincorporated enterprises, corporate profits, rental income, and interest.

Table 20-2 shows the computation of personal income for 1956 by making certain deductions from gross national product. Deducting depreciation and other capital consumption allowances from gross national product leaves *net national product*. The capital consumption allowances are part of the cost of the total output of the economy, but they do not represent amounts paid out to factors of production. When business taxes are deducted from net national product, the remainder is the national income (except for certain minor adjustments). Taxes paid are part of the gross national product since they are part of the cost of the goods produced; but government is not a factor of production, so amounts paid in taxes are not part of the national income.

Although national income represents the factor costs of production, it is not identical with the income distributed to individuals; some of the income belonging to the factors of production is not paid to individuals. Corporations do not always distribute all of their earnings as dividends. Social security payments are turned over to the government by business concerns and are not available to individuals as spendable income. These items must be deducted from national income in computing the amount of income paid to individuals.

Some individuals receive money payments that are not income in the sense that they are payments for productive service. One group of such payments, called transfer payments, includes social security payments, veterans' benefits, pensions, and certain other payments for services not currently rendered. Veterans' benefits may actually be payment for services rendered in the past, but they are not payment for a factor of production in the period in which they were paid. Government payments of interest to persons who hold government securities are not made for a factor of production. A third form of payments, business transfer payments, are charges against business production for which no return in the form of factor services is received. Bad debts and business gifts to nonprofit organizations are important business transfer payments. These three

Table 20-2

RELATION OF GROSS NATIONAL PRODUCT, NATIONAL INCOME, AND PERSONAL INCOME, 1956

	Billions of dollars	
Gross national product		414.7
Less: Capital consumption allowances		34.3
Net national product		380.4
Less: Indirect business taxes	35.0	
Other adjustments	1.8	36.8
National income		343.6
Less: Corporate profits retained	28.6	
Contributions to social security	12.3	40.9
		302.7
Plus: Government transfer payments	17.2	
Net interest paid by government	5.7	
Business transfer payments	1.3	24.2
Personal income		326.9
Less: Personal tax and nontax payments		39.7
Disposable personal income		287.2
Less: Personal consumption expenditures		267.2
Personal saving		20.0

SOURCE: *Survey of Current Business*, July, 1957.

groups of payments must be added to national income in the computation of the amount of income paid to individuals.

The total arrived at after making the above deductions from and additions to national income is *personal income*, which is the sum of all payments to individuals. This series is one of the most important of all the national income figures, since it represents the flow of purchasing power that has resulted from the workings of the economic system.[2]

[2]The personal income series are published monthly in the *Survey of Current Business*, instead of quarterly or annually as is the case with the other national income and product series. Figures for total personal income and considerable detail as to the source of the income are available for each state from 1929 to date in *Personal Income by States Supplement to the Survey of Current Business.*

After personal taxes and nontax payments that must be made to federal, state, and local governments are deducted from personal income, the remainder is *disposable personal income*. Disposable personal income represents the amount that individuals have available to spend or to save. Next, consumer expenditures are deducted, and the balance equals *personal saving*.

The data on national income and product, when published for periods less than a year, are given both with and without seasonal adjustment. Series for less than a year that are adjusted for seasonal variation are published as annual rates rather than as quarterly or monthly totals. For example, the values of personal income and the seasonally adjusted quarterly figures at annual rates for the first half of 1957 are as follows:

Table 20-3

PERSONAL INCOME: FIRST HALF, 1957

Quarters	Personal income, unadjusted (Billions of dollars)	Personal income, seasonally adjusted at annual rates (Billions of dollars)
I.....	82.9	337.7
II.....	85.1	342.4

SOURCE: *Survey of Current Business*, August, 1957.

MEASURES OF INDIVIDUAL BUSINESS PROCESSES

The following sections will describe a limited number of statistical series that show changes in individual industries and components of business activity. The series chosen represent a very small percentage of the total number of time series that are published in the United States. Many lists of the important business series have been compiled; and while some series are included on practically every list, there is always a substantial variation from one selection to another. The following selection should be considered as but one opinion of the important barometers of business activity. The *Survey of Current Business* has 40 pages in the section, "Monthly Business Statistics," and each page carries more than 50 different series. With this much information available every month, it is extremely important that a rather rigorous selection be made. Every published statistical series representing an individual industry might be considered of some importance in measuring the total business situation, but the relative importance of the different series varies greatly.

Industrial Production

The index of industrial production compiled by the Board of Governors of the Federal Reserve System measures the change in the physical volume of manufacturing and mining rather than the change in dollar value or prices. Separate indexes are published for a detailed classification of industries and groups of industries. The indexes, which are published monthly, are given both with adjustment for seasonal variation, and without the seasonal adjustment. The index of industrial production does not include production on farms, construction, or the output of public utilities.

The fluctuations in over-all business activity reflect to a considerable degree the variations in output of the industrial system. Manufacturing activity is geared rather closely to the demand for goods, with the result that changes in industrial activity are an integral part of any change in total business activity. This means that the index of industrial production is watched with great interest as a sensitive indicator of change in the business situation. Although some lists of business barometers include the index as one of the measures of total activity, it does not measure the extent of the change.

The classification by individual industries permits the businessman to follow the changes in strategic industries in considerable detail. Durable and nondurable manufacturers are shown separately, and a special category shows the fluctuations in major consumer durables. Mining activity is shown for each of the major minerals.

The index is expressed as relatives, with the average month of 1947 to 1949 equal to 100. It is available in most business periodicals that publish statistical data on business.

Business Sales, Inventories, and Orders

The Bureau of the Census compiles monthly data on sales, inventories, and new orders. Since these series are expressed in current dollars, they are influenced by fluctuations in the price level as well as volume of business. The information, which is available for major industry groups, supplies a rather complete picture of the current level of manufacturing, as well as furnishing a basis for projection into the future. The level of inventories at the end of the month, when compared to sales, indicates whether production is moving into distribution channels or piling up in the manufacturers' storerooms. The changes in the level of unfilled orders serve as a forecast of future production rates. The relationship between new orders received during a month and the sales for the month gives additional indication of the course of future volume of manufacturing.

These series are published in the *Survey of Current Business* and *Economic Indicators.*

Retail sales represent a major proportion of consumer expenditures, and the Bureau of the Census makes an estimate of retail sales each month from a sample survey. These estimates give an early indication of the direction in which the consumer expenditure component of the gross national product is moving. Estimates are also made for each major kind of business, with separate estimates for durable-goods and nondurable-goods stores. These estimates represent the sales of the two types of stores and are not the sales of durable and nondurable goods. Furniture and appliances sold by a department store are sales of durable goods, but these sales would be included in the total sales of department stores, which would be reported as sales of nondurable-goods stores. In spite of the fact that the classification does not exactly parallel the consumer expenditure groups in the gross national product, the retail sales are extremely valuable as indicators of the shifts in consumer spending. Since expenditures for durable goods represent one of the strategic elements in the business situation, the sales estimates are particularly important in measuring the changes in business activity.

At the end of each month the Bureau of the Census also estimates retail inventories, which give an indication of the movement of goods through the channels of distribution. The same survey also estimates sales and inventories of wholesale businesses. When the inventories of retail and wholesale establishments are added to manufacturers' inventories described on page 485, the resulting total of business inventories represents the amount of goods already produced that has not moved into the hands of consumers. By watching the changes in this total from month to month, it is possible to anticipate fairly accurately the change in inventories that will appear in the quarterly estimates of the gross national product.

Construction

The expenditures on new construction represent one of the most important segments of the gross national product. Together with the expenditures for producers' durable equipment, new construction accounts for most of the gross private domestic investment; and the fluctuations of the expenditures for these capital goods are highly significant factors in the over-all level of business.

The Department of Commerce publishes a monthly series showing the dollar value of new construction put in place, classified, on the basis of ownership, as private and public. Construction put in place represents the amount actually spent on construction work during the period covered. For example, a house costing $25,000 may require five months to build, so

the value of new construction put in place each month would be one fifth of the total, or $5,000. Residential construction includes nonhousekeeping facilities such as hotels and dormitories, as well as dwelling units. Nonresidential construction includes building and nonbuilding structures such as dams, reservoirs, docks, highways, airfields, and utility lines.

The estimates of construction put in place are derived by converting data on work started into estimates of new construction activity for private and nonfederal public construction. Progress reports on federal projects are received from the various agencies. Since the data are derived from many different sources and estimates are made by many different methods, there is probably considerable error in the series; and conclusions probably should not be drawn from small month-to-month changes. It is important to remember that a series showing new construction put in place is entirely different from one showing new work started.

Two sources of information on new construction started have been used for many years. The F. W. Dodge Corporation compiles the number of building contracts that have been awarded, and the Bureau of Labor Statistics collects reports from building inspectors in the areas of the country that require permits for building.

The series on contracts awarded are compiled by the F. W. Dodge Corporation from field reports of individual projects. Recently the coverage of these series has been extended to all 48 states; previously only 37 states were included. Number of projects, floor area, and valuation are included for all types of construction, except that the classifications of public works and utilities include certain types of projects, such as highways, pipelines, and subways, which cannot be measured in terms of floor area. The major classifications of kinds of construction are published in the *Survey of Current Business*. A more completely itemized record of contracts awarded and a bimonthly reporting service, listing construction contracts awarded by type and region, are available from the F. W. Dodge Corporation.

A seasonally adjusted index of the value of construction contracts awarded is compiled by the Board of Governors of the Federal Reserve System. Two separate indexes are compiled — one for residential and one for all other types of construction. These indexes are published in the *Federal Reserve Bulletin* and the *Survey of Current Business*.

The Bureau of Labor Statistics compiles estimates of the number of new permanent nonfarm dwelling units on which construction is started each month, basing the estimates on building permits issued. The valuation of building authorized by permits is published by type of building construction by regions. These series are published in *Construction Review*, and some of the series are available in the *Survey of Current Business*. The

number and valuation of building permits issued are available by cities and represent an important source of information on local business conditions.

Plant and Equipment Expenditures

Since World War II the Department of Commerce and the Securities and Exchange Commission have been compiling data on capital outlays and anticipated capital expenditures by business. Each quarter a sample of business firms, covering all private industry except agriculture, is asked to report plant and equipment expenditures for the past quarter and planned expenditures for the coming two quarters. At the end of each year these firms are asked to report their anticipated expenditures for the next year. Quarterly estimates based on these reports are available back to 1945. The anticipated expenditures are reported regularly in the *Survey of Current Business.* The estimated expenditures for a year in advance have been reasonably accurate and represent an important series for the business forecaster. Although the data have been available for too short a period of years to determine exactly how reliable they will be, in this short period the series has become one of the most widely used barometers. Its use in making business forecasts will be discussed further in Chapter 21.

Commodity Prices

The average level of the prices of groups of commodities is a significant barometer in the analysis of general business conditions. The level of prices is probably as important a factor in business decisions as business volume. The technical problems of constructing price indexes were discussed in Chapter 16; this section will describe the most important published indexes that are available to the businessmen.

Wholesale prices. The index of wholesale prices compiled by the Bureau of Labor Statistics is designed to measure the changes in the composite of price movements in primary markets. The term *wholesale* refers to sales in large quantities rather than to the prices paid by retailers to wholesalers or jobbers. The prices used in constructing this index represent the first important commercial transaction for each commodity. For manufactured goods, prices are the selling prices that have been secured from a sample of manufacturers. For raw materials, prices are generally those that are quoted on the organized exchanges or central markets.

Approximately 2,000 commodities are included in the index, which is computed by multiplying the current price relative by the assigned value weight and dividing the sum of these products by the aggregate of the value weights. The weights used since 1947 are based on the value of primary market transactions in 1947. The monthly average for the three years 1947 to 1949 is the base of the index.

Indexes are published showing the changes in commodities classified according to the following broad commodity groups: farm products, processed foods, and commodities other than farm products and foods (basically industrial commodities). Each of these major groups is broken down into smaller commodity groups. It is almost inevitable that a special problem will require a commodity classification not provided in the index, but a great deal of detail is given and the series are extremely important barometers.

Retail prices. Consumer expenditures are made largely at retail, so the fluctuations in retail prices become an important element in the business situation. The Consumers' Price Index, which is issued monthly by the Bureau of Labor Statistics, measures the average change in the prices of goods and services purchased by families of wage earners and salaried clerical workers living in cities. It is important to note that the index undertakes to measure changes in prices only, not the change in consumer expenditures, which would also reflect changes in patterns of consumption.

The index is constructed so that it measures changes in prices by comparing the total amounts of money required from month to month to purchase a fixed quantity of goods and services. In representative cities, surveys of how families spend money were made to determine the quantities of each of the goods and services that are used as weights.

Agricultural prices. The Department of Agriculture compiles an index of prices received by farmers based on prices paid for farm products in local markets. These prices represent the first prices at which the products change hands, as distinguished from the prices in the central markets used by the Bureau of Labor Statistics. Index numbers are published monthly with the five-year period 1910 to 1914 representing the base.

The Department of Agriculture also compiles an index of prices paid by farmers for commodities used in family living and for production. This index is now known as the *Parity Index*, and the formulation of major agricultural policies is related to its fluctuations. Both of the above indexes are published monthly in *Agricultural Prices*, a release of the Department of Agriculture. The *parity ratio*, the ratio of prices received to prices paid, is published with the two indexes. This ratio is widely used as a barometer of the economic status of farmers.

Financial Barometers

The cyclical fluctuations of business, to a considerable extent, are brought about by the fact that purchasing power in the form of credit is extremely elastic and can expand or contract to meet the needs of business. The expansion and contraction in the volume of credit used by business

can be expected to reflect the fluctuations in the volume of business; but, in addition, credit is one of the strategic forces in the business situation. Interest rates (the price of credit) are significant to the businessman for the same reason that all prices are important. Other financial aspects of business such as security prices, new security issues, number of new business firms incorporated, and business failures, are all significant barometers of business. Brief descriptions of some of these series are given below. The *Federal Reserve Bulletin* and the releases of the Board of Governors of the Federal Reserve System are basic sources of financial data.

Bank loans. The volume of credit outstanding reflects the level of business operations, and changes in the amount of short-term credit granted to business concerns should be taken into consideration in analyzing the business situation. The *Federal Reserve Bulletin* publishes monthly data on loans and deposits of all banks, based on reports from member banks and estimates for nonmember banks. Supplementing the monthly reports for all banks are the weekly statistics compiled for a group of approximately 400 member banks referred to as the "weekly reporting member banks." These banks are not a random sample of banks; they account for more than half of the loans and investments of commercial banks. The monthly reports are intended to reflect promptly the developments in demand for bank credit, and the data are shown in considerable detail. Classifying loans into the three categories, "commercial, industrial, and agricultural," is particularly valuable as a measure of business trends.

More detailed information on business loans is available in the *Federal Reserve Bulletin* and in a weekly release showing changes in commercial and industrial loans, classified into eleven categories according to the industry of the borrower. These data are tabulated from a sample of the weekly reporting member banks. About 210 banks are included in this sample, but they hold over 90 per cent of the total commercial and industrial loans of all weekly reporting member banks and nearly 70 per cent of the loans of all commercial banks.

Consumer credit. Credit used by individuals to finance the purchase of commodities and services for personal consumption, or to refinance debts originally incurred for such purposes, should be taken into account in the analysis of consumer expenditures. Credit is particularly important in the market for expensive consumer durable goods such as automobiles. The *Federal Reserve Bulletin* publishes monthly data on the amount of consumer credit outstanding at the end of the month as well as the amount extended and repaid during the month.

The relationship between disposable consumer income, consumer credit, and consumer expenditures is a significant factor in the analysis of changes

in business conditions. The correlation between consumer expenditures and income will be high, but when consumer credit increases substantially, for example, it makes possible a greater rise in expenditures than can be explained by the level of income. After analyzing the level of consumer spending that is expected to be associated with the existing level of income, it is wise to modify this estimate to allow for any change in the volume of consumer credit. If an expansion of consumer credit is expected, sales may be higher than the forecast on income would indicate. Conversely, when consumer credit is expected to contract, some of the consumer income will be absorbed and the level of expenditures may be less than was expected.

Consumer finances. The subject of consumer credit is part of the broader subject of consumer finances, in which income, savings, credit, and expenditures are all related. Income, expenditures, and savings are included in the national income and product accounts, but a considerable amount of additional information is available from other sources. One of the most significant statistical summaries on this subject is the annual Survey of Consumer Finances. This survey has been conducted since World War II in the early part of each year by the Board of Governors of the Federal Reserve System in cooperation with the Survey Research Center of the University of Michigan.

The 1957 survey included interviews with approximately 3,000 spending units (related persons living together who pool their incomes) in 66 sampling areas throughout the United States. Information was secured on the financial position of the respondents, including income received in the past year, past change in income rate, and liquid assets owned. Questions were asked concerning expected changes in income, and planned expenditures for durable consumer goods (housing, home improvement, new and used automobiles, furniture, and appliances). The results of the survey are published in the *Federal Reserve Bulletin* during the first half of the year, and are widely quoted in business publications.

Real estate credit. The volume of mortgage debt outstanding is published monthly in the *Federal Reserve Bulletin*, together with a detailed classification by type of property mortgaged and the type of mortgage holder. Since the purchase of real estate is financed to a great extent with long-term loans, real estate credit is a vital element in the construction industry. Tightening credit may slow down building, while plenty of credit available at favorable rates is a stimulating factor.

Security issues. Information on the financing of capital expenditures is necessary for the understanding of changes in business activity. The national income and product accounts indicate the amount of business savings that is available for financing the expansion of fixed assets, but the

information on new securities issued gives more specific information on the money made available. The Securities and Exchange Commission compiles monthly estimates of the gross proceeds from new securities issued in the United States. These estimates are published in the Securities and Exchange Commission *Statistical Bulletin*, classified in various ways, such as by industries, type of security, and proposed use of proceeds.

Security prices. The trading in corporate securities already issued does not provide funds for business expansion, but is merely a change in ownership of the securities. However, the amount of trading and the prices at which stocks sell have long been considered important indicators of business expectations. The prices of common stocks are particularly sensitive to changes in the prospects for profits. Since the common stockholders have the residual equity in a corporation, any rise in profits means an increase in the earnings going to the holders of common stock, while a decrease in profits means a decline in the earnings of the common stock.

Measurement of changes in common stock prices have been published for many years in the form of index numbers based on a sample of stocks. These indexes are rather generally referred to as *stock market averages*, which is not an illogical title since all index numbers of prices are averages of individual price series. Averaging the prices of a number of stocks eliminates the unusual movements that may be present in the behavior of individual stocks.

A number of indexes of common stock prices are compiled. The Dow-Jones averages, which have been compiled for many years, give a continuous record of the fluctuations in common stock prices. Standard and Poor's Corporation compiles index numbers that give information on the fluctuations of detailed industry groups. Standard and Poor's indexes are expressed as percentage relatives with 1935 to 1939 equal to 100, while the Dow-Jones indexes are published as dollars per share. The day-to-day fluctuations of these indexes are widely publicized in newspapers, newscasts, and financial journals.

STUDY QUESTIONS

1. Some books on business forecasting have distinguished between "business thermometers" and "business barometers" on the basis of the ability of statistical series to forecast changes in business conditions. The series that tended to turn ahead of changes in business activity were called "barometers," while those that merely measured changes in business activity were called "thermometers." Do you consider this distinction worth making, rather than calling all business data "barometers"?

2. Why is it worth while to express the internal data of a business in the form of index numbers that can be compared with published business barometers?

3. Is it logical to use a statistical series for a single industry, such as freight cars loaded or steel production, as a measure of changes in total business activity? Give reasons for your answer.

4. The text states that total economic activity might be defined in the three following ways: (1) total value of business transactions, (2) total number of persons employed, and (3) total value of final goods and services produced. What are the advantages and disadvantages of using each of these definitions of total business activity?

5. Why did business statisticians use bank clearings as one of the first measures of total economic activity? Although bank debits are considered a better measure of business activity than clearings, they were not used as statistical measures until much later than the first use of clearings. Why was this the case? Why are bank debits considered a better measure of total business activity than clearings?

6. Why are the figures for the gross national product considered such a good measure of business activity?

7. Personal consumption expenditures represent the largest single component of the gross national product. Does its size automatically make it the most important segment? Should other factors also be considered? Explain.

8. The three major subdivisions of gross private domestic investment are important measures of business activity because they are subject to rapid and substantial change. Why are the three divisions of gross private domestic investment subject to sharper fluctuations than consumer expenditures?

9. Why is it important to separate federal government expenditures from state government expenditures?

10. Why are the figures for disposable personal income particularly important in the study of consumer markets?

11. Compare the relative importance of the series on national income and the gross national product as measures of economic activity.

12. Before the gross national product estimates were published, many business analysts used the index of industrial production as a measure of changes in total business activity. How good a measure of the over-all changes in business activity is the index of industrial production?

13. What is the significance to the businessman of data on unfilled orders held by manufacturers?

14. What are the relative advantages of the series showing construction started (such as permits issued or contracts awarded) and construction actually put in place during a month?

15. Is it reasonable to expect that business concerns can report accurately on their expenditures for new plant and equipment in coming months?

16. Information on changes in the level of prices is considered of major importance by business analysts. What is the reason for this concern about changes in prices?

17. Fluctuations in the prices of common stocks on the organized exchanges are watched with a great deal of interest by business analysts. Why is so much attention paid to this barometer?

18. Why is information on the volume of the various kinds of credit outstanding an important factor in studying the business situation?

PROBLEM

Compile a statistical summary of construction since 1956 in as much detail as possible. This should include the volume and value of all types of new buildings started (either permits issued or contracts awarded), construction put in place, construction costs, financing, and any other phase of the industry that is considered significant. Present the data in properly designed statistical tables and draw charts of the most significant series.

From the information compiled write a brief summary of the course of construction activity since 1956.

21 BUSINESS FORECASTING

It has been frequently stated in previous chapters that one of the chief uses of statistical data in the form of time series is to forecast the level of business of some future period. Forecasting business change involves more than an analysis of statistical data — it also embodies the prediction of economic change, such as secular trend, seasonal variation, or cyclical movements, and a consideration of cause and effect. These three types of fluctuations and the methods of measuring each have been described in the preceding chapters. In this chapter, methods of forecasting the three types of fluctuations are presented.

Broadly speaking, the forecasting of business fluctuations, whether it be for secular trend, seasonal variation, or cyclical movements consists of the following steps:

1. *Understanding why changes in the past have occurred.* If an attempt is made to forecast business fluctuations without understanding why past changes have taken place, the forecast will be purely mechanical, based solely upon the application of mathematical formulas, and subject to serious error. On the other hand, if an understanding of the fluctuations has been secured, there is at least a basis upon which to start building a projection.

2. *Determining which phases of business activity must be measured.* After it is known why business fluctuations have occurred, or if there is a reasonable supposition, it is necessary to measure certain phases of business activity in order to predict what changes will probably follow the present level of activity. In this chapter it will be seen that, after certain phases of business activity, such as the components of gross national product, have been measured and projected, it is possible to predict rather successfully the total business activity.

3. *Selecting and compiling data to be used as measuring devices.* The business barometers described in Chapter 20 and some of the examples presented in the chapters dealing with time series analysis form the basis of statistical data that is available to measure the fluctuations in business activity. There is an interdependent relationship between the selection of statistical data and the determination of why business fluctuations occur. Statistical data cannot be selected and compiled in an intelligent manner unless there

is a sufficient understanding of business fluctuations; likewise, it is important that the reasons for business fluctuations be stated in such a manner that it is possible to secure data that are related to the reasons.

4. *Analyzing the data.* In this step the data are analyzed in light of one's understanding of the reasons why change occurs. For example, if it is reasoned that a certain combination of forces will result in a given change, the statistical part of the problem is to measure these forces and, from the data available, to draw conclusions on the future course of business. The methods of drawing conclusions may be called forecasting techniques, which represent any one of a large number of analytical devices for summarizing data and drawing inferences from the summaries.

SECULAR TREND

During the decade of the twenties the forecasting of secular trend received much less attention than the forecasting of the business cycle, but since World War II there has been an increasing interest in trend forecasts. While the war was still going on, the Department of Commerce published a report, "Markets After the War," which pointed out the importance of trend forecasts to the businessman. However, the thinking of business economists at the end of the war was almost entirely concerned with the study of the business cycle, with particular attention focused on the expected postwar depression. Gradually it became evident that the expected depression was not developing, but for a while it was not generally realized how rapidly the economy of the United States was expanding. Instead of there being an excess of productive capacity at the beginning of the postwar period, it gradually became clear that the growing population and the greatly increased productivity of labor were resulting in a volume of business that was continuing to grow at a rate that had been entirely unforeseen. It was primarily the rapid rate of growth that stimulated increased interest in forecasting the secular trend of business. This was a natural development, since the more rapid the change, the greater the need for a forecast of the future.

Although considerable interest in forecasting secular trend had developed by 1950, the expansion of industry during the Korean War stimulated an even more intense interest in the future trend of business. There was a fear that, at the end of the rearmament expansion program, industry would be left with an excess of capacity which would bring on a major depression. This feeling seemed to voice again the opinions that had existed toward the end of World War II. However, the projection of the secular trend gave a much more encouraging picture of the potential of

business, and still greater activity in studying the potential growth of business was stimulated, as evidenced by the forecasts of secular trend which were published in 1952.[1]

Later studies of the secular trend have been made, particularly for specific industries. A typical study was "America's Demand for Wood, 1929–1975," prepared by Stanford Research Institute for Weyerhaeuser Timber Company in 1954. This study analyzed the factors that would affect the long-range demand, supply, and price of forest products, and projected wood requirements for the years 1960, 1965, 1970, and 1975.

Any industry that is showing a substantial rate of growth needs to keep a forecast of that growth available for long-range planning purposes. This is particularly true of industries with a large investment in fixed assets, since the expansion of capacity and gearing for full production normally require a considerable period of time after the decision to build additional facilities is made.

The problem of distinguishing between an increase in demand that is the result of growth and one that is temporary, resulting from cyclical expansion, sometimes requires careful analysis. If growth is incorrectly interpreted as a cyclical expansion that will not be permanent, it is likely that business will be permanently lost to competitors. But if a temporary cyclical expansion is mistaken for growth, it is possible to make serious mistakes in building unneeded capacity.

As a part of long-range planning in industries with large capital investments, the budgeting of capital expenditures represents policy making of great importance. Not enough is known about the process of making decisions on capital expenditures, but recent interest in the use of secular trend forecasts in planning for the future of an individual business reflects a growing belief that statistical methods can make a significant contribution to the solution of management's problems. The contribution of statistical methods to the forecasting of secular trend was discussed in Chapter 13 on pages 314 and 315. In that chapter it was pointed out that the only basis for a forecast of the future is a knowledge of past happenings and that the first step in a forecast of secular trend is to measure the trend in the past.

[1] *Resources for Freedom*, A Report to the President's Materials Policy Commission, U. S. Government Printing Office, Washington, D. C., June 1952. (Generally known as the "Paley Report.")

Gerhard Colm, "The American Economy in 1960," National Planning Association Washington, D. C., 1952.

"Markets After the Defense Expansion," United States Department of Commerce, U. S. Government Printing Office, Washington, D. C., 1952.

"The Sustaining Economic Forces Ahead," Materials prepared for the Joint Committee on the Economic Report by the Committee staff, 82d Congress, 2d Session, U. S. Government Printing Office, Washington, D. C., 1952.

Forecasting the Trend of Total Gross National Product

The problem of forecasting secular trend will be illustrated with a forecast of the trend of total gross national product made by projecting certain basic factors that contribute to the trend of the total. The basic elements in the total level of economic activity can be reduced to the following factors: population, labor force, hours worked, and productivity per man-hour. In this example, the secular trend of each of these elements will be projected to 1975; and from these projected values a forecast of the total gross national product will be made.

The projection of the labor force is based on a study of the forecast of population in 1975, made by the Bureau of the Census in the summer of 1957. The Bureau of the Census makes forecasts of the population based on several assumptions as to the rate of increase.

A total population of 221,000,000 was selected although it was somewhat less than the highest projection. A projection of present age groups indicated that the population 14 years of age and over in 1975 would be approximately 154,687,000; and it was estimated that 55.2 per cent of this group would be in the labor force. In May, 1957, 58.8 per cent of this group was in the labor force; but it was assumed that, with the increasing percentage of the population in the older age groups, a somewhat smaller percentage of all persons over 14 years of age would be working or looking for work. Increased social security benefits would also tend to influence older persons from seeking employment. Further, with an increasing percentage of the population going to college, there would be some reduction in the proportion of young persons working.

A total of 4,000,000 in the armed forces was assumed, leaving a civilian labor force of 81,387,000. The normal rate of unemployment, not attempting to forecast whether or not 1975 would be a period of prosperity or recession, was estimated to be 4 per cent of the total labor force, which was approximately the postwar average. With 3,415,000 unemployed in 1975, the number employed, not including the armed forces, was estimated to be 77,972,000.

Total civilian employment of 77,972,000 was classified into three major groups: government, agricultural, and nonagricultural private employment. Civilian employees of the federal government have averaged approximately 3.5 per cent of the labor force during the postwar period, but a rising trend in the percentage of the population employed by the federal government suggested that 4 per cent be projected for 1975. State and local government employment was projected at 6.5 per cent, which was very close to the postwar average.

Agricultural employment in 1975 was projected on the basis of the need for farm products. The per capita consumption of agricultural products in the postwar period was estimated at $136 (in 1953 prices). No increase in per capita consumption in 1975 was assumed. With a projected population of 221,000,000 in 1975, the total value of agricultural production in 1953 dollars would be $30 billion. The labor force needed to produce this amount of farm products in 1975 was estimated to be 5,555,000. The value of farm products per man-hour produced in 1953 was $1.437. Between 1910 and 1953 the output per man-hour had increased an average of about 2 per cent per year; but, in view of the improving technology in agriculture, it was assumed that production per man-hour would increase at an average rate of 2.5 per cent each year until 1975. This increase would raise the output per man-hour to $2.473 in 1975. The average work week in the postwar period was approximately 47 hours, but it was assumed that it would drop to an average of 42 hours each week by 1975. This would amount to 2,184 hours each year for each employee; and, with an output per man-hour of $2.473, the value of the output of each employee in agriculture would be $5,401 (in 1953 dollars). It was stated before that the output needed in 1975 would be $30 billion, which would require 5,555,000 employees.

The employment not included in the four groups described above (armed forces, federal government, state and local government, and agriculture) represents the total number of persons employed in private nonagricultural business. The classification of the total labor force projected for 1975 is given in Table 21-1.

Table 21-1

AN ILLUSTRATIVE PROJECTION OF THE 1975 LABOR FORCE

	Number of employees
Government employment:	
Federal	3,255,000
State and local	5,290,000
Total government employment	8,545,000
Agricultural employment	5,555,000
Nonagricultural private employment	63,872,000
Total civilian employment	77,972,000
Unemployment	3,415,000
Total civilian labor force	81,387,000
Armed forces	4,000,000
Total labor force (55.2 per cent of population 14 years of age and older)	85,387,000

Table 21-2

AN ILLUSTRATIVE PROJECTION OF 1975 GROSS NATIONAL PRODUCT

		Billion 1953 dollars
Government gross product:		
Military (4,000,000 × $4,319)		17.3
Federal civilian (3,255,000 × $4,935)		16.1
State and local (5,290,000 × $4,916)		26.0
Total government gross product		59.4
Private gross product:		
Agricultural		30.0
Nonagricultural:		
Average annual hours worked per employee (38 hrs. × 52 weeks)	1,976	
Employment	63,872,000	
Number of man-hours worked per year	126,211,000,000	
Output per man-hour	$4.826	
Total nonagricultural gross product (126,211,000,000 × $4.826)		609.0
Total gross national product		698.4

From the projected 1975 labor force it is possible to project the basic factors of gross national product. This illustrative projection is shown in Table 21-2, which gives gross national product in 1953 dollars. The military forces were projected at 4,000,000, and it was estimated that average income in 1975 would be approximately 50 per cent higher than in 1953, or $4,319 per year. Federal civilian employees were estimated to receive an average income of $4,935 per year, and state employees, $4,916. Both of these average incomes were also 50 per cent higher than in 1953. When the average income for each employee was multiplied by the number of government employees, the total government gross product was estimated to be $59.4 billion in 1975. The agricultural product was estimated to be $30 billion, based on the needs of the estimated 1975 population.

The projecting of nonagricultural private gross product requires more computations, based on the information summarized in Table 21-2. The average number of hours worked by each person in 1975 was estimated to be 1,976, based upon an average of 38 hours each week for the 52 weeks of the year. Since the time taken as vacation would be deducted from the total hours worked in a year, the length of the work week would be somewhat more than 38 hours. The trend in the hours worked each week has been declining steadily during the twentieth century, and it was assumed that this trend will continue through 1975. With an estimated employment of 63,872,000, the total number of man-hours worked in 1975 was computed to be 126,211,000,000.

After computing an estimate of the total number of man-hours worked in nonagricultural private business, it was necessary to estimate the output per man-hour in 1975. The output per man-hour is one of the most fundamental statistical series in the measurement of secular trend of overall economic activity The value of goods and services that are produced for each hour of work is of fundamental importance in the economy of any country. The economic progress that has been made depends primarily upon this one factor. As better methods of production and more efficient uses of capital have been employed, the output for each hour of work has increased. Continued increase in the standard of living depends upon a continuation of the increase in the productivity of labor. One estimate places the value of goods and services produced by each hour of labor in private nonagricultural industry in the United States during 1953 at $2.804. This amount was computed by dividing the total value of private nonagricultural gross product by an estimate of the number of man-hours used in producing this total value.

The past trend of productivity per man-hour has shown an average increase of nearly 3 per cent per year. A conservative projection of the productivity at an annual rate of 2.5 per cent each year gave a value of $4.826 in 1975. When this factor was multiplied by the number of man-hours estimated for 1975, the total value of nonagricultural private gross product was $609.0 billion. Since no allowance was made for a change in the price level, this estimate is given in terms of 1953 purchasing power. The total of all the separate estimates of parts of the gross national product was $698.4 billion.

It is possible to raise valid objections to any of the projections, since no matter how many historical data are used in projecting a trend, it is still characterized by a considerable element of subjective judgment. The technique of splitting the forecast into a number of factors and projecting each factor separately does not make the method any less subjective, but such a technique does give each person using the forecast an opportunity

to evaluate the individual projections. For example, a different projection of the number of hours in the work week would change the estimate. Changes in the trend of productivity, percentage of the population in the labor force, or any of the other factors would be reflected in a different estimate of the gross national product. In fact, one way in which this type of analysis can be used is to experiment with different values of the factors to determine what effect the changes would have on the final estimate.

Forecasting Components vs. Forecasting the Total

Table 20-1 (page 479) shows a classification of gross national product from 1946 to 1956 by components, consisting of the four major purchasers of the goods and services produced by the economy: consumers, businesses, foreigners, and governments. Instead of forecasting the total gross national product by breaking it into factors to be projected individually, it would be possible to forecast separately each of the major components or even the subdivisions of the major components. No example of such a procedure is given in this chapter since the projection of the total as illustrated in the preceding section seems to be the most effective method. If forecasts of the individual components are wanted, it is better to secure them by breaking down the total forecast, either by assuming that the percentage distribution by components will remain unchanged or by developing a forecast of the relationship between the various components of the total.

Forecasts of the components of gross national product are generally more valuable than the total in making projections for individual industries. For example, the demand for lumber would logically be related to the forecast of the construction component; household appliances would be expected to show a correlation with the durable consumer goods component; and machine tools should behave in much the same way as producers' equipment. In other words, the technique of long-range projections considers economic growth to be an interrelated series of components; and once the total level has been projected, the various subdivisions of the total can be related in a consistent manner.

SEASONAL VARIATION

Because the forecasting of seasonal fluctuations is chiefly a problem of computing an accurate measure of the average seasonal pattern, this subject was discussed in Chapter 14 as part of the problem of using an

index of seasonal variation. If the reasons for the seasonal fluctuations are known, it is usually not difficult to reach a conclusion as to whether or not the past seasonal pattern will continue in the following year. Gradual shifts in established seasonal patterns will be detected by computing an index of changing seasonal variation. Sudden shifts are somewhat in the nature of erratic variations and can normally be explained only after a study of the specific circumstances involved. When the shift is understood, it is usually possible to project the new seasonal pattern; but there is very little opportunity to develop general rules for making such a forecast if each change in seasonal pattern is a special case.

The probability that a seasonal pattern will continue in the future is basically related to the reason for the pattern, and since the causes of seasonal fluctuations in economic activity are fairly well understood, the forecasts of seasonal variation that are based on carefully observed past behavior are reasonably accurate. It is to be recommended that an objective, quantitative method of measuring seasonal fluctuations be used instead of subjective judgment, although it is possible for a businessman, through long experience and observation, to acquire a reasonably accurate knowledge of the seasonal patterns in a business.

CYCLICAL FLUCTUATIONS

A forecast of the cyclical fluctuations of business is a basic element in a wide range of business decisions, since the variations in the level of business volume have a profound effect on practically all phases of business. Budgeting is a systematic method of relating the forecasts to the various problems of the business and bringing those problems into an integrated form. The budget for the coming year represents a formal statement of the forecast upon which business plans are based and a statement of the operations that are to be undertaken in light of these forecasts and plans. Whether the basic forecast of the course of business for the period is stated or is merely implied in the plans that are made, business management makes a forecast of the business cycle in the mere act of making commitments to be carried out in the future. Making any decision that must be carried out in the future, such as ordering a given supply of goods, hiring employees, arranging for a line of credit, or scheduling advertising, implies a specific forecast of the level of business activity in the future.

There is good reason for stating the forecast and accepting it as the basis for other decisions after taking maximum care to insure that the forecast is a reasonable one. There have been many cases in which the

forecast could have been recognized as unsatisfactory if it had been stated specifically at the outset. A basic element such as a forecast of cyclical fluctuations that has such far-reaching influence on business decisions should not be taken for granted, but should be given careful thought to be sure it is the best possible forecast that can be made from the available information.

Forecasting Cyclical Fluctuations

The attempt to develop a systematic method of forecasting the cyclical fluctuations in business has occupied the attention of a large number of persons over a long period of time. The original goal of discovering a simple formula that would give several months' advance warning of changes in economic conditions has generally been abandoned with the realization that the problem is extremely complex and that a simple solution is probably not to be found. However, the search for a forecasting formula has brought about the development of methods of measuring changes over a period of time, and has stimulated the collection of better data to measure the changes in business activity.

Much study by economists has been aimed at understanding the causes of the business cycle, and all advances in this field should be taken into account by the business forecaster. A very substantial body of statistical data is available and economists now have a reasonably adequate understanding of the causes of the cyclical fluctuations in business activity. One goal of the economist in the study of the business cycle is to provide a basis for doing something to control cyclical variations. The wide fluctuations in employment and output are a serious flaw in our system of free enterprise, and many economists are hopeful that these cyclical fluctuations can be brought under control. The businessman must accept the fluctuations of the business cycle as one of the forces over which he has no control, but a force that must always be taken into account in making decisions. The job of the business statistician is to make the most effective possible use of existing information and understanding of the economic system to improve the decisions made by businessmen.

Lead Series

The cycles of most economic series are related to the cycles of total business activity — they expand in general when business is expanding and contract when business is contracting. Those series that turn up or down ahead of the measures of total business activity are known as *lead series,* while those that turn up or down afterwards are called *lag series.*

Probably the simplest approach to the construction of a forecasting formula is to find a statistical series that will indicate the course of business activity by turning up or down several months ahead of the measures of total business activity. Although it is possible to find a lead series for almost any past period of time, no series has been *consistently* successful in forecasting the turns of total business activity. One of the long-time favorites that has shown a tendency to lead has been the prices of common stocks. Other lead series that have been popular are: indexes of prices of commodities that are sensitive to changes in demand, such as scrap steel; unfilled orders or new orders, particularly for capital goods; production of capital goods, such as machine tools and housing; consumer durable goods; new incorporations; business failures; and hours worked per week.

An extension of the use of a lead series as a forecasting formula would be to average a number of series that have tended to turn early, with the hope that the average would consistently turn early even though all of the component series did not always indicate the turn. However, the forecasting series based on the averages of several series has not been any more successful than individual series in predicting the turning points of the business cycle.

The failure of various forecasting indexes to give reliable forecasts of changes in total business activity does not necessarily mean that the study of lead series has been wasted. Although no one series can be depended upon to furnish a forecast of total business activity, these series are useful in a limited way since they do reflect promptly many of the forces that have an influence on the volume of business. When the individual lead series are studied in connection with the forecast of the various components of total business, they are often valuable indicators of coming changes in direction.

A technique of forecasting changes in over-all business activity by studying components of the total is discussed in detail in the following section, and the use of important lead series is incorporated into this method of forecasting. The assumption upon which the use of lead series is based is essentially sound even though it has not proven to be so reliable as was once hoped it would be. When applied to individual components of the total economy, lead series may prove to be more valuable than when applied to a forecast of over-all business activity; but even if lead series do not furnish a dependable forecast of individual components, they may give valuable information that can be used with other data. Since the individual components of business are more homogeneous than total business activity, there is reason to believe that lead-lag relationships will be more reliable when used in connection with the forecasting of individual components.

Gross National Product Analysis

The development of reliable statistical data for the gross national product and its components has made possible a systematic approach to the problem of forecasting the business cycle. In a sense, gross national product analysis is not really a method of forecasting; certainly it is less definite than the systems based on lead series that are expected to indicate the changes in the business cycle in a purely mechanical fashion.

An analysis of the components of gross national product makes it possible for the forecaster to follow the changes in the various segments of the economy that are of particular importance. It is now generally accepted that certain factors in the business situation, such as government expenditures, expenditures for new plant and equipment, building, and net foreign investment, are strategic in that the instability in the economy seems to be centered in them. Other factors tend to experience variations that are derived from the fluctuations in these strategic sectors. Accumulation of inventories stimulates business in much the same way as capital expansion, while liquidation of inventories has the opposite effect. Expenditures for durable consumer goods in many respects are more like capital expenditures than expenditures for consumption.

Even though the expenditures for consumer nondurable goods and services are larger than those for the strategic categories mentioned above, their fluctuations tend to be derived rather than to initiate changes in the level of total activity. Attention is first directed to the components of the gross national product that contribute to the instability, and the expected changes in these factors are traced through the various stages to derive estimates of all the components, and, as a final result, an estimate of over-all activity.

Approaching the problem of forecasting changes in total business activity through gross national product analysis makes possible the effective use of a wide variety of sources of information. All kinds of data can be made to throw light on one or more of the strategic factors; without some unifying scheme of analysis, the details of the information available become so numerous that it is difficult to determine if any significant relationship exists.

A fairly popular method of forecasting has been to prepare a list of favorable factors and another list of unfavorable factors. The two lists are then balanced against each other and the result is assumed to show which group is the stronger. The actual working of this method has usually not been very satisfactory, in part at least because it was almost impossible to give any accurate weight to the various factors. The method of gross national product analysis may be essentially this same method of

balancing favorable and unfavorable factors, but it provides a better mechanism for integrating the effects of the factors into a composite result.

It is impossible in the short space of this chapter to describe all of the information that may be used in arriving at conclusions on the behavior of the strategic elements in the business situation. A brief description of some of the most significant factors to consider for each component is all that will be attempted.

As a starting point, it is well to remember that gross national product includes a classification of the total output of goods and services according to the group that buys the final product. This classification places the emphasis where it belongs, since the function of the economic system is to produce goods and services for final consumption. The forecast of over-all business activity consists of the total forecasts of the individual components of this total demand for goods and services.

Many factors must be taken into consideration in the determination of final demand. Among these factors are wage payments, profits, dividends, savings of corporations, savings of individuals, taxes and governmental services, prices, interest rates, fiscal policy of governments, controls over money and credit, the attitude of governmental agencies towards business, and the international situation. All of these factors need to be interpreted in light of their effect on the final demand for the different kinds of goods and services represented by the major subdivisions of gross national product.

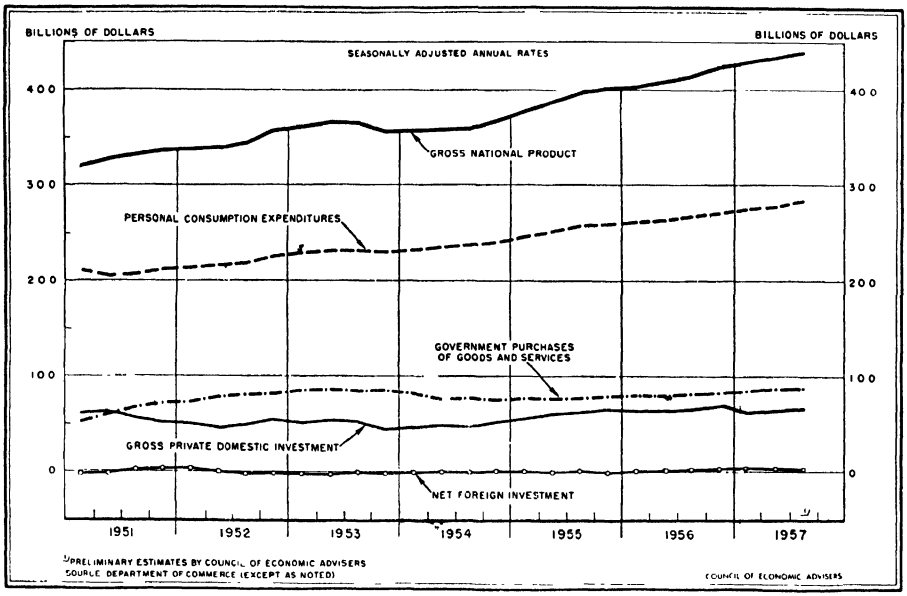
Gross national product analysis can be used for forecasts of varying degrees of detail. For example, if a decision calls for a forecast of business activity for a year in the future to be made on short notice, it would be possible to spend the time available considering as much information as possible on the probable course of each component of gross national product. Even the most hurried projection of the components could be checked for consistency and there would be no danger that an important element had been overlooked, since the components of gross national product make up the total output of the economic system, and in terms of total dollars, each component is automatically weighted.

At the other extreme, a detailed study could be made of each of the components by the most elaborate statistical analysis. The major components could be broken down into minor subdivisions and each could be forecast separately. The framework of gross national product would make sure that all of the subdivisions were weighted correctly in building up the total, whether they were studies in detail or in extremely brief fashion.

The fluctuations in the gross national product and its four major components are illustrated in Chart 21-1 by quarters for 1951 to 1957. The

Chart 21-1

GROSS NATIONAL PRODUCT CLASSIFIED BY MAJOR COMPONENTS, 1951-1957



Economic Indicators, October, 1957

following sections will discuss the forecasting of the four major components according to the outline below:

I. Gross private domestic investment.
II. Government purchases of goods and services.
III. Net foreign investment.
IV. Personal consumption expenditures.

Gross private domestic investment. The three major subdivisions of gross private domestic investment that are discussed below are:

1. Change in business inventories.
2. Residential building.
3. Plant and equipment expenditures.

It will be noted that the residential portion of new construction is described in a separate section; the remainder of construction is included in the section, "Plant and equipment expenditures." Such a division of the domestic investment component is desirable because of the publication of the series on new plant and equipment expenditures and the reported

plans on these expenditures. (This series was described briefly on page 488 in Chapter 20.) The statistical series on inventory changes, new construction, and expenditures for new plant and equipment include most of the expenditures in the gross private domestic investment component, and it is possible to make a satisfactory forecast of this component from these three subdivisions.

CHANGE IN BUSINESS INVENTORIES. The changes in business inventories represent an especially significant component of gross private domestic investment. In one sense the change in inventories is different from the other components of the gross national product in that changes in the amount of goods in inventories do not represent the final purchase of goods, but rather a disturbance in the movement of goods to the final consumer. As the volume of business grows, the amount of goods in inventory must also grow, which means that over a long period the increases in inventory will be greater than the decreases. The spending for inventories is a fixed investment in the same way that the purchase of fixed assets represents a permanent investment.

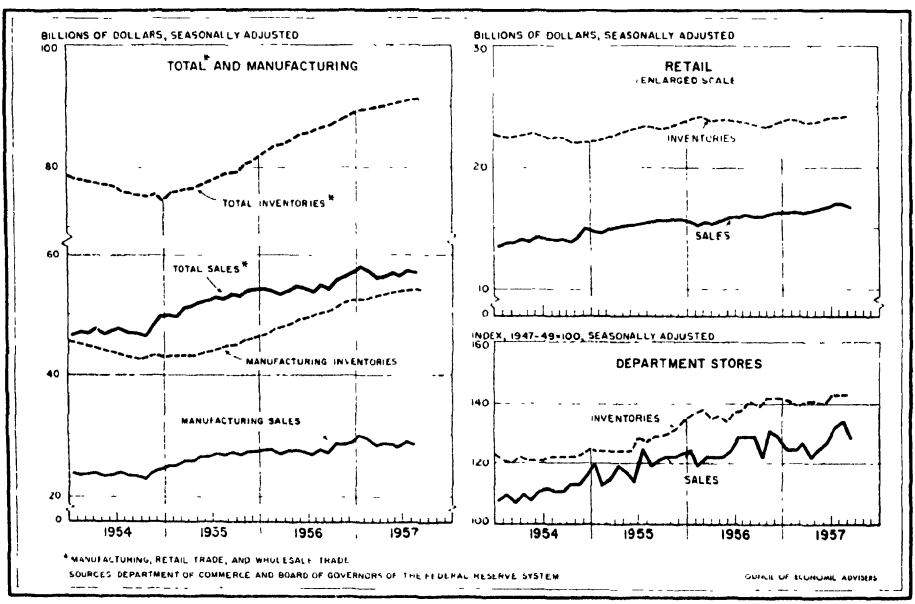
Over short periods inventories may increase or decrease in rather substantial amounts; in some years the major portion of the changes in gross national product have been the result of changes in the size of inventories. When inventories are increasing, it means that the stream of income in the form of wages and other costs of production is greater than the expenditures for the goods purchased. The money used to buy the goods carried in inventory is generally furnished by an expansion of bank credit, which means that new purchasing power is created. As long as inventories increase, this stimulus to business continues; but once businessmen decide that inventories are too high, the reverse process gets under way. Less is paid out as income to the factors of production than is required to purchase the goods that businessmen are trying to sell. This process reduces the total gross national product directly, and by curtailing production of additional goods, it may bring about a reduction in income. This can develop into a self-generating spiral unless the reduction in inventories results in an increase in production to replenish stocks that may have dropped too low.

Changes in inventories frequently generate sharp fluctuations in gross national product, but if they are corrected by a reversal when the total value of inventories gets too far out of line, the result is a minor cyclical swing in business. However, if the decline in inventories generates a decline in capital expenditures, a much more serious type of decline may get under way. A decline in investment expenditures will have far-reaching effects on the total volume of business, and once started down, it may generate a further decline that tends to gain momentum.

The problem of forecasting the change in inventories can be attacked by studying the trend of inventories in the immediate past and evaluating the level of inventories in relation to the volume of business currently being done. The ratio of inventories to monthly sales can be used to help decide whether or not the goods in stock are excessive in relation to current demand. The ratio of inventories to sales can be computed for retail, wholesale, and manufacturing establishments for past periods and a judgment can be made as to whether the level is critical. After a period of depression, a rising ratio of inventories to sales is good for business, but when the ratio begins to reach an unusually high value, serious question should be raised as to the length of time the rise can continue. An excessive ratio is likely to be corrected by a reduction in inventories. The Department of Commerce publishes detailed data on inventories and sales by kinds of business, and a careful watching of these series will usually give warning of a change in direction of this component. (The data currently available are described in Chapter 20.) Monthly data on sales and inventories from 1954 to 1957 are shown in Chart 21-2.

Chart 21-2

MANUFACTURING AND TRADE SALES AND INVENTORIES
1954-1957



Economic Indicators, October, 1957

RESIDENTIAL BUILDING. Expenditures for residential building show an extreme instability and fluctuate with a much wider amplitude than personal consumption expenditures. There is some indication that building expenditures fluctuate in long cycles of fifteen to twenty years' duration, as well as following the swings of the cyclical fluctuations in total business activity. Even with a high degree of correlation between building and the business cycle, the extremely wide amplitude of the fluctuations in building exerts a strong influence on the total gross national product.

The amount of new housing constructed in any year is a small fraction of the total supply of housing in existence, with the result that major shifts in demand can have an exaggerated effect on the amount of building in a given year. A growing population is one of the important factors in determining the demand; but shifts in the population can have an equally important effect, since housing cannot be moved with the population. When one area grows in population at the expense of another area, a demand is created for new housing just as if there had been an increase in population.

The availability of credit is an important factor in demand, since residential housing is seldom bought out of consumer income. The influence of the federal government on credit supply and interest rates has become of major importance in considering the future of building activity. Shortage of credit can effectively curb an upswing in building, and a relaxing of credit restrictions can stimulate activity when other aspects of the demand situation are favorable. It is debatable whether or not plenty of credit will in itself cause building to increase, but it has been demonstrated many times that the availability of credit is one of the necessary conditions for an expansion of building activity.

A study of the factors influencing the demand and supply of credit is a necessary start in the forecasting of residential building. Official forecasts of the number of new houses to be started in a given period are made by governmental agencies; in fact, if one is willing to take a forecast made by someone else, there is never any problem of finding a prediction on the rate of building activity. The statistical series showing the number of houses started, described on pages 486 and 487 in Chapter 20, represent some of the best sources of information, since the amount of constructon put in place in a given month is related to the number of houses started several months earlier. The data on building permits are actually a lead series since the issuance of a permit precedes the actual work on a house.

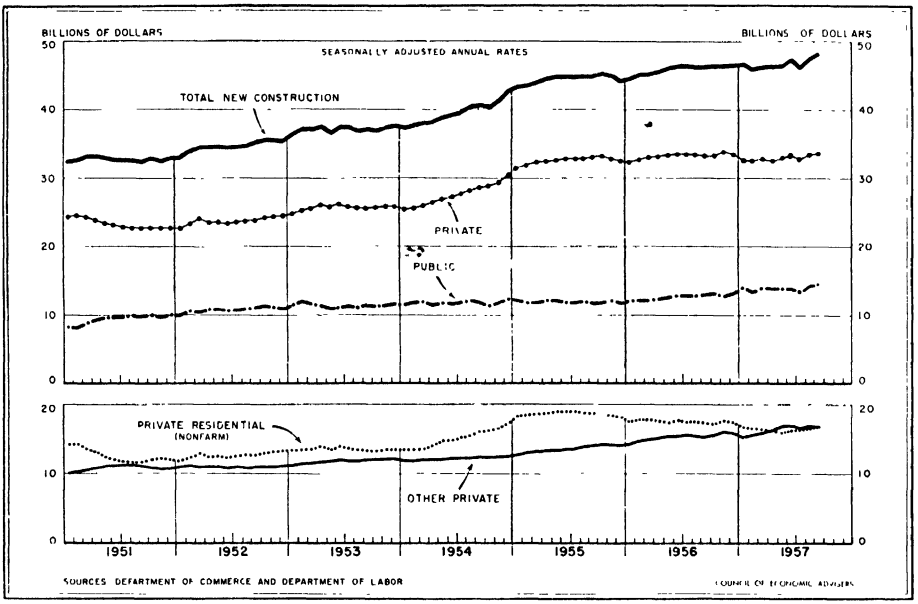
The forecasting of short-range variations in building is rather subjective, no matter how carefully one has compiled all the information on trends in building permits, rents, costs, number of vacant houses, interest rates, availability of credit, and population changes. The residentiai

building category should be watched constantly for changes in direction, since the probability of predicting its course for very many months is not great. The fact that a large number of statistical series relating to this industry are available makes it possible for the forecaster to learn of changes promptly and to make adjustments in the forecast very quickly.

Chart 21-3 shows the monthly series on total new construction put in place, classified as public and private, and with private nonfarm residential shown as a separate class.

Chart 21-3

NEW CONSTRUCTION EXPENDITURES, 1951-1957



Economic Indicators, October, 1957

PLANT AND EQUIPMENT EXPENDITURES. The estimates of new plant and equipment expenditures by the Securities and Exchange Commission and the Department of Commerce differ from the gross national product estimates in that the two agencies mentioned omit the following categories: capital expenditures charged to current accounts, farm construction and capital equipment, professional equipment and instruments, and private nonprofit construction. In a very detailed forecast it might be important to make estimates for each of these categories, but it is usually satisfactory to use the rate of increase in the estimated expenditures for new plant and equipment to project the remainder of gross private domestic investment, after separate forecasts have been made for inventory change and residential nonfarm building.

The level of investment in plant and equipment appears to be almost completely independent of the level of income, and there appears to be no relationship with any series that can be used for forecasting purposes. A considerable portion of expansion projects is financed by funds from outside the business and is therefore less dependent on income than consumer expenditures. The need for capital goods is not necessarily related to the current demand for the product, since expansion of plant and equipment would be based more on expected future demand. If the future appears to be uncertain, expenditures might be limited almost entirely to replacement needs.

Spending for plant and equipment is of such a nature that plans must be formulated considerably in advance, and after a certain stage, there exist commitments that are not likely to be canceled except in unusual circumstances. When expansion is planned to take care of secular trend, or when new equipment is installed to reduce costs, expenditures may be made with very little regard to the current demand situation. Such a situation has made possible the collecting of reasonably reliable information on the investment programs of business. In spite of the danger that plans might change, the reported plans form the basis for the best forecast of business investment that is available at this time. In addition to the series collected by the Department of Commerce and the Securities and Exchange Commission, which were described in Chapter 20, the McGraw-Hill Publishing Company, Inc., makes similar surveys that are widely used along with the government estimates.

The reported plans for new plant and equipment expenditures are generally sufficient for making a forecast of this portion of the gross national product; and unless a very detailed study were being made, it would seem unnecessary to do more than to study these data. Even though the classifications for new plant and equipment expenditures are not exactly the same as the gross national product categories, the data should be sufficiently accurate to project the gross national product components by using the percentage increase shown in the reported plans.

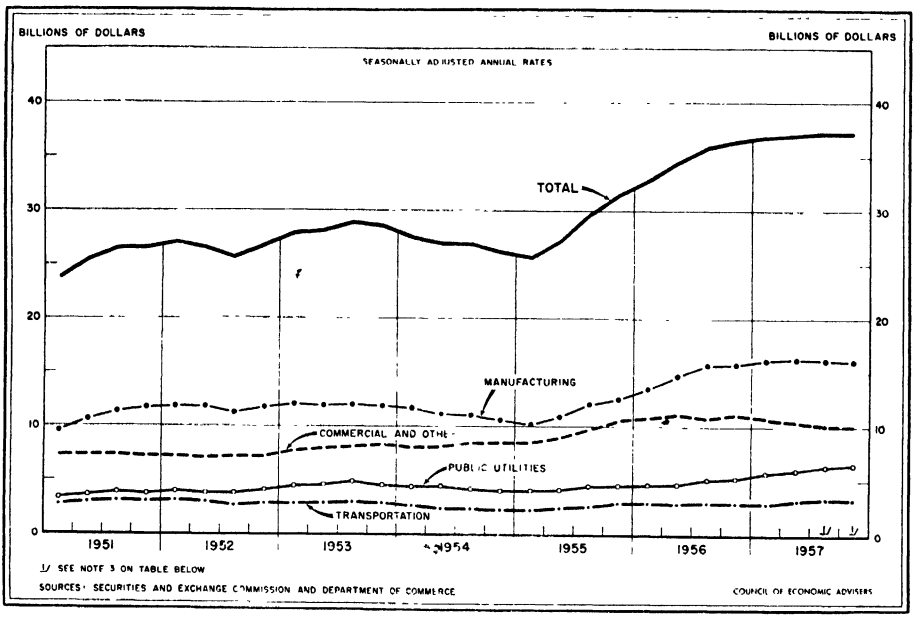
Chart 21-4 shows the expenditures for new plant and equipment, classified by major industries. Chart 21-5 shows gross private domestic investment and the three major subdivisions. This chart summarizes the fluctuations in the strategic factors that have just been discussed.

Government purchases of goods and services. The purchases of goods and services by government should be classified into at least three groups:

1. State and local government.
2. Federal government — nondefense.
3. Federal government — national defense.

Chart 21-4

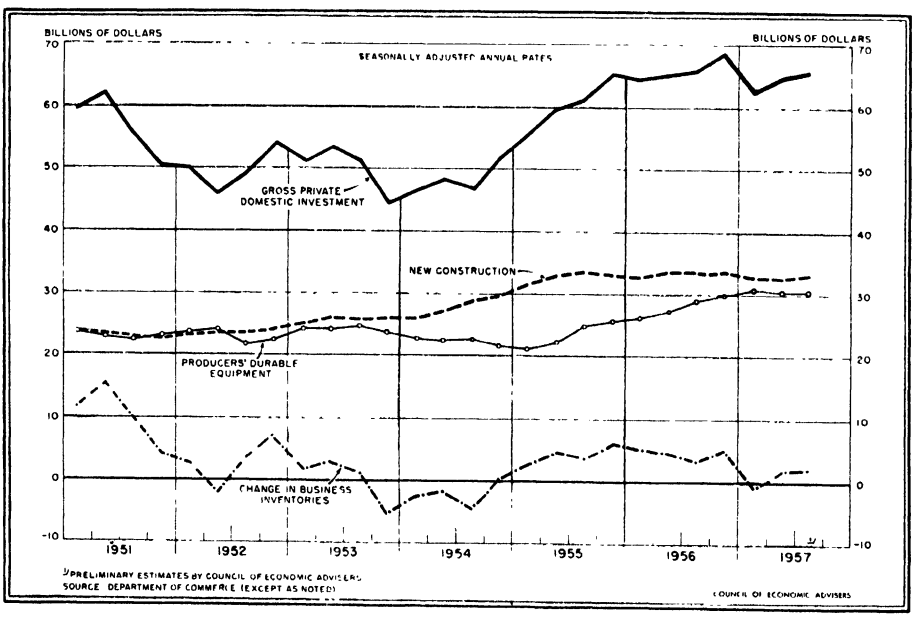
EXPENDITURES FOR NEW PLANT AND EQUIPMENT 1951-1957



Economic Indicators, October, 1957

Chart 21-5

GROSS PRIVATE DOMESTIC INVESTMENT, 1951-1957



Economic Indicators, October, 195

Expenditures of state and local governments are not subject to rapid change. This is also generally true of nondefense expenditures of the federal government, since most of the programs to which the government is committed are fixed by law and any change is likely to come about slowly. The defense expenditures, however, are subject to sharp fluctuations, with the result that they can have a significant effect on the over-all level of business. In addition to representing a substantial part of total expenditures, the fact that defense expenditures are autonomous rather than being derived from other movements in the gross national product makes them strategic factors in the forecast of the level of business activity.

The steady upward trend of expenditures by state and local governments in recent years permits a forecast of this series to be made by a rather simple extrapolation of the past trend. If the average percentage increase of the previous three years were projected one year, it would probably be a satisfactory forecast. Even if a severe cyclical decline occurred during a year, it is almost certain that state and local expenditures would not reflect this decline, since funds are appropriated for at least one year in advance and many of them are made two years in advance.

The federal budget is given a great deal of publicity, and official estimates of both defense and nondefense categories can be used to forecast these two classifications. The official estimates are, of course, subject to error and the forecaster can try to modify the official estimates; but such a forecast will inevitably be highly subjective. The amount of spending by the federal government, particularly the defense expenditures, depends upon decisions that are determined by domestic and international conditions, which cannot be predicted by any kind of analysis. Anyone who aspires to making a better forecast than the official estimates needs to follow the political as well as economic developments both at home and abroad and try to take them all into account in making a decision.

Net foreign investment. Net exports or imports of goods and services represent the smallest of the components of gross national product, although net foreign investment is a strategic factor in forecasting the total. In recent years net foreign investment has been such a small percentage of total gross national product that it could have little effect on the total forecast. If a detailed forecast is undertaken, a complete analysis of the prospects for both exports and imports should be made. In general, imports would be expected to show a positive correlation with the level of income in this country, while exports are more closely related to the economic conditions of foreign countries. For a less detailed forecast, it is usually satisfactory to project the trend of the last two or three years, unless there is specific reason to believe that a change has occurred.

Personal consumption expenditures. In Chapter 20 it was shown that personal consumption expenditures were 64.4 per cent of the total gross national product in 1956. This segment of total business activity is the largest component; but since its fluctuations are the result of changes in other parts of the economy, the discussion already presented will serve as the basis for a forecast of consumer expenditures.

The fluctuations in personal consumption expenditures are, in general, the result of variations in the stream of income paid out to consumers in the course of the operations of the economic system. However, at times there are other forces that modify the effect of the income payments; and in making a forecast of consumer expenditures for a given year, it is always wise to determine whether or not these forces are exerting any unusual influence on the basic relationship between income and expenditures.

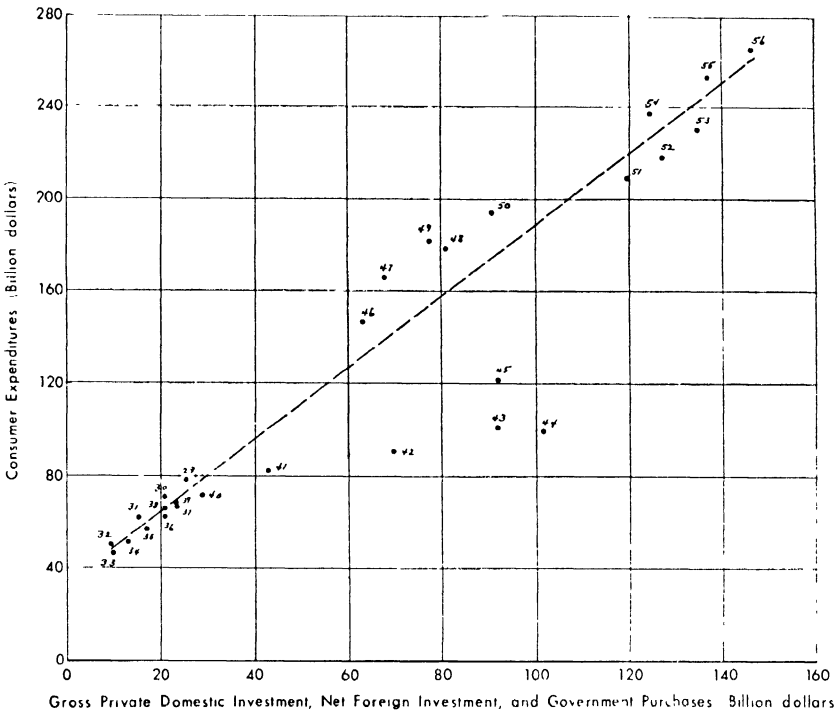
Chart 21-6 is a scatter diagram showing the relationship between consumer expenditures and the total of the other three components of gross national product. This scatter diagram shows the accuracy that would have resulted from using this relationship to forecast consumer expenditures for the years 1929 to 1956. For the years prior to 1941 the correlation was very high, but if the regression line based on these years is projected, the years since 1940 show rather pronounced deviations.

For 1941 and the following four war years, consumer expenditures were substantially below the amount that would be indicated by the expenditures in the three other components. This can be explained largely by the shortage of goods for consumers, particularly durable goods. Personal savings increased greatly and consumer debt declined sharply during these years.

During the five years immediately after the end of the war, consumer spending exceeded the amount projected by the regression equation. Spending during this period was influenced by accumulated savings of the war years and the incurring of a substantial volume of consumer debt. The years 1951 to 1956 fall very close to the regression line, which would be a reasonably good estimate of consumer expenditures without any adjustment for changes in the savings rate or the volume of consumer credit. A more accurate forecast of expenditures would be secured if these factors were given consideration, and in periods such as a war and the immediate postwar period such adjustment would be very important. Since information on consumer credit and personal savings is readily available, it should always be considered in making the forecast of consumer expenditures, even though it usually does not show much change from the volume computed from the relationship to the autonomous components of the gross national product. Even though consumers can generally be depended upon to spend the income they receive in a given period, there are enough

Chart 21-6

RELATIONSHIP OF CONSUMER EXPENDITURES TO TOTAL OF OTHER COMPONENTS OF GROSS NATIONAL PRODUCT, 1929-1956



Source of data Survey of Current Business, July, 1957

possibilities of variation in this pattern to make it worth while to look for random variations and make allowance for them when possible. A number of forecasts of consumer spending have been seriously wrong because no allowances were made for consumers who substantially changed their previous saving and borrowing practices.

FORECASTS FOR AN INDUSTRY AND INDIVIDUAL BUSINESS

The businessman is interested in a forecast of total business activity more because of its importance in forecasting for his individual business firm than because of an interest in the general forecast itself. Budgets are made for the individual business concern, and the forecast of sales is the ultimate goal of the businessman's measurement of changes in business activity and the forecasting of gross national product. The relationship of a company's sales to gross national product or one of the major components is a basic consideration in formulating a sales forecast. Often it is

better to forecast the level of the industry first and from this forecast to arrive at a projection of the individual company's sales. This procedure involves the additional step of knowing the position of one's business in relation to the industry of which it is a part. After a firm has determined the percentage that its sales represent of the total market, it can project this percentage to correspond with the level of the industry forecast. If there is reason to believe that the company's percentage of total sales will change, a different percentage may be used. The decision as to the relative position of a company in relation to the industry must be based on an appraisal of the strength of competitors, the company's products, sales promotion plans, prices, and any other specific information that can be secured.

The use of a scatter diagram and either a graphic or mathematical analysis of the correlation between the different factors is basic to this method. Because a considerable element of judgment enters into the forecast after the regression analysis has been made, it is probably just as satisfactory to use graphic analysis as to perform the calculations needed for an arithmetical analysis. The forecast from the regression analysis should be considered as a first approximation that, on the basis of additional information, can be modified as seems wise.

As a final word of caution, it should be emphasized that no matter what methods of forecasting are used, it is essential that the forecasts be checked against the judgment of individuals who are familiar with the business. While it is true that the use of statistical data is an attempt to substitute facts for subjective judgments, it does not mean that knowledge gained through experience in a given situation should be ignored in favor of quantitative data It is particularly important to take into consideration any specific plans of the business that might affect the pattern of sales in relation to the indicators used for forecasting.

Finally, it is important to emphasize that any forecast should be reviewed frequently and revised in light of the most recent information. The techniques of flexible budgets have been developed to permit the revision of budget estimates, and everyone dealing with forecasts should be alert to the need for constantly checking to see if anything has happened to change the outlook. Keeping accurately informed about the current level of business is probably the simplest insurance that can be secured against making wrong decisions regarding the future. It is possible that business statisticians have not emphasized enough that a precise, up-to-date measurement of the present level of business may be more useful than an inaccurate forecast of the direction in which business is headed.

STUDY QUESTIONS

1. Does forecasting the cyclical fluctuations in business require more knowledge of statistical methods or of economics? Explain.

2. How successful a method of forecasting secular trend is the projection of past trends into the future? What assumptions are implicit in a method that relies solely on the projection of past trends? Do you think these assumptions are valid?

3. Is it correct to conclude that the forecasting of seasonal variations is easier than the forecasting of either secular trend or cyclical fluctuations? Give reasons for your answer.

4. Sales of life insurance show a regular seasonal rise in December and a sharp drop in January. What do you consider a logical explanation of this seasonal pattern?

5. Do you consider it important to understand the reasons for a seasonal pattern before using it as a forecast of the seasonal variation in a coming year? Why?

6. What characteristics of an industry make a forecast of the secular trend particularly important? Give an example of an industry that has a particularly urgent need for an accurate forecast of secular trend.

7. The text states that in forecasting secular trend it is important that the series be broken into as many subdivisions as possible, with special attention being paid to securing as great a degree of homogeneity in each group as possible. Why is it logical to place greater confidence in the projection of the more homogeneous of two series, assuming that other factors are essentially the same?

8. What factors account for the greater interest in the forecasting of secular trend at the present time than twenty years ago?

9. What contributions do statistical methods make in the forecasting of secular trend? Do you consider these forecasts more a statistical problem or an economic problem?

10. It has been suggested that the cyclical fluctuations of business activity are purely random variations, such as the variations that would occur in the number of heads secured if ten coins were tossed a large number of times. If this view of the business cycle were accepted, what would be a logical approach to the problem of forecasting cyclical fluctuations? Do you believe that this explanation of the cyclical fluctuations of the business cycle is correct?

11. Why has the theory that the business cycle could be forecast by securing a series that would always turn ahead of general business had such an appeal for business analysts? Do you believe that anyone at the present time knows of any series that will accurately forecast the fluctuations in total business activity?

12. Do you consider it to be more or less likely that a series could be found that would lead the changes in the level of an individual industry? Give an example of a situation in which you think such a relationship might exist.

13. What are the limitations of the forecasting technique that attempts to set up a balance sheet of favorable and unfavorable factors to determine the direction that the cyclical swings of business are likely to take?

14. In preparing the sales forecast of a business concern, it would be easy to secure from each salesman an estimate of the sales he expects to make during the period covered by the forecast. How accurate a sales forecast would you expect this method to give?

15. A great deal of attention is being paid to surveys that ask for information from business concerns on their capital expenditure plans. Do you consider the information secured from the surveys useful in forecasting the cyclical fluctuations in business? Why?

16. How valuable is the information that is secured by asking consumers what their spending plans are for next year? Would the information on certain types of spending be more valuable than others? Give reasons for your answer.

PROBLEM

This problem consists of preparing a report, the purpose of which is to forecast the secular trend of demand for petroleum in the United States to 1975. The projection of the gross national product for 1975 was made in this chapter on pages 498 to 502, and these projections may be used in the forecast of the demand for petroleum. The forecast should be supported by as detailed an analysis of the factors that may affect demand as it is possible to make in the time you are allotted for making the forecast. Give all the statistical data on which the forecast is based, state all your assumptions, and show all computations.

22 MARKETING RESEARCH

The application of research methods to the problems of marketing has become so general in business that marketing research is recognized as an important specialized and professional service. The methods used in marketing research are so predominantly statistical that they are properly considered a part of the field of business statistics as well as marketing, and in this text many of the illustrations of the use of statistical methods have been drawn from marketing research applications. The statistical methods used in the analysis of marketing problems have sometimes been designed to meet a particular situation, although generally it has been a matter of adapting standard statistical techniques to the study of a particular marketing problem.

Most of the statistical methods used in the study of marketing problems have already been described; but, at the risk of some duplication of material presented earlier, this chapter will briefly review some ways in which the methods of collecting and analyzing statistical data can be applied to the problems of marketing goods and services. Attention will be given to the agencies involved in marketing research, and a simple outline of the types of marketing problems to which statistical methods have been successfully applied will be given.

AGENCIES AND FUNCTIONS OF MARKETING RESEARCH

Marketing research makes use of internal data compiled from the records of the individual concern; general information collected and published by public agencies, such as government departments; data collected and sold by private statistical agencies; and studies made especially for a specific problem and for the exclusive use of one business.

Statistical methods used in marketing research include the collection of statistical data as well as the analysis of the data after they have been collected. The use of the data in making decisions, of course, is the all-important reason for collecting them; and the amount of analysis that should be applied is governed by what is needed to make the data effective for use in making decisions. The quantity of data needed usually is so

large that the problems of collecting it become a major consideration in the organization of the market research function. External data are particularly burdensome to collect, and there has been a definite trend for specialized agencies to assume this function. Some large business concerns attempt to collect part of the external data they need, but it is difficult for an individual business to collect large quantities of external data.

The analysis of the data on marketing problems may be performed by the individual company that will use it, or it may be done by specialized market research organizations. Some of these organizations collect data and also offer skilled, professional services in performing analysis and even make recommendations on the decisions. Other organizations believe it is better for the executives making the decisions to interpret the information they will use.

Since collecting, analyzing, and interpreting statistical data tend to be so closely related, it is practically impossible to classify agencies according to the functions they perform. Instead of attempting such a classification, the following discussion will describe several types of organizations that are actively engaged in conducting some phase of marketing research, and will point out some of the most important ways that each one contributes to the whole field of marketing research.

Government

The sources of statistical data were described in Chapter 4 without any attempt being made to point out specifically the sources that are devoted to supplying marketing data. The basic sources of data that are useful in marketing research are the major statistical compilations of the federal government. Also, many state agencies collect data of value in marketing research. In this country, it is considered a proper function of government to collect and publish basic statistical data on the population and the economy of the country.

Trade Associations

The activities of trade associations are directed to a considerable degree towards the collection and publication of statistical data on markets; and since the members of a trade association represent a reasonably homogeneous group, this is a particularly useful source of data on marketing problems. A trade association is in a position to concentrate on more specific problems than a government agency, with the result that the data compiled are more directly applicable to the problems of an individual business.

Private Statistical Organizations

Another agency engaged in the collection of basic statistical data on markets is the private statistical organization that collects data on markets on a regular basis and sells the results to a number of subscribers. When the data needed are expensive to collect and of interest to only a limited number of business concerns, it is unlikely that either a government department or a trade association will be willing to spend the money necessary to collect the information. This means that the cost to a small group of users will be higher than if more concerns were interested in the information, but the data can still be bought from a statistical agency cheaper than if it were compiled by each business.

The terms under which information is supplied by private organizations vary widely. At one extreme is the organization that publishes a statistical compilation that is sold to anyone who will pay the subscription price. At the other extreme is the firm that sells tabulations of data on an exclusive basis to only one concern in each kind of business, which makes the information more valuable to the individual business than if all its competitors had access to the same data. Obviously such a service could be made available only by a private organization; agencies supported by public funds have an obligation to make the information available to any concern that wants to use it. As a general rule, the more limited the circulation of the information, the more possibility there is that the supplying organization will provide analysis and interpretation.

A typical example of the type of information supplied by the private data-gathering agency is the information furnished by store audits. Reports furnished subscribers give detailed data on sales to consumers, inventories, and prices for a wide variety of items. Manufacturers have been generally unable to determine from the analysis of their sales records how fast the goods were being bought by consumers in different markets, and how well their products were selling in relation to competitors' products. The A. C. Nielsen Company has been an outstanding source of this type of information in the food and drug industries.

Consultants

The major functions of marketing research consultants are interpreting information and using it in the solution of marketing problems. A consulting organization might consist almost entirely of professional staff members who furnish only interpretation and advice; it might also include a complete statistical organization that is equipped to collect, tabulate, and analyze the statistical data needed in arriving at conclusions. There are also a number of large, well-staffed market research firms that will take

over a complete study and deliver a final report with advice on the decision to be made. In some cases it might be difficult to decide whether a particular organization belongs in the consultant category, or whether it should be considered primarily a data-collecting agency; the dividing line between the groups is not too distinct.

Individual business concerns find it economical to use the services of consulting organizations even though the fees are relatively high. Except for the largest concerns, the cost of maintaining a staff capable of supplying all of the specialized services that a firm of consultants can provide would usually be prohibitive. When large amounts of information must be collected by personal interviewers, a data-collecting organization can be used effectively; but when specialized professional skills are needed, the consulting organization is generally the most efficient agency to use.

There are disadvantages in engaging the services of an agency or persons not familiar with the problems of the business, but to a certain extent this disadvantage is offset by the advantage of having the problem attacked by someone with an outside viewpoint and wide experience in different situations. The use of independent consultants in the field of marketing research is similar to their use in other areas of business, such as law, finance, etc. However, the practice of employing consultants is relatively new in marketing. Experience has indicated that businesses can most effectively use independent market research consultants when, in addition to the consultants, the individual business concern has a staff in its organization that possesses a sound knowledge of how to use the results of research in making decisions. This leads to the problem of organizing a business concern in order to make the most efficient use of marketing research techniques.

Marketing Research Departments of Individual Concerns

If marketing research is used to any important extent in reaching decisions on marketing problems, it is essential that the business have some formal plan for the organization of marketing research. For the business concern that is able to afford the expenditure on a continuing basis, it is possible to set up within the company a complete statistical department to collect, tabulate, and interpret all the data needed by executives in making decisions on marketing problems. Or, varying amounts of the work of compiling the information may be turned over to outside organizations.

Although outside organizations may collect and analyze the data, it is the responsibility of management to understand the techniques of market research and to make effective use of research information in the operations of the business. Without an executive who knows what is needed before

a study is made and then understands how the results can be used, there is the danger that an elaborate report may be purchased and then never used in arriving at decisions. Generally, a concern using marketing research determines by experience which functions can most profitably be performed by outside organizations, taking into consideration the volume and regularity of the work and the type of outside services that are readily available.

Advertising Media and Agencies

Magazines, newspapers, and radio and television networks have been important agencies in the development of marketing research. The Commercial Research Division of the Curtis Publishing Company was one of the earliest and most influential marketing research organizations. Selling space to advertisers requires that a salesman know what market is represented by the readers of his publication. Thus, many magazines and newspapers have entered into marketing research. When radio and television became important advertising media, they immediately became equally interested in determining the characteristics of the persons reached by their programs. The major networks have set up extensive marketing research organizations, whose functions have been supplemented by the individual stations.

Advertising agencies also became a major source of marketing research data as the services rendered by the agencies expanded beyond merely buying space and writing copy. Research in the problems of advertising has become an integral part of the planning of sales campaigns, and many advertising agencies maintain large and competent research staffs that study a wide range of marketing problems for their clients.

There has been some criticism leveled at the research performed by advertising media on the grounds that it tends to be biased. The fact that the results of research are used by salesmen of advertising media to sell space or time implies an inherent danger that the results presented will be only those that favor the use of the medium being represented. However, many research departments have developed the reputation for objectivity and for accurately reporting the facts. Research by advertising agencies also has sometimes been labeled as superficial. Since the research performed by the agency is often but a small part of the total service rendered the client, for which a commission is charged, there may be a resulting temptation to cut costs by reducing the amount of research carried on. But as standards of good marketing research have become better established, the quality of the service rendered by advertising agencies has improved and the research departments of agencies are among the important sources of marketing research.

TYPES OF MARKETING RESEARCH

The field of marketing research covers a wide variety of problems and uses a number of different research techniques. Consumer surveys are probably the best-known examples of marketing research, but much significant research is done that does not fall under this classification. Business concerns represent an important segment of the total sales market, and the study of the problems in reaching these markets is another category of marketing research. Still another important type of research is composed of the various quantitative methods of analysis related to the problems of sales control. This includes the analysis of internal sales and cost records and external statistical data relating to markets. Advertising research, which may overlap some of the other types of marketing research, is such a well-recognized category that it may well be considered a separate type of marketing research.

Consumer Surveys

Probably the most widespread type of marketing research in the United States is the study of what consumers like and dislike about the products and services offered for sale by business. Consumer surveys are designed to find answers to the questions that marketing executives raise as they attempt to make intelligent decisions. Such questions as who uses the product, why consumers prefer one brand over another, how consumers would react to a change in design or a new package, and where consumers prefer to buy a given product are typical. The investment required to place a new product on the national market is so great that few companies would attempt to do so without finding out in advance the answer to the basic question: How will the product be accepted by the consumer?

Examples of the collection of statistical data to supply information to marketing executives have been used in various places throughout this text. The diagram on page 87 was used to show the importance of providing for a channel of information that would promptly relay the acceptance, or nonacceptance, of a product to the management of the business. By the time the sales figures show that customers do not like a product, a great deal of business may have been lost. In the highly competitive field of marketing consumer goods, information about the consumer market is essential, and more reliable and valid information is being sought constantly.

Consumer surveys are continually discovering that the factors a manufacturer may have rated as important are completely ignored by the customer. Similarly, a survey may show that some characteristics never

before considered by the manufacturer would add greatly to the appeal of the product. It continues to surprise manufacturers to learn that neither their salesmen nor retailers know exactly what the consumer prefers.

Collecting information directly from consumers by the survey method has become a standard procedure in the marketing practices of many major producers of consumer goods, and under the competitive system by which goods are marketed it appears likely that the survey method will continue to grow in importance. Since the information gives one concern an advantage over its competitors, it becomes practically necessary for other firms to take steps to get similar information.

Closely related to consumer surveys are the various kinds of opinion polls. The development of opinion polls on social, economic, and political questions has paralleled the development of consumer surveys for business use. The statistical methods are essentially the same, and business concerns are making an increasing use of the results of opinion polls, as distinct from the typical market survey. With business management becoming more and more concerned with the opinion of various public groups, the taking of opinion polls for business use is increasingly important. There is still a great deal of interest in the political polls, particularly in presidential election years, in spite of the error in the 1948 election forecasts.

Industrial Marketing Research

The basic statistical techniques of collecting data are used in both consumer surveys and industrial marketing research, and the only real differences between the two types of research grow out of the fact that industrial marketing research relates to business concerns rather than individual consumers. There is much less variation in the importance of the individual units in the consumer market than between units in the industrial market. In some industries a relatively small number of concerns represents the major portion of the demand, while in other industries the market is widely dispersed among many small companies. This variation in the industrial market must be taken into consideration when collecting data.

The consumer market generally represents a much more homogeneous group than the industrial market. In selling machinery and equipment, for example, it is necessary to analyze the needs of each industry in detail to determine whether or not there is any potential market for a given product. If a business concern wants to know the market for a certain type of a compressor, it is frequently extremely difficult to locate all of the industrial classifications that might need this particular product. Such a situation creates a real demand for marketing research to supply information on the location of the market and how to reach it.

Quantitative Methods of Research in Sales Control

In the description of consumer surveys and industrial marketing research, the discussion was concerned chiefly with processes involved in collecting data. However, it was emphasized in the early part of the text that the first step in any statistical analysis is to be certain that a complete inventory of all existing information be made before additional information is collected. In the study of either consumer or industrial markets, a great deal of valuable information can be found in the records of the business concern and in published statistical form. The records of a business will supply quantitative data on sales and costs, while published data are a valuable source of information on potential markets. The analysis of these two kinds of statistical data will provide a degree of control over sales and sales effort that will improve substantially the marketing efficiency of a business.

Sales analysis. Of all the information that a marketing executive should have available, probably a detailed record of past sales, costs, and profits is the most fundamental. Without such a record of the performance of the business, it is difficult to make any plans for the future. The more detail that is furnished about the significant parts of the sales record, the better the executive can understand and solve the problems facing him. The information needs to be available promptly to be of the most use in making decisions; and, as business operations have become larger and more complex, mechanical tabulating devices have speeded up the processing of the data needed for decisions. Punched-card tabulation made it possible to analyze a large volume of statistical data in a short time, and now, electronic data processing offers a method of making still more data available in a shorter time.

Since an invoice is usually written for every sale, the basic information for an analysis of sales is available in every business. Although it is not always so easy to make as an analysis of sales, a detailed analysis of costs, in addition to the sales analysis, will give still better control, since there is almost always a variation in the profit on different products. An analysis of sales and costs will make possible a much more accurate determination of the products that are making a profit and the conditions under which the most profitable operations are being conducted. The analysis can be made for different products, salesmen, individual customers, classes of customers, regions, or any other basis of classification that appears to be significant. Any management that neglects to secure all of the information possible from an analysis of its own records is open to criticism because modern methods of collecting data are available to convert sales and cost figures into a usable form to provide the basis for making decisions.

Market potential. The measurement of market potential makes extensive use of detailed statistical data, classified on a geographic basis, on such subjects as the number of people and their characteristics, the number of business concerns, the amount of economic activity, and the flow of income to consumers. The statistical data measuring the amount of potential business, combined with the information supplied by the analysis of sales records, give a complete basis for evaluating sales performance.

Measures of market potential are basic to practically any kind of sales planning, since, in controlling sales, usually the first thing that must be known is the amount of the product that can reasonably be expected to be sold. Not only must the amount of total sales be available, but the amount must also be broken down for rather small areas, since the administration of sales activity is generally organized on a geographical basis. The setting of sales quotas summarizes all of the available information on potential sales. The quotas may be determined for individual salesmen's territories, or they may form the basis for setting up a distributor's territory. The quotas may also be used to locate the strongest markets or to eliminate the weak markets. The same information on potentials may be used to allocate advertising expenditures.

Measuring the market potential for any product or business should always start with the published statistical data available. Although the information needed is probably never complete, the detailed statistical data now made available by public organizations give a surprisingly complete picture of the market for a large part of the total industrial output. The number of statistical series needed to measure the potential market varies widely. Convenience goods such as soap, packaged food staples, and drugs are primarily related to the number of people in an area, although specific consumption habits may modify somewhat the basic relationship between sales and population. Luxury goods are correlated with the number of people with higher incomes. Industrial products are related to the volume of the industries that use a given product. Agricultural machinery is designed for specific crops, so the potential market for a given piece of equipment depends on the volume of production of that crop in different regions. For practically any product it is possible to find one or more statistical series that measures in some manner the sales potential for the product. These statistical series are known as *market factors* or *market indicators*, and the first step in the measurement of market potential is to determine the best market indicator available.

The simplest measure of the potential market for the product of a given concern is the total sales of the product by areas. For example, figures showing the amount of gasoline consumed within each state are available from the sales tax records that are maintained, and the potential market

of a given company could be considered as following the same distribution as the total sales of gasoline. In many states, sales tax collections on gasoline sales are available by counties, and these data can be used in the same manner to measure the potential sales. In many ways data on the total sales of a commodity are the best possible measure of market potential, but there are not many commodities for which such information is available on a detailed geographic classification. A number of series are available on a state basis, but to be of maximum usefulness a classification of sales by counties is needed.

If the total sales of a product are not available for the areas needed, a good substitute method is to use some product that shows a high correlation with the product for which a measurement of the potential market is desired. For example, if county data on gasoline consumption were available, but data for lubricating oil were not, it might be reasonable to assume that the consumption of oil would follow the same pattern as gasoline; and the potential market for oil could be computed on the basis of this relationship. This is essentially the method of correlation analysis, whether the relationship between the two commodities is derived from the formulas in Chapters 18 and 19 or by a simpler calculation. The accuracy of the computations of the market potential depends on the reliability of the assumed relationship between the two factors, as was discussed in detail in Chapters 17 to 19.

If the record of sales by geographical areas is available, it is possible to measure the correlation between sales and one or more market factors and use this relationship to compute the potential sales for additional areas. The use of multiple correlation makes it possible to derive a regression equation for computing sales potentials from a number of market factors. The example on page 465, showing the correlation between heating oil consumption, number of burners, and temperature, could be extended to a geographical distribution and the two independent variables could be used as the market factors for computing the potential sales of heating oil by counties, cities, or other areas.

One of the most widely used compilations of data measuring market potentials is the *Survey of Buying Power*, published annually by *Sales Management*. The consumer market potential is measured by counties and cities, giving estimates of population, retail sales in total and for selected kinds of business, and effective buying income. As an example, the market potential represented by farmers is measured by the value of farm products sold. In a recent issue the value of farm products sold is given for each state and for the 400 leading agricultural counties. The data given for the farm market vary somewhat in different years, since estimates are made for years in which census data are not available.

The amount of information on industrial markets included in the *Survey of Buying Power* gradually increased from year to year, until in 1957 a separate publication, *Survey of Industrial Buying Power*, was issued. Industrial potentials are measured in terms of manufacturing employment in all counties with 1,000 or more employees in industrial establishments. The 100 leading counties are ranked according to 20 manufacturing classifications and 6 nonmanufacturing categories. The industrial market is more highly concentrated than the consumer market and is generally subject to more dynamic change.

The publishers of *Sales Management* sell sets of punched cards containing not only the data published in the *Survey of Buying Power*, but also a considerable amount of data upon which the estimates are based. Since a large volume of statistical computation is involved in measuring sales potential for many areas, it is a distinct advantage to be able to secure the data in a form that permits the use of mechanical equipment. All of the materials on market potentials published by *Sales Management* are widely used in marketing research and are generally accepted as the most detailed allocation of the potential market that is available. Obviously in making such extensive estimates, many of the results are only approximate; but these estimates have been accepted by marketing executives as valuable help in making decisions requiring a knowledge of market potentials.

Advertising Research

Although advertising is an integral part of marketing, there are enough specific kinds of analyses used almost exclusively in relation to the problems of advertising to justify giving special attention to the subject. The decisions to use magazines, newspapers, radio, television, direct mail, outdoor advertising, or some other medium require a great deal of careful study if the most effective expenditure of the advertising budget is to be accomplished. Also many decisions, such as those regarding copy, the use of color, and the layout, must be made in the course of every advertising campaign. The use of radio brought a new set of problems related to developing programs that would hold the attention of listeners; television added to these problems.

Printed media. When consideration is given to buying advertising space in magazines and newspapers, the fundamental data needed is an analysis of circulation. The Audit Bureau of Circulation is the basic source of data, and its audit of circulation statements insures an honest reporting of the number of subscribers With this basic statistical data available, it is possible to make detailed analyses of the extent to which

any combination of publications will cover the market that the advertiser wants to reach. The correlation between circulation and any measure of market potential may be used to check how accurately the selling effort is spread over the market.

If a more precise measure of the impact of an advertising campaign is desired, it may be necessary to determine the number of readers of a publication, instead of merely taking a count of the number of copies sold. Surveying the number of readers has been receiving an increasing amount of attention, and the results of these studies are used to supplement the basic data on circulation. Information showing the various classes of readers of a publication is also useful in planning an advertising campaign. Since the duplication of circulation is a key problem in building a list consisting of a number of magazines, there is a great demand for data that will show the number of new readers that will be gained by adding different combinations of publications.

The above subjects are only a few of the topics on which research is continually being undertaken in an effort to give the advertiser more information about the audience that his campaign is reaching. The expenditures for advertising are so large that some method of checking on the effectiveness of these expenditures is considered a necessity. The tremendous volume of research on printed media that is constantly being undertaken reflects the intensity of the need for such data.

Radio and television. Measuring the size of the audience reached by a radio or television station offers special problems not encountered in determining the number of subscribers to a publication. The problem resembles that of determining how many persons read a publication, as distinct from the problem of determining the size of circulation. The only practicable approach seems to be to find out from the owners of sets how many were tuned to each station in an area at a given time. Because of the size of the universe, it is necessary to use a sampling technique rather than a complete enumeration.

There are four common methods of securing information on whether or not a television or radio set is tuned to a given program.

1. One method is to make personal interviews, asking what programs were watched or listened to the previous day. Usually a roster of programs is provided to aid the respondent in recalling the program, but the method places heavy reliance upon memory.

2. Another method is to arrange to have a diary kept by set owners showing when the set was turned on and to what station, assuming that the diary will be kept currently and will eliminate the danger inherent in asking for remembered facts.

3. A third method places no reliance upon memory but asks, by means of telephone calls, if the radio or television set is turned on, and if so, to what station. Enough calls must be made during each period of time to give a large enough sample for that period.

4. A fourth method uses a recording device that can be attached to a radio or television set in such a way that it records on tape or photographic film when a set is turned on and the station to which it is tuned. This method has the advantage of recording observed data and thus avoids the possibility of inaccuracies that may arise when persons are asked to supply information. The recording device gives a complete record for each unit to which it is attached; and if proper sampling techniques are used, satisfactory estimates of the universe can be made.

Copy testing. The cost of an advertising campaign is so great that much care is normally taken to test the campaign in every possible way before it is conducted on a large scale. After an advertisement has been run, it is often tested again for effectiveness even though whatever is learned about it cannot be used in the current campaign. Since advertising and selling are a continuous process, it is always possible to profit by a study of past experience. The continuous study of the effectiveness of advertisements by research departments has provided copy writers with an increasing understanding of why advertising sells goods.

Before a proposed series of advertisements is run, the series may be tested on a group of typical consumers that make up a consumer jury. However, a more objective method of testing the effectiveness of an advertisement is to run a number of different campaigns in test markets, using audits of store sales to determine the relative increase in sales in the different areas. If store audits are correlated with specific advertising campaigns, it is possible to obtain more objective evidence of the effectiveness of the advertising. Another approach to the testing of copy is to include a coupon with a special offer or a premium. An analysis of the coupons returned may be used to measure the effectiveness of different advertisements. The interpretation of the coupon returns, however, is subject to question since returning a coupon is not the same as buying the product. Nevertheless, decisions both as to copy and the selection of media are based on an analysis of coupons returned.

COMMENTS ON SOME STATISTICAL TECHNIQUES USED IN MARKETING RESEARCH

The discussion in this chapter and in earlier chapters of the text has emphasized that marketing research represents an important application of statistical methods to business problems. As stated before, most of the

statistical methods used in marketing research are no different from those generally used by business statisticians. For example, the methods of collecting and tabulating data are generally the same in marketing research as in any other kind of statistical work. Punched-card tabulating machines were adopted early for use in marketing research, and electronic data processing is now being used. The analysis of time series, forecasting techniques, and correlation methods are standard analytical tools in using marketing data. Sometimes the fact that a procedure is known by a different name gives the impression that the procedure itself is different. Where a special method appears to be used, it is usually only a modification of a standard method. Some statistical techniques that require special consideration in marketing research are discussed below.

Sampling

The problems of sampling have loomed large in marketing research from the beginning; and as more efficient sampling methods have been devised, they have been adapted to the collection of marketing data. Although statisticians long understood the importance of random samples, there was a tendency to underestimate the difficulties of securing a random sample of consumers or business units. Mailed questionnaires were widely employed in the past because they represented a simple and relatively inexpensive method of collecting data; the difficulty of securing a true probability sample was generally ignored. The failure of the *Literary Digest* election poll to predict the presidential election of 1936 had a profound effect on the sampling techniques employed in collecting data from business and consumer populations.

The Gallup poll predicted the election of 1936 correctly by using what is now called a *quota sample*. This method of selecting a sample recognized that the interviewer was not well equipped to make a random selection of the individuals in the universe. In order to insure that the sample included all the different kinds of individuals in the universe, a system of quotas was set up, usually proportional to the major categories in the universe that were believed to be significant in the analysis of the problem. For example, if home owners were expected to give significantly different answers from those given by tenants, it would be necessary to insure that the correct proportion of both home owners and tenants be interviewed. This was done by determining the percentage of each group in the universe and requiring that the number of interviews secured be in that proportion. Within the two categories the interviewer was left free to choose those individuals that would be included in the sample.

The quota sample worked so much better than the older haphazard method of selecting a sample that it became the standard method of

marketing research and the opinion polls. However, statisticians became interested in the development of practical methods of designing probability samples; and after the methods had been tested in various surveys made by governmental agencies, they gradually became accepted as the proper methods to be used in marketing research. Improvements in the efficiency of probability samples continue to be made; and even though there is still too much data collected by poor sampling methods, good sampling procedures are now used in the majority of cases.

It is very expensive to take a probability sample each time a survey is made, so frequent recourse is made to the permanent sample, usually called a *panel*. The greatest advantages of using a panel are the reduced cost and the ease with which data can be compiled. A panel can be set up and used regularly to test the behavior of consumers in many different situations. If the same questions are asked over a period of time, it is possible to measure significant trends much easier than if a new probability sample were taken each time. The panel is also widely used in collecting data from business concerns to measure changes in business conditions. The weekly reporting member banks, discussed in Chapter 20, comprise a panel and are used to collect a comprehensive series of data on banking conditions.

Even if the panel members were selected by probability methods, the data collected from them might not give a completely accurate measure of the universe. A fixed panel fails to include new units that become a part of the universe, and this fault may lead to serious error. If a panel of retail stores is used to measure changes in sales from month to month, the new stores in the growing suburban shopping centers may not be given enough weight in determining the trend of sales. Efforts to remedy this situation by adding new stores to the panel may cause difficulty due to lack of accurate information as to the number of new stores that should be added. Also, consumer panels are criticized on the grounds that panel members become conscious of the fact that they are making reports and they cease to behave as they would have if they had not been chosen on the panel.

Motivation Research

Understanding why consumers behave as they do is basic to many decisions relating to marketing problems. The whole strategy of a sales campaign reflects the opinion as to the motives back of the consumer's decision to buy or not to buy a product. The use of the term *motivation research* is relatively new in marketing research and is subject to a considerable amount of disagreement. In a broad sense it may be used to

describe a major portion of the marketing research that is being carried on; and when used in this sense, it is only a new name for procedures that have been used for a long time. In a more restricted sense, it refers to a rather specialized type of investigation, only recently undertaken, that tries to probe intensively into human motives. When used in this latter sense, marketing research has borrowed techniques developed by psychologists and sociologists, and it is very likely that significant advances will be made in the use of these techniques to aid in the solution of marketing problems.

STUDY QUESTIONS

1. What is the importance of each of the following agencies to the business executive looking for information on which to base marketing decisions?

(a) Government agencies.
(b) Trade associations.
(c) Private statistical organizations.
(d) Consultants in special fields.
(e) Marketing research departments of individual concerns.
(f) Advertising media.
(g) Advertising agencies.

2. (a) Under what circumstances would you recommend that a manufacturer of consumer convenience goods set up a marketing research department in the company?

(b) What circumstances would make it undesirable for a company to operate its own marketing research department?

3. To some people the term "marketing research" means the same as "consumer survey." This usage is generally considered to be incorrect, but how do you explain its common acceptance? Why is it wrong to use the two terms synonymously?

4. What significant differences are there between consumer marketing research and industrial marketing research?

5. What valuable kinds of information can be secured from the internal records of a business by means of a marketing research study?

6. What are the chief problems encountered in making a detailed analysis of sales and of distribution costs?

7. How important is the determination of the relationship of sales to external market indicators? What use can be made of this relationship?

8. Why do salesmen of advertising space in magazines or newspapers and salesmen for radio and television time need statistical data?

9. How can a statistician help an advertising copy writer produce more effective copy?

10. Why has the field of marketing research been a leading user of sampling techniques?

PROBLEMS

1. The rapid expansion of the population of cities in the United States has resulted in a wide dispersal of retail shopping centers throughout the residential areas of cities and suburbs. The decision of an investor on whether or not to put money into the development of a new shopping center, or the decision of the management of a retail chain on whether or not to open a new store in a given shopping center, requires a great deal of specific information on population, purchasing power, retail sales, and competition.

Assume that the owner of a tract of land in a newly developed residential section of a city of 200,000 proposes to develop a shopping center. In order to secure the necessary financing, he needs accurate information on the potential volume of business that can be secured by a shopping center at this location. Before space can be sold or rented to merchants, they must have information that indicates that potential sales will be large enough to be profitable.

Prepare an outline of a study that will provide the information needed for the basic decisions regarding the development of the shopping center.

2. Assume that a chemical manufacturer has developed a synthetic fiber with the following characteristics. It is water-repellent; is resistant to bacteria, mildew, and a wide range of chemicals; and will not burn. However, the fiber shrinks at temperatures above 140 degrees, is too weak to be spun into yarn fine enough for weaving, and does not take dye well. The latter characteristics seem to eliminate the fiber from use in the conventional textile fields, but it has been successfully used in a few situations where its peculiar characteristics were an advantage.

Outline a study that will determine whether or not there is a potential market for the synthetic fiber.

3. Since markets are people with money, information on the distribution of population and income by geographic areas is basic to all marketing research.

(a) Compile from *Personal Income by States, A Supplement to the Survey of Current Business*, statistical data showing the total income received by residents of your state for the most recent year available and for 1939.
(b) Compile also a classification of this income by sources (such as agriculture, manufacturing, and trade) for the most recent year and for 1939.
(c) Present this data in a table and graphically.

4. (a) For each of the states that comprise the region in which you live, compile the statistical data outlined below. (Use the geographic classification of states that is used by *Personal Income by States, A Supplement to the Survey of Current Business.*)

1. Total personal income.
2. Population.
3. *Sales Management* estimates of total retail sales and sales by kind of business.

(b) From the data compiled, compute the per capita personal income and retail sales for each state. (All data should represent the most recent year available.)

23 STATISTICAL QUALITY CONTROL

A major application of statistical methods to the solution of business problems has been made in recent years by management in controlling the quality of manufactured products. Inspection has always been a vital part of the manufacturing process, and for many years the function of inspection has been to sort out the acceptable product from the unacceptable. Beginning soon after World War I, significant progress was made in applying statistical methods to the problems of inspection with the result that better control of quality of the product has become possible, frequently at a savings in cost.

The term *quality control* has sometimes been used to describe the field of activity represented by the application of statistical methods to the problems of manufacturing, but there is valid objection to using such a general term to apply to one specific method of controlling quality. The term that appears to describe best the use of statistical methods in the control of the quality of manufactured products is *statistical quality control*, which is the expression that will be used in this chapter.

VARIABILITY IN THE MANUFACTURED PRODUCT

One of the basic considerations that is always present in producing goods is the fact that it is impossible to make all of the items exactly alike. If accurate enough measurements could be taken, it would be discovered that no two items are ever exactly alike, although approximate measurements may not show that a difference exists. Specifications recognize this fact by giving a tolerance range within which measurements must fall. Individual items falling outside this range are not acceptable; if they are final products they will not perform satisfactorily, or if they are component parts, they will not fit properly when assembled with other parts. The basic principle of mass production is the fact that the individual parts are all near enough the same size to be interchangeable.

A simple and direct method of insuring that all of the parts fall within the specification limits is to inspect each item and reject any that fail to fall within the limits specified. The rejected items may be reworked to make them conform to the specifications or, if this is impossible, they must be

scrapped. Inspection of each item produced, a method which is probably as old as manufacturing, will give control over the quality of manufactured products by eliminating most of the defective items. However, the use of statistical methods has proven to be so much more efficient in controlling quality that statistical quality control is one of the very significant uses of statistical methods in business management. Its proper use requires such a detailed knowledge of manufacturing as well as statistical methods that it is now recognized as a specialized professional activity. This chapter will merely introduce the subject by describing a few basic applications.

The variations in a given characteristic of a manufactured product may be grouped into two classes on the basis of the causes of the variations. One group of variations comprises those resulting from specific causes that can be identified. Variations in the product due to mistakes of inexperienced workmen, worn tools, machines in need of adjustment, and defective raw materials are examples of this class. These causes of variations are called *assignable causes* and, since they cause a relatively large variation in the product, they should be identified and removed. For example, a machine which needs adjustment should be located and the adjustment corrected, or an employee who is performing an operation incorrectly should be given further instruction.

A second type of variation may result from a random combination of circumstances that cause slight differences in the individual units produced, differences that individually have so little effect on the result that it is impracticable to try to locate them or to trace their effects. The variations that result from the numerous minor causes are called *random variations*, and they may be considered as simply a characteristic of the manufacturing process. Even though the same machines, materials, labor, and manufacturing techniques are used, it is to be expected that some variations will occur in the product. However, it is not worth the cost of trying to find the reason for each of these variations since they are in reality the result of chance.

Once it has been established that a given degree of random variation is inherent in a process, a decision must be made as to whether or not this variation in the product is greater than can be tolerated. If the variations are greater than can be permitted, a change in the process should be made to the extent necessary to make it possible to turn out an acceptable product. This may mean such changes as purchasing better machines, hiring more skilled labor, providing labor with better training, or buying better raw materials.

Once it has been determined that the random variations in the product can be tolerated, the problem becomes that of preventing any assignable cause from introducing variations. If variations due to an assignable cause

should appear, it is important that the cause of the variations be located and removed. Since the fundamental problem in this connection is distinguishing between the random variations and variations that can be attributed to specific assignable causes, the inspector who has the responsibility for passing on the acceptability of the product needs some guide that will enable him to distinguish between causes.

CONTROL CHARTS FOR VARIABLES

Control charts are used to furnish information on the nature of the variation in the characteristics of the manufactured product. When the control chart gives warning that an assignable cause is present, steps can be taken to find the trouble and correct it before more defective products are produced. As long as the control chart indicates that only random variations are present in the product, no further inspection is necessary. The control charts for variables that are most commonly used are $\bar{X}$ charts and R charts.

$\bar{X}$ Chart

The $\bar{X}$ chart is constructed to show the fluctuations of the means of samples about the mean of the process and can be used to determine whether or not the fluctuations are due to random causes or to an assignable cause. The data given in Table 23-1 will be used to illustrate the construction and interpretation of the $\bar{X}$ control chart. A critical dimension was measured to the nearest ten thousandth of an inch. The first item measured .5025. Then, .5000 inches was subtracted from .5025, and the decimal was moved four places to the right (.5025 − .5000 = .0025, or 25). Thus, the first measurement is 25.

It is not necessary to convert the data as described above, but since it results in considerably less clerical work, it is rather standard practice in many types of statistical work. The amount to subtract from the individual observations (.5000 in this case) was chosen because it seemed to effect the greatest saving in calculations. This is true because, by subtraction, each individual value was converted from a four-digit number to a two-digit number, with an obvious reduction in clerical work. Expressing the dimensions in .0001 inches had the additional advantage of eliminating the need to work with decimals. It is, of course, necessary to show exactly what operation was performed on the data; the subtitle in Table 23-1 is used to explain the conversion in this example.

Table 23-1

MEASUREMENT OF A CRITICAL DIMENSION OF 20 SAMPLES OF 4 ITEMS

UNIT: .0001 INCHES IN EXCESS OF .5000 INCHES

Subgroup number	Measurement of individual items				ΣX	X	ΣX^2	σ^2	σ	R
1	25	14	19	18	76	19.00	1,506	15.5000	3.9	7
2	22	16	20	19	77	19.25	1,501	4.6875	2.2	6
3	24	12	15	24	75	18.75	1,521	28.6875	5.4	12
4	18	17	23	21	79	19.75	1,583	5.6875	2.4	6
5	26	19	16	21	82	20.50	1,734	13.2500	3.6	10
6	18	17	16	15	66	16.50	1,094	1.2500	1.1	3
7	19	22	15	14	70	17.50	1,266	10.2500	3.2	8
8	18	20	21	23	82	20.50	1,694	3.2500	1.8	5
9	17	21	20	17	75	18.75	1,419	3.1875	1.8	4
10	20	16	22	17	75	18.75	1,429	5.6875	2.4	6
11	18	19	21	20	78	19.50	1,526	1.2500	1.1	3
12	19	13	20	18	70	17.50	1,254	7.2500	2.7	7
13	19	22	21	18	80	20.00	1,610	2.5000	1.6	4
14	21	16	17	19	73	18.25	1,347	3.6875	1.9	5
15	15	23	15	16	69	17.25	1,235	11.1875	3.3	8
16	20	22	20	19	81	20.25	1,645	1.1875	1.1	3
17	17	23	18	19	77	19.25	1,503	5.1875	2.3	6
18	22	19	21	18	80	20.00	1,610	2.5000	1.6	4
19	17	25	24	20	86	21.50	1,890	10.2500	3.2	7
20	18	16	14	21	69	17.25	1,217	6.6875	2.6	7
Total	..	..	..	..	1,520	380.00	29,584	143.1250	49.2	121

SOURCE: Confidential.

The measurements in Table 23-1 are given for 20 subgroups, each containing 4 items. At intervals of 30 minutes, 4 items were taken from the assembly line for inspection and the measurements of the critical dimension were entered in the table. The subgroup numbers represent the order in which the items were produced. If we assume that there were no assignable causes of variation in the dimension present while these items were being produced, the mean of each of the 20 samples of 4 is an unbiased estimate of the universe mean, although a sample of 4 would be expected to produce an estimate with a wide confidence interval. The mean of the 20 sample means would, however, be a more reliable estimate of the universe mean than any of the individual sample means. The mean of the sample means is represented by $\bar{\bar{X}}$ and is used as an estimate of the mean of

the universe. A method of checking on the assumption that no assignable causes of variation in the dimension were present during the period in which the samples were taken will be explained later. Any sample that is known to have been taken when an assignable cause was present should be eliminated from the calculation of $\bar{\bar{X}}$. The values of the 20 sample means are given in Table 23-1, and from these means the value of $\bar{\bar{X}}$ was found to be 19.0 $\left(\frac{\Sigma \bar{X}}{20} = \frac{380.00}{20} = 19.0\right)$.

Table 23-1 gives also the values of σ^2 and σ for each of the samples. It was explained in Chapter 10 that the standard deviation and the variance of a small sample are biased estimates of the universe values. The bias may be removed by multiplying the sample variance (σ^2) by the ratio $\frac{n}{n-1}$, after which the average variance can be computed. Since the same correction is applied to each of the variances being averaged, it is less work to average the uncorrected variances and then correct the average. The average variance is found to be 7.15625 by dividing the sum of the 20 variances (143.1250) by 20.

$$\bar{\sigma}^2 = \frac{143.1250}{20} = 7.15625$$

Using the value of $\bar{\sigma}^2$ as the variance (σ^2) in equation (24), the estimate of the universe standard deviation (σ') is made as follows:

$$\sigma' = \sqrt{\bar{\sigma}^2 \frac{n}{n-1}} = \sqrt{(7.15625)\left(\frac{4}{4-1}\right)} = \sqrt{9.5417} = 3.1$$

If the mean of the universe, in this case all of the production of this particular item, is assumed to be .5019 inches (using only the last two digits, it becomes 19), the problem is to determine within what range the means of samples of four will fluctuate. The standard error of the mean was computed in Chapter 10 by the following equation (ignoring the correction for a finite universe):

$$\sigma_{\bar{x}} = \frac{\sigma'}{\sqrt{n}}$$

For a sample of four items and a universe standard deviation of 3.1, the standard error of the mean is computed as follows:

$$\sigma_{\bar{x}} = \frac{3.1}{\sqrt{4}} = 1.55$$

It is almost universal practice in statistical quality control work in the United States to use a confidence interval of three standard errors. The probability of a sample falling outside the three-sigma confidence interval is so small that it is considered practically certain that the mean of a random sample will fall within the limits of $\pm 3\,\sigma_{\bar{x}}$ from the mean of the universe. In other words, it is assumed that a sample of four items will fall below $\bar{\bar{X}} + 3\,\sigma_{\bar{x}}$, the upper control limit (UCL), but above $\bar{\bar{X}} - 3\,\sigma_{\bar{x}}$, the lower control limit (LCL).

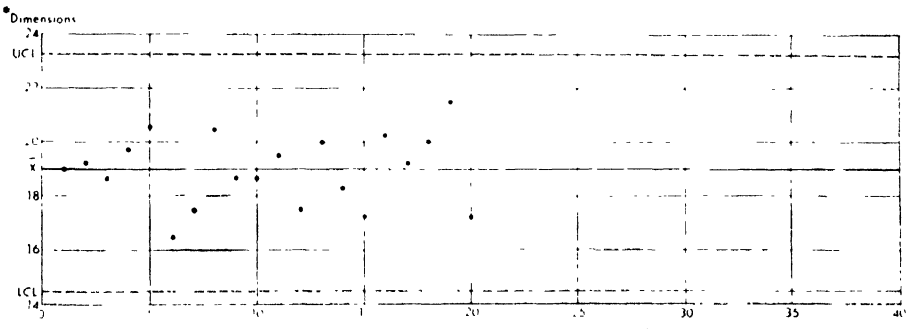
$$\text{UCL} = \bar{\bar{X}} + 3\sigma_{\bar{x}} = 19.0 + (3)(1.55) = 23.65$$

$$\text{LCL} = \bar{\bar{X}} - 3\sigma_{\bar{x}} = 19.0 - (3)(1.55) = 14.35$$

Since the samples were taken in chronological order, it is possible to set up a time series chart showing the universe mean, upper and lower limits of the confidence interval, and the mean of each of the samples. The chart will then show graphically whether or not any of the sample means fall outside the confidence interval. Chart 23-1 is a control chart for $\bar{X}$ and is used to determine each time a sample mean is plotted whether the variation of this sample mean from the universe mean is to be ascribed to random causes or assignable causes. All of the 20 sample means in Chart 23-1 fall within the control limits, which indicates that the process is in control, and that only random variations were present in the production at the time the samples were taken. If a sample mean had fallen outside the control limits, it would indicate that such a large deviation from the universe mean was the result of something more than the forces causing random variations. The very small probability that a sample mean will deviate more than three sigmas from the universe mean justifies drawing this conclusion. It is considered that such an extreme variation from the universe mean can be explained more logically by some assignable cause than by

Chart 23-1

CONTROL CHART FOR MEANS



Sample number

*Unit: 0001 inches in excess of 5000 inches

Source of data: Table 23-1

assuming that it was merely due to random variations. It would not be correct to say that it is *certain* that such a deviation is not due to random variations, but the probability that the deviation would be due to a random variation is very small. As a practical matter, it is much more reasonable to conclude that some nonrandom cause has influenced production, and a search should be made for this cause of variation. When an assignable cause is believed to be present, the process is said to be *out of control.* As long as only random variations are present, the process is *in control.*

If the mean of any of the subgroups plotted in Chart 23-1 had been either above or below the control limits, the $\bar{\bar{X}}$ and $\bar{\sigma}$ should be recomputed without including the data for the subgroup that was out of control. The control chart should be revised on the basis of the new value for $\bar{\bar{X}}$ and the new upper and lower control limits. There is a possibility that the new control limits will show another subgroup out of control; if so, the process is repeated until all of the sample means fall within the control limits.

After the control limits have been established, the chart may be used to check on the manufacturing process during subsequent production. The chart may be extended as shown in Chart 23-1, with a sample of four items taken at 30-minute intervals, and the average for the sample plotted on the chart. If the plotted mean falls outside the control limits, it indicates that an assignable cause is present; and when this occurs, the usual procedure is to start an immediate search for the cause of the variation. When the cause has been found and corrected, the average of the next sample of four items should fall within the control limits. It might be wise to continue more frequent inspections until it is certain that the cause has been removed and production is again in control. If it appeared that production had been out of control for some time, all items produced during this period might be inspected to remove those that did not meet specifications.

The control limits are set at three standard deviations from the mean, so that the probability of the mean of a sample exceeding the limits due to random variations will be very small (99.73 per cent of the area of the normal curve lies between the three-sigma limits). The small probability of exceeding the control limits when the process is actually in control has been adopted to avoid issuing numerous "false alarms." If a control chart gave frequent indications that the process was out of control but investigation established that only random variations were present, the chart would quickly cease to be used with any confidence. Using the wide control limits based on three standard deviations may result in warning not always being given as soon as the process goes out of control, but against this disadvantage must be weighed the danger of giving a warning of trouble when no trouble exists. The control limits based on three standard deviations have been accepted as a satisfactory compromise.

Computing control limits from $\bar{\sigma}$ instead of $\bar{\sigma}^2$. It is possible to compute the control limits of the X chart by using the standard deviation (σ) to estimate σ' instead of using the variance (σ^2) as explained above. In statistical quality control the preference has been to use the standard deviation instead of the variance. The standard deviation of the universe can be estimated by dividing $\bar{\sigma}$ by the following term, which has been designated as c_2*.

$$c_2 = \sqrt{\frac{2}{n}} \frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-3}{2}\right)!} \quad \ldots\ldots\ldots\ldots (71\dagger)$$

The value of c_2 when $n = 4$ is computed as follows:

$$c_2 = \sqrt{\frac{2}{4}} \frac{\left(\frac{4-2}{2}\right)!}{\left(\frac{4-3}{2}\right)!} = \sqrt{.5}\,\frac{1!}{.5!}$$

$$c_2 = \sqrt{.5}\,\frac{1}{.5\sqrt{\pi}} = .707107 \frac{1}{.8862}$$

$$c_2 = .7979$$

From the value of c_2 the mean of the universe can be estimated from the mean of the standard deviations in Table 23-1. The mean of the 20 standard deviations ($\bar{\sigma}$) is 2.46 ($\bar{\sigma} = \frac{49.2}{20} = 2.46$). The estimate of the universe standard deviation (σ') is found to be 3.1 by dividing $\bar{\sigma}$ by c_2 ($\sigma' = \frac{2.46}{.7979} = 3.1$). This estimate of the universe standard deviation calculated from the mean value of the distribution of sample standard deviations is the same as the estimate secured on page 542.

The value of c_2 for each value of n from 2 to 15 is given in Table 23-2; tables are available in books on statistical quality control for larger values of n. The calculation of the control limits from $\bar{\sigma}$ may be further simplified by the use of the values of A_1 given in Table 23-2. As was stated on page 543, the equation for the upper control limit is $UCL_{\bar{x}} = \bar{\bar{X}} + 3\sigma_{\bar{x}}$.

*W. A. Shewhart, *Economic Control of Quality of Manufactured Product* (New York: D. Van Nostrand Co., Inc., 1931), p. 184.

†In formula (71) the symbol (!) represents the factorial of a number. The factorial of a number is the product of the integer multiplied by all the lower integers: the factorial of 4 (written 4!) is $4 \times 3 \times 2 \times 1 = 24$. The factorial of 3.5 (written 3.5!) is $3.5 \times 2.5 \times 1.5 \times .5 \times \sqrt{\pi} = 11.6314$.

Since $\sigma_{\bar{x}} = \dfrac{\sigma'}{\sqrt{n}}$

$$UCL_{\bar{x}} = \bar{\bar{X}} + 3\frac{\sigma'}{\sqrt{n}} = \bar{\bar{X}} + 3\frac{\dfrac{\bar{\sigma}}{c_2}}{\sqrt{n}} = \bar{\bar{X}} + \frac{3}{c_2\sqrt{n}}\bar{\sigma}$$

Letting $$A_1 = \frac{3}{c_2\sqrt{n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (72)$$

$$UCL_{\bar{x}} = \bar{\bar{X}} + A_1\bar{\sigma} \text{ and } LCL_{\bar{x}} = \bar{\bar{X}} - A_1\bar{\sigma}$$

The use of the value of A_1 from Table 23-2 gives the same control limits as computed on page 543, except for rounding error.

$$UCL_{\bar{x}} = 19.0 + (1.880)(2.46) = 23.6$$
$$LCL_{\bar{x}} = 19.0 - (1.880)(2.46) = 14.4$$

Table 23-2

VALUES FOR COMPUTING CONTROL LIMITS FOR $\bar{X}$ CHARTS FROM $\bar{\sigma}$

n	c_2*	A_1†
2.....	.5642	3.760
3.....	.7236	2.394
4.....	.7979	1.880
5.....	.8407	1.596
6.....	.8686	1.410
7.....	.8882	1.277
8.....	.9027	1.175
9.....	.9139	1.094
10.....	.9227	1.028
11.....	.9300	.973
12.....	.9359	.925
13.....	.9410	.884
14.....	.9453	.848
15.....	.9490	.816

*c_2 computed from formula (71).
†A_1 computed from formula (72).

Computing control chart limits from $\bar{R}$. Since the number of control charts needed in any quality control program is large, any substantial savings in the time required to compute the necessary values is an important consideration. The mean of the ranges ($\bar{R}$), instead of the mean of

the standard deviations of the subgroups, can be used in computing the control limits; and the range of a sample can be computed in considerably less time than the standard deviation. The relationship between the standard deviation of the universe and the range (R) of a sample is shown by the values d_2 in Table 23-3 for samples with values of n from 2 to 15. The value of d_2 for a given value of n shows the ratio of the average value of the range of all samples of size n to the standard deviation of the universe from which the sample was taken. When $n = 4$, $d_2 = 2.059$, which indicates that the average range of samples of 4 is slightly more than two times the standard deviation of the universe from which the samples were taken ($d_2 = \frac{\bar{R}}{\hat{\sigma}} = 2.059$). If the value of $\bar{R}$ computed from the data in Table 23-1 is taken as the average value of the distribution of ranges from samples of 4, the standard deviation of the universe from which the samples were taken can be estimated by the following equation

$$\sigma' = \frac{\bar{R}}{d_2}$$

The value of $\bar{R}$ is 6.05 and the estimate of the standard deviation of the universe is as foliows

$$\sigma' = \frac{6.05}{2.059} = 2.94$$

Table 23-3

VALUES FOR COMPUTING CONTROL LIMITS FOR $\bar{X}$ CHARTS FROM $\bar{R}$

n	d_2	A_2
2.....	1.128	1.880
3.....	1.693	1.023
4.....	2.059	.729
5.....	2.326	.577
6.....	2.534	.483
7.....	2.704	.419
8.....	2.847	.373
9.....	2.970	.337
10.....	3.078	.308
11.....	3.173	.285
12.....	3.258	.266
13.....	3.336	.249
14.....	3.407	.235
15.....	3.472	.223

SOURCE: American Society for Testing Materials, *A.S.T.M. Manual on Quality Control of Materials.* (used with permission)

This estimate of the standard deviation of the universe may be used instead of the estimate based on $\bar{\sigma}$ to compute the control limits for the $\bar{X}$ chart. Because of the savings in time required to calculate R instead of σ for each sample, constructing the control limits from $\bar{R}$ is preferred in quality control work. The mean of the standard deviations gives somewhat more accurate values for the control chart; but when the subgroups are small, the ranges give satisfactory values. Just as the calculation of the control limits from $\bar{\sigma}$ is simplified by the use of the values of A_1, the calculation of the control limits from R is simplified by the use of the values of A_2 given in Table 23-3.

Since

$$\sigma' = \frac{\bar{R}}{d_2}$$

$$\text{UCL}_{\bar{x}} = \bar{\bar{X}} + 3\frac{\frac{\bar{R}}{d_2}}{\sqrt{n}} = \bar{\bar{X}} + \frac{3}{d_2\sqrt{n}}\bar{R}$$

Letting

$$A_2 = \frac{3}{d_2\sqrt{n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (73)$$

$$\text{UCL}_{\bar{x}} = \bar{\bar{X}} + A_2\bar{R} \text{ and } \text{LCL}_{\bar{x}} = \bar{\bar{X}} - A_2\bar{R}$$

Using the values of A_2 and $\bar{R}$, the control limits are found to be only slightly different from those computed from A_1 and $\bar{\sigma}$ on page 546.

$$\text{UCL}_{\bar{x}} = 19.0 + (.729)(6.05) = 23.4$$

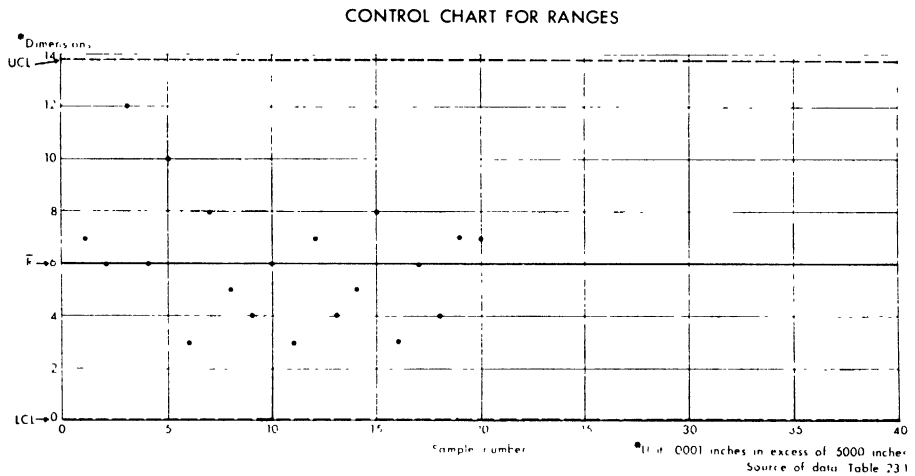
$$\text{LCL}_{\bar{x}} = 19.0 - (.729)(6.05) = 14.6$$

R Chart

The R chart can be used to show the fluctuations of the ranges of the subgroups about the average range $\bar{R}$. The construction of an R chart follows the same general principle as the $\bar{X}$ chart described above, using the mean of the ranges of the samples as the center line and the $3\sigma_R$ limits. The distribution of the ranges of all possible small samples from a normal universe is not normal, and a larger proportion of the cases may exceed the upper $3\sigma_R$ limit than would be true for a normal distribution. The lower control limit for small samples in many cases would be negative; but since the range cannot have a negative value, the lower control limit in such situations is set at zero. Although the distribution of the ranges is not normal, the three-sigma control limits indicate when the vari-

ability is so great that it may be assumed to be the result of an assignable cause rather than random causes. With no lower control limit, it is possible for the process to go out of control only by exceeding the upper control limit. Such an increase in variability might occur without the mean going out of control, and in such a situation the R chart would give warning of an assignable cause of variation when the $\bar{X}$ chart did not. It is much more common for an assignable cause that increases the variability also to throw the process out of control with respect to the mean, in which case the R chart would not be needed. Sometimes the R chart is not used; but since it can be constructed with very little more work than is needed for the $\bar{X}$ chart, it is generally kept along with the $\bar{X}$ chart.

Chart 23-2



The control limits for the R chart are computed as follows:

$$\text{UCL}_R = \bar{R} + 3\sigma_R$$

$$\text{LCL}_R = \bar{R} - 3\sigma_R$$

If

$$D_4 = 1 + \frac{3\sigma_R}{\bar{R}}$$

and

$$D_3 = 1 - \frac{3\sigma_R}{\bar{R}}$$

$$\text{UCL}_R = D_4\bar{R}$$

and

$$\text{LCL}_R = D_3\bar{R}$$

For the data in Table 23-1, the following control limits are computed, using the values of D_4 and D_3 from Table 23-4.

$$\text{UCL}_R = D_4\bar{R} = (2.282)\,(6.05) = 13.8$$

$$\text{LCL}_R = D_3\bar{R} = (0)\,(6.05) = 0$$

Table 23-4

VALUES FOR COMPUTING CONTROL LIMITS FOR R CHARTS FROM $\bar{R}$

n	D_3	D_4
2.....	0	3.267
3.....	0	2.575
4.....	0	2.282
5.....	0	2.115
6.....	0	2.004
7.....	.076	1.924
8.....	.136	1.864
9.....	.184	1.816
10.....	.223	1.777
11.....	.256	1.744
12.....	.284	1.716
13.....	.308	1.692
14.....	.329	1.671
15.....	.348	1.652

SOURCE: American Society for Testing Materials, *A.S.T.M. Manual for Quality Control of Materials.* (used with permission)

With the value of $\bar{R}$ plotted as the center line and the control limits just computed, Chart 23-2 is constructed for the 20 subgroups. By extending the chart, the range of each sample may be plotted to determine whether or not the process is in control with respect to variability. This involves computing the range of each subgroup as well as the mean, which is needed for plotting the $\bar{X}$ chart.

CONTROL CHART FOR FRACTION DEFECTIVE

The control chart for variables can be used whenever the quality characteristic can be measured and expressed as a quantity. In many situations, however, the quality characteristics can be expressed only as meeting or not meeting specifications. If an item is either accepted or rejected on the basis

of whether or not it passes a given test, a control chart for fraction defective (also called a p chart) may be used instead of $\bar{X}$ and R charts. Acceptance of an item may be the result of an elaborate analysis, or it may be based on the checking of dimensions by "go" and "not-go" gages. As long as the inspection of an article results in the classification of the item as accepted or rejected, the fraction defective chart may be used to analyze the data.

Control charts for fraction defective may be based on 100 per cent inspection or on samples, although if sample data are used, it is important that a relatively large sample be used. The small samples that are used with control charts for variables are not satisfactory in an analysis of fraction defective.

The data in Table 23-5 will be used to illustrate the use of the p chart. The output of an electrical component was checked in 100 unit lots and the number of rejected items tabulated. The results of the inspection of the first 25 lots, representing a total of 2,500 items, are given in Table 23-5. The total number of defective items was 270, or .108 of the total produced. If the process were under control, this fraction could be used as an estimate of the process fraction defective, (designated as p'). Each of the lots inspected may be considered a sample of 100 from a universe in which the fraction of defective items is .108.

Using equation (29) from Chapter 10, it is possible to compute the three-sigma confidence interval for a sample of 100 taken from a universe that has a fraction defective of .108. (The correction for a finite universe is omitted since the universe is large in relation to the size of the sample.)

$$3\sigma_p = 3\sqrt{\frac{\bar{P}\bar{Q}}{n}} = 3\sqrt{\frac{(.108)(.892)}{100}} = 3\sqrt{.00096336}$$

$$3\sigma_p = (3)(.031) = .093$$

Using the following values, Chart 23-3 was constructed as a control chart and the fraction defective for each of the 25 lots was plotted.

$$\text{Center line} = p' = .108$$

$$\text{UCL}_p = p' + 3\sigma_p = .108 + .093 = .201$$

$$\text{LCL}_p = p' - 3\sigma_p = .108 - .093 = .015$$

Since lots number 5 and 20 had a fraction defective higher than the upper control limit, it was assumed that assignable causes had been operating when these lots were processed. In order to compute a better estimate of the universe fraction defective, these two lots were eliminated from the

computation. The remaining 23 lots contained 227 defective items out of a total of 2,300 inspected, or a fraction defective of .099. This new value of p' was used to recompute the value of $3\sigma_p$ and the control limits.

$$3\sigma_p = 3\sqrt{\frac{P\bar{Q}}{n}} = 3\sqrt{\frac{(.099)(.901)}{100}} = 3\sqrt{.00089199}$$

$$3\sigma_p = (3)(.0299) = .0897$$

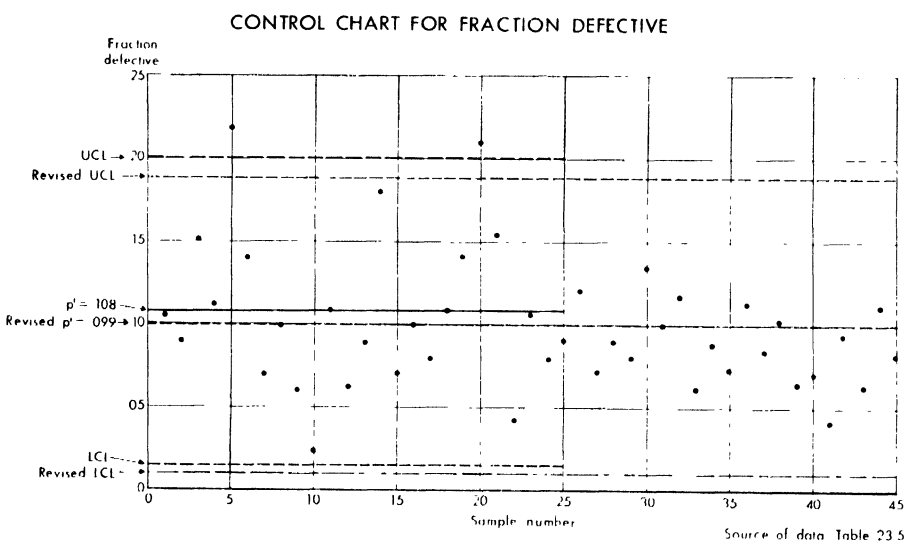
$$\text{Center line} = p' = .099$$

$$\text{UCL}_p = p' + 3\sigma_p = .099 + .090 = .189$$

$$\text{LCL}_p = p' - 3\sigma_p = .099 - .090 = .009$$

The new value of p' and the new control limits were plotted on Chart 23-3 and were extended beyond the first 25 lots to be used to check each succeeding lot for assignable causes of variation in the fraction defective. The fraction defective for each of the next 20 lots was plotted and a definite downward trend was evident. On the assumption that a smaller fraction defective could be maintained in the process, a new center line and control

Chart 23-3



limits should be computed and an effort made to keep the process in control at this better quality level. It has been a common occurence for the fraction defective to drop after a control chart is put into use; and when this happens, the level of p' is usually revised and an attempt is made to hold the new quality level.

Table 23-5

NUMBER OF DEFECTIVE ELECTRICAL COMPONENTS IN 25 LOTS OF 100

Lot number	Number inspected	Number defective *np*	Fraction defective *p*
1.....	100	11	.11
2.....	100	9	.09
3.....	100	15	.15
4.....	100	11	.11
5.....	100	22	.22
6.....	100	14	.14
7.....	100	7	.07
8.....	100	10	.10
9.....	100	6	.06
10.....	100	2	.02
11.....	100	11	.11
12.....	100	6	.06
13.....	100	9	.09
14.....	100	18	.18
15.....	100	7	.07
16.....	100	10	.10
17.....	100	8	.08
18.....	100	11	.11
19.....	100	14	.14
20.....	100	21	.21
21.....	100	16	.16
22.....	100	4	.04
23.....	100	11	.11
24.....	100	8	.08
25.....	100	9	.09
Total.....	2,500	270	...

SOURCE: Confidential.

When the subgroup size varies, new control limits should be computed for each subgroup. The control limit will be represented by a broken line, moving farther from the center line when the subgroup is smaller, and closer to the center line for larger subgroups. Except for the additional work of computing the control limits for each subgroup, the interpretation of the *p* chart offers no new problems. It is possible to compute the control limits for a subgroup of average size and then compute the control limits for individual subgroups only when it is not possible to determine by inspection where the fraction defective for a given lot falls with respect to the control limit.

ACCEPTANCE SAMPLING

The use of sampling inspection by a purchaser to decide whether or not to accept a shipment of product is known as *acceptance sampling*. A sample of the shipment is inspected and if the number of defective items is not more than a stated number, known as the *acceptance number*, the shipment is accepted. If the number of defective items exceeds the acceptance number, one of two procedures may be used. In an *acceptance-rejection* sampling plan, the shipment is returned to the supplier. In an *acceptance-rectification* plan, all of the items are inspected and the defective ones removed, with the cost of the additional inspection usually charged to the supplier.

Acceptance-rejection sampling plans were developed for use by the military services in procuring materials. The acceptance number is selected to give adequate protection against accepting lots below the required quality level. It is the responsibility of the supplier to correct any lot that fails to pass inspection.

Acceptance-rectification plans are used in industry to protect the purchaser against inferior quality. If a supplier delivers materials that have too large a percentage of defective items, the purchaser protects himself by one hundred per cent inspection. As long as the sampling inspection shows no more defectives than the acceptance number, the buyer takes the lot on the basis of the sample and the supplier is saved the cost of complete inspection.

Sampling acceptance plans make possible a prompt decision to accept or reject a lot, with knowledge of the probabilities of making a mistake. Deciding on which plan to use requires a decision as to the risk one is willing to take in making two types of mistakes. A mistake may be made by (1) accepting a lot as satisfactory when in fact it is below the quality level, or by (2) rejecting a lot as unsatisfactory when it is of acceptable quality.

Sampling acceptance plans may be classified as single, double, or sequential, depending upon whether one, two, or more samples are taken before reaching a decision. In *single sampling*, the decision is made on the basis of the evidence furnished by one sample. *Double sampling* provides a method for taking a second sample if the results of the first sample are not conclusive. A lot may be rejected on the basis of the information supplied by the first sample if it is bad enough, or it may be accepted after the first sample has been taken if it is good enough. If the first sample does not give a clear enough indication of the quality of the product to make a decision, a second sample is taken. *Sequential sampling*, which provides for taking as many samples as needed to reach a decision,

results in small samples being used when the quality of the product is either very good or very bad. A larger sample is taken only when the quality of the product is close to the rejection level. Sequential sampling plans generally result in less total inspection for a given degree of precision. This type of sampling, developed first in the inspection of manufactured products, is being adapted to all types of situations where the decision between two possible actions is made on the basis of sample data.

STUDY QUESTIONS

1. What is the advantage in using the term "statistical quality control" rather than the shorter term "quality control"?

2. Why is it important to distinguish between variations due to assignable causes and those due to random causes?

3. It has been said that quality must be built into a product and that a control chart cannot cause a product to have high quality. Although it is true that a control chart itself does not put quality into a product, how does it make a contribution towards better quality?

4. Would a control chart based on samples of 8 give better control of quality than a chart based on samples of 4? Give reasons for your answer.

5. Explain the significance of the mean of a sample falling above the upper control limit of a $\bar{X}$ chart.

6. Explain the significance of a sample range falling above the upper control limit of an R chart. Would you expect both the mean and the range of a sample to fall above the upper control limits of their charts, or would it be more likely for only one to exceed its upper control limit?

7. Why does quality control practice in the United States use the three-sigma control limits instead of some other multiple of sigma?

8. What are the advantages and disadvantages of computing the control limits of the $\bar{X}$ chart from $\bar{R}$ and $\bar{\sigma}$? Which method of computation is more common?

9. Explain the significance of a control chart for fraction defective.

10. What is the significance of a fraction defective falling below the lower control limit? What action should be taken when this happens?

11. In what way does statistical quality control represent a significant departure from the older philosophy of inspection?

12. Why did World War II provide a strong incentive for adopting statistical quality control practices?

13. What is meant by the statement that control charts are primarily a diagnostic device?

14. Assume that a process can be kept in control with a dispersion about the mean dimension represented by a standard deviation of .0032 inches. The standard deviation was computed from a large sample of items manufactured while the process was in control with respect to the mean dimension. How could this information be used in determining the tolerances to set for this dimension?

PROBLEMS

1. The average length of life of the incandescent lamps manufactured by a certain plant was estimated from a large sample to be 1,406 hours, and the standard deviation was estimated at 220 hours. These values may be considered to be the universe mean and the standard deviation. Compute the three-sigma upper and lower control limits of a $\overline{X}$ chart based on this information, using samples of 5.

2. Thirty samples of 4 items each were taken from the output of a machine, and a critical dimension was measured. The mean of the 30 samples ($\overline{\overline{X}}$) was .6550 inches, and $\overline{R}$ for the 30 samples was .0021 inches. Compute the upper and lower control limits for the $\overline{X}$ and R charts.

3. A large sample of a product gave an average fraction defective ($\bar{p}$) of .073. Compute the upper and lower control limits for lots of the following sizes:

(a) 100
(b) 175
(c) 450

4. The table at the top of the opposite page gives the results of taking 20 samples of 4 cans of soup from the output of an automatic filling machine. The data recorded represent the weight in hundredths of an ounce less 11.00. (For example, the first item in sample number 1 is 46, which represents a weight of 11.46 ounces. The decimal has been omitted to simplify calculations.)

(a) From the data in the table, compute the trial limits for $\overline{X}$ and R charts. Plot these charts.
(b) Do the measurements show statistical control? If the charts show the process out of control, compute revised control limits; if not, extend the control limits already computed.

5. The table at the bottom of the opposite page gives the results of inspecting 10 additional samples of 4 each, from the filling machine described in Problem 4.

(a) Plot the proper values on the $\overline{X}$ and R charts.
(b) Explain how the charts will be used by management.

6. (a) Compute the standard deviation for each of the samples in Problem 4 and using the σ values, prepare a σ chart.
(b) Plot the standard deviations of the 10 samples in Problem 5. Does the chart show statistical control?

WEIGHT OF CANS OF SOUP IN EXCESS OF 11 OUNCES
(IN HUNDREDTHS OF AN OUNCE)

Sample no.	Weight of individual items				ΣX	ΣX^2	$\bar{X}$	R
1	46	31	43	34	154	6,082	38.50	15
2	20	38	32	28	118	3,652	29.50	18
3	39	17	29	10	95	2,751	23.75	29
4	28	34	38	18	118	3,708	29.50	20
5	25	23	17	36	101	2,739	25.25	19
6	42	33	47	13	135	5,231	33.75	34
7	27	38	22	20	107	3,057	26.75	18
8	30	26	25	31	112	3,162	28.00	6
9	32	28	46	27	133	4,653	33.25	19
10	21	39	28	33	121	3,835	30.25	18
11	58	34	36	56	184	8,952	46.00	24
12	34	20	13	15	82	1,950	20.50	21
13	33	29	42	41	145	5,375	36.25	13
14	22	32	39	26	119	3,705	29.75	17
15	22	23	45	20	110	3,438	27.50	25
16	24	26	47	30	127	4,361	31.75	23
17	35	51	29	27	142	5,396	35.50	24
18	35	11	33	24	103	3,011	25.75	24
19	28	23	31	43	125	4,123	31.25	20
20	35	39	17	29	120	3,876	30.00	22
Total	..	..	..	..	2,451	83,057	612.75	409

SOURCE: Company records.

WEIGHT OF CANS OF SOUP IN EXCESS OF 11 OUNCES
SAMPLES 21–30
(In hundredths of an ounce)

Sample no.	Weight of individual items			
21	47	32	44	35
22	35	23	10	5
23	37	32	36	39
24	31	16	17	10
25	21	17	9	29
26	52	36	19	50
27	20	46	32	50
28	26	5	43	20
29	38	40	49	37
30	47	32	64	42

SOURCE: Company records.

7. Make two estimates of the standard deviation of the universe from which the samples in Problem 4 were taken. Base one estimate on $\bar{\sigma}$ and the other on $\bar{R}$. Which of these two estimates do you consider more accurate?

8. The following table gives the number of defective units of product in ten lots of 400 units each. Compute the control limits and plot a p chart. Does the chart show statistical control?

NUMBER OF DEFECTIVE UNITS IN 10 LOTS

Lot no.	No. of units inspected	No. of defective units
1	400	17
2	400	15
3	400	14
4	400	24
5	400	9
6	400	4
7	400	19
8	400	12
9	400	11
10	400	14

SOURCE: Company records.

24 COMMUNITY ECONOMIC SURVEYS[1]

The period since the end of World War II has been one of unusually rapid growth in the United States, and during the postwar years a great many changes have taken place in the economy. Improvements in agricultural technology have resulted in much greater production per man-hour with the result that much less labor has been needed on farms. The rapidly expanding manufacturing and service industries have drawn persons from the rural areas to the cities, but at the same time manufacturing plants have been looking for new locations in all parts of the country. The availability of raw materials, a supply of labor, and better access to markets have been among the factors that have induced industry to look for locations in areas that had previously been primarily engaged in agriculture. The movement of industry into new areas has stimulated the collection and analysis of economic data on a regional basis. This type of statistical analysis makes use of many of the techniques described in previous chapters and calls for a wide familiarity with the sources of published statistical data.

NATURE OF COMMUNITY ECONOMIC SURVEYS

The movement from the farms to the cities and the spreading of industry into new areas have been proceeding at such a rapid pace since the end of World War II that many communities have not been aware of the revolution that is taking place. Where the leaders in a community have realized the nature of the change that has been occurring, they have found that their first need was for complete information about the resources of the area in order to plan intelligently for its future development. Information about the resources of an area constitutes an economic survey and should provide facts that explain the present basis of the community's economic life. An economic survey is basic to any intelligent effort being made to develop the economy of an area in the light of the changing environment, but the survey should be more than a collection of data on all aspects of the economy of the community. In one sense the collection of information is basic to any use that may be made of the data, but the data

[1]This chapter is based on *An Economic Survey Method for Small Areas* by Alfred G. Dale, Bureau of Business Research, The University of Texas, Austin, 1955.

will be assembled more intelligently if there is some understanding of its potential use. The purpose of this chapter is not only to describe the various kinds of information that need to be collected but also to point out the relative importance of the different kinds of information.

The Economic Base of an Area

Not all types of economic activity are of equal significance in the development of a community, and the first step should be to establish the relative importance of the various elements which comprise the economic base of an area. Some types of economic activity are more important than others in their effect upon an area, and prime consideration should be given to identifying the type or types of activity upon which the economy of the area has been developed. For example, the sale of goods and services at retail constitutes the major business activity of most communities, but the level of this type of business is related to the total amount of income generated in the community. If more retail stores come into existence, it merely means that the total business is divided among more concerns; the establishing of new retail outlets in itself does not stimulate additional economic activity after the necessary investment in building, equipment, and inventory has been initially made. All local services (shoe repair, beauty parlor, dry cleaning, etc.) are in the same general category as the retail store.

It is not intended to imply that retail and service businesses are not important economic activities, for they employ a large number of persons, and the goods and services distributed by them are essential to the community. The important factor to consider is that in an individual community, retail and service businesses represent a type of economic activity that is essentially dependent on the spending and respending of income generated by producing and selling goods or services to persons outside the area and thus bringing income into the community. Our modern economy is built on specialization in the production of goods, which means that in general something must be exported from an area in order to create the income needed to buy goods and services from other areas. The community in which the people made their living by doing each other's laundry cannot exist in the modern economic world.

Basic Industry and Service Industry

In this discussion the term *basic industry* will be used to describe the type of activity that is based on the sale of goods and services outside the local area and thereby brings income into the local area. Economic activity that is derived from the spending of the income brought into the area by

basic industry will be designated, for purposes of this discussion, as *service industry*. Frequently the terms *nonbasic industry* and *service industry* are used interchangeably. Service industry includes, in addition to retail and service businesses, the local manufacturing concerns that sell their products entirely in the local market. It should be pointed out that basic industry is not inherently superior to service industry, as has been held by certain economic theories. The purpose of all economic activity is to supply consumers with goods and services, and every stage in this activity is an essential occupation.

It is important, however, in studying the existing economic development and in deciding on a course that will stimulate growth in the economic activity of an area, to distinguish between the two types of activity. The only way in which the growth of a community can be stimulated is to increase the activity of the basic industry in the area. When basic industry expands, it increases the income flow in the community, and the accompanying spending and respending generate an increase in the amount of sales of service industries. If a concern that has been selling its product outside the area reduces the volume of its business, a reduction in the volume of service business supported by this activity inevitably follows. Any study of the economy of an area should, therefore, be primarily concerned with the trends and opportunities in the basic industries. Action designed to stimulate total economic development should be applied to the development of basic industry, for the resultant development of the service industry will be almost automatic.

The determination of the basic industry of an area cannot be made accurately without a detailed study of the source of business for all major types of activity. The study almost inevitably involves collecting data from individual business concerns, although an approximation may be secured by analyzing the various categories of economic activity in the area. Agriculture is largely basic, since the major portion of farm products is sold in central markets rather than being consumed locally. Mining, lumbering, and fishing are normally classed as basic industries. Manufacturing plants that ship their products outside the area also represent basic industry. Spending by government agencies (other than local government units) is basic income to the community. Educational institutions, hospitals, resorts and tourist attractions, transportation facilities serving more than the local community, and even retail trade, when it draws customers from other trading areas, should be considered basic industries.

If the amount of employment in industries classified as basic for a given community is subtracted from the amount of total employment in the area, the ratio between basic- and service-industry employment

can be determined. The more goods and services that a community produces, the higher will be the ratio of service-industry employment to basic-industry employment. Such an analysis will yield important information on the nature of the economy and will serve as the basis for action directed towards changing the nature of the economic activity of the area.

After a community has determined as accurately as possible the nature of the economic development of the area, the next step is to study the resources of the area in the light of information on the demand for products in the markets that are accessible to the area. The study will form a basis for an intelligent effort to stimulate the development of those basic industries that appear to have a chance of operating successfully. The efforts may take the form of trying to persuade businesses to build new establishments in the community, or of suggesting that existing concerns be helped to expand the size of their operations.

USERS OF COMMUNITY ECONOMIC SURVEYS

One of the important users of economic surveys is the local chamber of commerce, which usually represents the organization that is specifically charged with the responsibility of promoting the economic development of a city. Most of the states also have a public body that deals with the problem of bringing industry into the state. Certain businesses that benefit directly from the industrial development of their areas are important users of economic surveys, and frequently the businesses conduct the surveys in order to secure the information needed. Banks, railroads, and electric power and gas companies are all actively engaged in promoting the economic development of the regions they serve; and many companies have well-staffed departments devoted entirely to industrial development work. The community leaders of any city that wants to develop its basic industries should work with the railroads and public utility companies that serve the city. Banks in the large cities usually have special departments working on industrial development, and in many small communities the banker is one of the most effective promoters of economic development. Community economic development is obviously a profitable activity for a bank, since its business will grow as the total economic activity of the area grows.

The agencies described above attempt to keep complete, up-to-date statistical data on the economy of their area. The information is made available to any business that is interested in the area, and it is common practice for the agencies to collect any specific data that are needed by a business in making a decision to expand or build a new plant. Although the information on the economy is always rather general, it is the basic

information upon which the decision to locate an establishment in a given region or city is based. The selection of a specific site upon which a plant or warehouse is to be built requires very specific data that may be compiled by the business making the decision, along with local agencies that are useful in this operation. Some large business organizations that are expanding rapidly have elaborate departments devoted entirely to the location of new building sites.

Many business concerns and public agencies also find economic surveys valuable in making long-range plans. Providing an adequate water supply for a growing community requires long-range forecasting of the population growth and economic trends in the area. The engineering projects required to supply water are major capital investments which should be based on as complete information as it is possible to secure. The provision of adequate schools, hospitals, and other municipal services within a community requires information in the form of surveys upon which to base decisions. Likewise, investment dealers are interested in economic surveys of communities that are trying to market bonds.

In general the statistical methods used in making economic surveys are not complicated. Most of the effort goes into compiling the needed information, although it has been pointed out earlier that a clear definition of the objectives of the study makes for greater effectiveness in the collection of data. The collection of adequate data on small areas is expensive; thus, there is a general lack of comprehensive information on small areas. The Bureau of the Census publishes its major collections of data by small areas, but many current statistical series are not classified in enough detail to give the information needed. Anyone desiring to make an effective economic survey for a community is usually faced with the problem of collecting original data.

BASIC DATA FOR COMMUNITY ECONOMIC SURVEYS

The following outline of important factors to be considered in preparing an economic survey is merely one of many possible ways of organizing the data. Specific problems may require a different organization of the needed information, but the basic requirements for all problems are not greatly different.

Regional History and Natural Features

It is easy to give undue emphasis to the history of an area, but an understanding of the manner in which economic development has proceeded in the past is generally helpful in analyzing the future trends of the area. The natural features of an area furnish the base upon which the economy rests.

Historical data. A condensed historical review of the area should provide a starting point for the detailed discussion of the current economic position. The date and circumstances of the first settlement may be of significance for present analysis, or the character of the economy may have changed so much that the factors influencing the first settlement are no longer important. In any event, a brief historical sketch seems to be an appropriate introduction to an economic survey.

Geographic setting. A brief description of the geography of the area, indicating its position relative to other major centers of population and industrial activity, serves to give a general introduction to the economic problems of the area. A simplified map may be an effective method of introducing the significant features of the area.

Climate. Information on climate should include statistical data on temperature and rainfall covering a number of years. The U. S. Weather Bureau publishes detailed information and averages based on long periods. Statistical tables will present most of this information, and there is little need for extended discussion. When agriculture is an important basic industry, the data on climate should cover the length of the growing season and the dates of the first and last killing frosts.

Topography. A review of the land conformation of the study area is important for practically all the purposes for which an economic study will be made. A topographic map is probably the most efficient method of showing the needed information. Any decision to locate a business establishment in an area will be concerned with the nature of the local terrain.

Soils and geology. The characteristics of the area being studied will determine the importance of this type of information. Where agriculture is an important basic industry, the pattern of land use will be related to the types of soils in the area. Where mineral resources are important to the economy of the area, a detailed inventory of the raw materials available will be of strategic importance. The increased production of goods has magnified the demand for raw materials, and if an area possesses commercial deposits of minerals, complete information on these materials should be included in any economic survey.

Water supplies. The growth of population and industry has increased tremendously the consumption of water, both by households and by industrial concerns. Information on the supplies of both surface water and ground water is of vital importance in any compilation of information on an area. Since many manufacturing operations require rather precise control of the quality of water used, it is worth while for a community to have detailed information on the chemical composition of the water supply.

Population and Labor Force

The size and characteristics of the population of an area represent fundamental facts that are of interest to everyone who might make use of an economic survey. People make up the labor force and the consumers in the area; both of these factors are of vital importance in the decision to locate a business. The discussion in Chapter 4 concerning sources of published information gave most of the sources of information on population and labor force. The decennial Census of Population must be relied upon for detailed information, although in a dynamic area the last census may be badly out of date. Communities have a great deal of interest in estimating population between census dates, but the methods used give estimates for total population only. Information on the characteristics of the population and changes in employment are much more difficult to secure.

Characteristics of the population. In addition to giving the total population by small areas, the Census of Population gives information on the age distribution of the population, racial characteristics, amount of education, composition of households, home ownership, and income. Except for the fact that the information is available only at ten-year intervals, the Census of Population provides all the information on population that can be used effectively. In a community that is undergoing rapid change, some thought should be given to the possibility of securing information more frequently than provided by the Bureau of the Census.

Employment pattern. The analysis of the population with respect to the pattern of employment involves a classification of individuals according to whether or not they are in the labor force. The labor force includes employed individuals and persons who are able to work and are looking for work but who do not have a job. The group not in the labor force includes those unable to work, inmates of institutions, students, retired individuals, and housewives. Only persons 14 years of age and over are included in the labor force.

The Census of Population classifies employment according to major industrial and commercial groups, thereby giving a detailed picture of the economic characteristics of the community. Employment is also classified according to major occupation groups, such as clerical, sales, and professional workers. The functional composition of the labor force is one of the major factors to consider in planning for the future economic development of an area.

Since the census data rapidly become out of date, many state employment agencies make local labor market surveys for communities. This type of survey gives up-to-date information on local labor supply and employment and is basic information for use in any industry location study.

The Bureau of Old Age and Survivors Insurance also tabulates information on payrolls and employment by kinds of business for all employers subject to the provisions of the Social Security Act.

Housing and Construction

The characteristics of the housing in a community are important factors in the economic development of the area. A favorable housing situation may be an asset to a city trying to attract industry, and in general, the quality of available housing is an important element in any study of the community.

The Census of Housing gives complete statistical data on occupancy, physical facilities, and value of housing. The information includes the proportion of occupied dwelling units that are owner occupied; the number of vacant dwelling units; the state of repair of housing units; classification according to year built; the percentage of dwelling units with refrigeration, radio, and television; and many other data that are useful under certain circumstances.

Since most cities require the issuance of a building permit for new construction or major repairs, information is available on construction of new housing units. The analysis of building permits makes it possible to keep the information from the Census of Housing up to date. An economic survey should compile a detailed record of building as reflected in the permits issued.

Civic Information

This category might be considered an analysis of the city not only as a place to carry on business but also as a place to live. Civic information should be secured about such diverse subjects as the schools, churches, recreation facilities, radio and television stations, newspapers, libraries, and cultural entertainment. A story is told of a large company that sent a representative to a city to determine whether or not it had a little theater, since certain executives of the company were interested in the little theater and did not want to move to a city where this interest could not be continued. An increasing number of businesses are giving serious attention to the living conditions in a city because of their influence on the labor supply. For example, climate has become an important factor in recruiting certain kinds of personnel, and industries that are not restricted as to their location by such factors as markets or availability of raw materials are locating in areas with an attractive climate.

Other civic factors that directly affect the desirability of a city as a business location include such things as local tax policy, police and fire protection, and regulatory practices. Tax laws and assessment practices

are particularly important to the business concern, and complete information on all such matters should be included in any economic survey. Somewhat less tangible is the attitude of city officials and community leaders towards industry. An unfriendly feeling towards business concerns is a serious liability to a community trying to secure new industry, for this attitude will be reflected in many ways. However, if the atmosphere is friendly towards business, the community possesses an asset in promoting industrial development.

The topics discussed above, although among the important ones, do not include all of the information that may at one time or another be important in the decision to locate a business in a given area. Any factual data on the general subject of the city as a place to live and in which to transact business is valuable material for the civic information of an economic survey.

Business Facilities

The decision to build a manufacturing plant must take into consideration the character of a number of services and facilities available at the proposed location. The most important of these are transportation, public utilities, financial services, industrial sites, and, in specific industries, certain specialized services that are peculiar to the business. An example of the latter type of facility is the geophysical services used by the oil industry.

Transportation. Since accessibility to markets is one of the most important factors in the location of an industrial concern, information on the transportation facilities of a community is necessary in making decisions on industrial location. An economic study should include a complete inventory of railroads, motor freight carriers, water transportation, air transportation, and warehouse facilities. Schedules and rates to selected points should be available for all forms of transportation. Adequate transportation and reasonable rates are a necessity in securing new industry; and if a community has an unusual advantage in transportation facilities, this fact should be emphasized in analyzing the economic potential of the area.

Public utilities. The ability of the local public utilities to meet current requirements and their potential for expansion to meet new demands should be studied in the preparation of an economic survey. The existing capacity of the utilities to supply electricity, gas, and water, and the rates for the various types of consumers should be given in complete detail. In addition to this information, the sources from which the water supply is drawn should be shown, and any limitations on the availability of additional supplies should be indicated. The willingness and ability of the

electric utility to provide additional generating capacity, if needed, should be reported. Similarly, the report should show the size of the gas utility's reserves and its ability to supply an increased demand.

Financial services. The availability of bank credit and other financial services needed by business should be described. The number of banks, their capital structures, volume of deposits, and amount of loans and investments should be given for a period of years. Descriptions should be given of other types of financial institutions, such as savings and loan associations, finance companies, mortgage loan companies, organized exchanges, security dealers, and insurance brokers.

Industrial sites. It has become almost a necessity for the community desiring to attract industry to have good industrial sites available at a reasonable cost. Any industrial concern that is seriously considering a given location must be assured that an adequate site is available at a fair price. If the community does not make any provision to set fair prices and to have such sites available, it is possible that all the owners of good sites will attempt to sell at too high a price. Such a situation can easily result in a different location being chosen by the concern.

The preferred method of insuring that good sites will be available is to create an industrial district in the community. In many communities a railroad provides industrial sites adjoining its lines. In other cities the industrial district may be owned by some type of nonprofit organization such as an industrial foundation. Besides insuring that sites will be available at reasonable prices, such an arrangement has the advantage of providing proper zoning for different kinds of business and also making available the business services needed by a business, such as railroad spurs, roads, public utilities, and waste disposal. Some industrial foundations will provide buildings for either rent or purchase.

Regardless of the method used by the community to make industrial sites available, an economic survey should describe in detail the situation with respect to sites. If no provision has been made for establishing industrial sites within the city, the economic survey might result in steps being taken to remedy this deficiency. The competition for new industries is so keen that it is almost hopeless to try to attract new industry without some provision being made for industrial sites.

Trade and Services

Since retailing and personal and professional services do not exert the same influence on the level of economic activity in a community as basic industry, it is generally not necessary to make an exhaustive analysis of the

trade and services segment of the economy. A complete statistical summary of trade and service is given in the Census of Business. If a number of years have elapsed since the latest census, more recent data may be obtained from *Sales Management, Survey of Buying Power.*

Wholesale trade may be included in the summary being made, although a careful examination of this type of business should be made to determine whether or not some of the wholesale trade is basic economic activity. In a few cases it may also be discovered that some retail sales represent basic industry. When certain retail stores become so well known that they bring in business from outside the normal trading area of the city, the sales should be classed as basic industry; however, such retail sales are probably exceptional cases that will not be encountered frequently.

Basic Industries

Since the growth of a small area reflects the expansion of the basic industry in the area, the portion of an economic survey that deals with basic industries is one of the most strategic sections. An inventory of the industries that may be considered basic in the area should be followed by an analysis of the advantages that the area offers for each type of basic industry.

Agriculture. The importance of agriculture varies considerably from one community to another, but for all cities, except the largest, the agricultural activity in the trading area represents a significant basic industry. The relative importance of agriculture in the total economic pattern of an area is shown by the amount of employment and the value of farm products. Over a period of time agricultural trends are reflected in changes in land use and in the value of the various crops, livestock, and livestock products sold.

Department of Agriculture reports and the Census of Agriculture represent the most important sources of data on agriculture. The study of statistical data on all phases of agriculture, if extended over a long enough period to give an indication of trends, should provide a basis for drawing conclusions on the future importance of the agricultural phase of the economy of the area.

Other raw-material industries. The importance of the other industries producing raw materials varies widely from one area to another. Mining, lumbering, and fishing represent important basic industries in some regions; and whenever the development of any extractive industry is considered to be important, all of the available statistical data should be compiled and analyzed.

Manufacturing. In the Census of Manufactures the statistical data on manufacturing industries are not classified according to basic or service industry, so the first step is to compile all the information available on manufacturing in the area. The second step is to analyze the industrial pattern to determine the amount of basic manufacturing and the number of employees engaged in this kind of activity. Although the expansion of basic industry has the greatest effect on the growth of a community, the inventory of manufactured goods sold in the local market can be useful in stimulating local economic activity. The inventory of locally consumed goods will permit the determination of the products that are imported for consumption in the community. Some of the products, which may be either consumer goods or supplies and capital equipment for local businesses, may offer an opportunity for the expansion of manufacturing within the local market. However, goods that would compete with well-advertised brands would have considerably less chance of success in a local market. Since any manufacturing concern that supplies only a local market would be rather small, the establishment of the concern would be particularly well suited for local financing. If the business concern is successful in supplying the local market, there is always the future possibility of expanding to reach neighboring markets.

The inventory of basic manufacturing industry in the area should be studied with particular care to determine which categories offer the most potential for expansion. Industrial classifications that are not represented in the list of basic industries should be checked for suggestions as to possible new industries that might profitably operate in the area. A new industry should be located on the basis of competitive advantage, and attempts to bring in such an industry without definite advantages to offer generally will be unsuccessful. It should be noted that a manufacturing concern supplying local demand would not need so decided a competitive advantage, since the closeness to the market would give the concern an initial advantage over outside competitors. An exception to the principle of locating on the basis of competitive advantage would be the concern that sells advertised brands with strong consumer acceptance, for goods of this type in general would have to compete in the local market in much the same manner as in other markets.

Other types of basic industry. Many communities have specialized basic industries that bring income into the area as effectively as the sale of minerals, farm products, or manufactured goods. Areas with a favorable climate, particularly if combined with scenic attractions and recreational facilities, draw visitors whose expenditures represent basic income for the area. As the length of the work week has decreased, the value of attractions that people enjoy in their leisure time has increased until many communities

depend heavily on recreational types of basic industry. Theaters, athletic events, and any other attraction that brings people into the city to spend money may be considered basic industry.

As mentioned earlier in this chapter, educational institutions that draw students from outside the local area, hospitals that attract patients from outside the community, and government activities that are supported from taxes collected outside the area are all examples of basic industry.

Since 1940, military establishments have increased in number so greatly that they have become important basic industries for many communities in the United States. There are disadvantages in the military establishment as a kind of basic industry, since a major curtailment in military expenditures and activity is felt immediately by all kinds of business in the community. Even business concerns that do not have any direct dealings with either the government or the employees of the establishment feel the effect of a reduction in operations, since all business concerns benefit from the respending of the income brought into the area. Government installations may be considered extremely desirable basic industries to secure for a community, but the fact that their operations are subject to decisions made outside the community is a considerable disadvantage. Many small communities have felt the depressing effects of the sudden closing of a government or military installation, especially when community facilities have been expanded extensively on the basis of the new economic activity.

The examples given above serve the purpose of illustrating the various sources of primary income for a community, and a complete economic survey of an area should point out all of the important sources. If a city has been growing, this growth can best be explained in terms of an increase in the basic industry of the area. Whenever it is predicted that an area will grow, it is necessary to specify what basic industry will support this growth.

ECONOMIC POTENTIAL

After all of the statistical data for a community have been assembled and analyzed, it should be possible to draw tentative conclusions regarding the type of economic activity that gives promise of being successfully expanded in the area. It is generally profitable to separate the factors that give promise of potential growth into a few broad categories and to consider each of these classifications in detail. There are so many diverse combinations of resources in the different communities of the United States that no set pattern can be furnished for finding an industry that can operate successfully in any given area; every situation is a special case and the logical approach is to study all the information available as the basis for a decision.

It is important to understand that it is meaningless to conclude that a community is a "good place for industry" without making the analysis in terms of specific types of industry. It is important to determine the locational factors in which the area has a relative advantage over other areas and to analyze the requirements of different industries to know which ones are good prospects. To do this, it is necessary to draw conclusions from the data previously described in order to arrive at judgments as to the relative advantages of the area with respect to various factors of importance in the location of industry.

One locational factor that may give an area high economic potential is the possession of raw materials needed by industry. Raw-material based industries have been the dominant factor in the development of many major industrial areas. As the population grows and the demand for the output of manufacturing industry increases, new supplies of raw materials are sought; a community that has access to raw materials has a valuable location of great interest to business. In recent years industry has been expanding extensively into regions that offer new sources of needed raw materials.

In a country that has been growing as fast as the United States, there is a continual shift in the market pattern; and when a community's location becomes strategic with respect to a growing market, this becomes an important locational factor. An economic survey may reveal the growth of markets in a region that is giving the community a strong appeal for industry. An intensive study of the region is particularly important in determining the potential for market-oriented industry. The transportation facilities and the pattern of rates play an important part in the evaluation of this factor.

Industries with a high ratio of labor to total costs are attracted to areas with a surplus of labor. The mechanization of agriculture has created a surplus of labor in many regions that previously had little manufacturing activity. Whenever this situation exists, a community should realize that its potential is greatest in the kinds of business that require large amounts of labor. Much of the dispersal of industry throughout the country has been in an effort to take advantage of a labor supply. Communities should not overlook this type of industry in studying their economic potential.

In addition to the major categories of locational factors described above, within any community there is the possibility of discovering some specialized advantage that will attract a particular type of business. The development of the full economic potential of an area requires that thorough study be made of all resources, being certain that all of the advantages of the area have been explored. There may be a tendency to

concentrate on the major features such as raw materials, labor, and markets, with the consequent neglect of valuable assets that are more specialized.

The contribution of the statistician to a study of the economic potential of an area consists of providing complete quantitative measures of the factors that are important in making decisions. The analytical techniques that have been used in this type of analysis in the past have tended to be extremely simple, but there is a discernible trend towards a more elaborate analysis of these factors. The recent development of input-output analysis, or the quantitative analysis of inter-industry relationships, offers promise of becoming a major tool of economic analysis, although there is still an active debate as to the significance of its contribution. Regardless of the amount of statistical analysis used in the study of the economic potential of a community, it is an important type of activity making use of statistical data and one with which a great many businessmen come in frequent contact.

STUDY QUESTIONS

1. What are the important factors to consider in the location of a manufacturing plant? Would the product to be manufactured have any influence on the relative importance of these different factors?

2. Which business organizations in a community have the greatest interest in securing new industry for the community?

3. Where would you look for information on the following subjects if you wanted it for a manufacturing concern that is considering the location of a plant in your area?

(a) Raw materials available.
(b) Amount and kind of labor available.
(c) Living conditions for employees of the company that would move to the community.
(d) Transportation to markets.
(e) Tax rates and assessment practices.
(f) Insurance rates.

4. How would you determine the amount of employment in basic industry and in service (nonbasic) industry in a community?

5. From the standpoint of the over-all value to a community, are some types of basic industry more valuable to a community than others? Give specific examples of different types of basic industry and rate them according to their relative value to the area.

6. Is it possible for a community to continue to grow in population without any growth occurring in basic industry?

POPULATION AND LABOR FORCE, DENTON COUNTY, TEXAS, 1950

Item	Number
Population	41,365
Labor force:	
Unemployed	449
Employed	14,963
Total labor force	15,412
Industry group — total employed	14,963
Agriculture	2,906
Forestry and fisheries	19
Mining	70
Construction	1,570
Manufacturing	1,793
Furniture, lumber, and wood products	111
Primary metal industries	12
Fabricated metal industries	67
Machinery	54
Transportation equipment	368
Stone, clay, and glass	113
Food and kindred products	339
Textile mill products	12
Apparel and other textile products	193
Printing and publishing	401
Chemicals	17
Other	106
Transportation	372
Telecommunications	136
Utilities and sanitary services	179
Wholesale trade	341
Retail trade	2,466
Finance, insurance, and real estate	334
Business services	47
Repair services	279
Private households	460
Hotels and lodging places	108
Other personal services	545
Entertainment and recreation services	171
Medical and other health services	205
Educational services	2,085
Other professional services	189
Public administration	386
Classification not reported	302

SOURCE: Bureau of the Census, *U. S. Census of Population: 1950*, Vol. 1: *Number of Inhabitants*, Part 43, Texas.

7. If a community produced everything it consumed, what would be the significance of the distinction between basic and service industry? Is the trend for individual communities to become more specialized or more self-sufficient with respect to the goods and services they consume?

8. Assume that you are in business in a community of approximately 100,000 and are appointed to the industrial development committee of the chamber of commerce. There is general agreement that the city should attempt to secure more industry. What steps would you propose that the industrial development committee of the chamber of commerce take?

9. How would you go about finding out whether or not there exists an opportunity for a new business to supply goods or services that are at present being bought outside the community?

10. Most of the population growth in the United States since World War II has been in the cities. Rural areas and a very large percentage of smaller towns have declined in population. How do you explain this shift in population?

PROBLEM

Denton County, Texas, is situated directly north of the Dallas–Fort Worth area, about midway between that area and the Oklahoma border. The south county line is 9 miles north of the city limits of Dallas and 17 miles north of Fort Worth. The population in 1950 was 41,365, an increase of 7,707 over the 33,658 reported in 1940. Both of these figures are census counts. The excess of births over deaths in the 10 years between the 1940 and 1950 censuses was approximately 3,732, indicating that immigration into the county was substantial. However, the fact that Denton is the site of two large state colleges requires an adjustment in these figures, since the 1950 census counted students as living where they went to school, but the 1940 census counted students at their homes. The 1940 population was adjusted by adding students from out of the city to the reported population, giving an estimated total population of 38,482 according to the 1950 definition.

The population increase was less than the natural increase, and the industrial committee of the chamber of commerce interpreted this to mean that persons were leaving the county to look for better economic opportunity. If the population of the county was to increase as fast as the natural increase, it would be necessary to secure additional industrial activity to furnish jobs. As a first step in trying to attract new industry, the following analysis of the county's economy was made.

The table on page 574 gives the composition of the labor force in the county: the number unemployed, the number employed, and the classification of employment by industry groups.

A detailed analysis of the concerns doing business in the county indicated that out of the total employment in the county, 5,839 should be considered basic employment, distributed by industries as shown below. As nearly as could be

determined, the employees classified as basic produced goods or services for which money from outside the county was paid.

Agriculture, forestry and fisheries, and mining (all employment basic).
Furniture, lumber, and wood products (50 employees basic).
Fabricated metals (all employment basic).
Machinery (all employment basic).
Transportation equipment (all employment basic; residents of Denton County worked in the plants in the Dallas-Fort Worth area).
Stone, clay, and glass (101 employees basic).
Food products (285 employees basic).
Apparel and other textile products (all employment basic).
Printing and publishing (306 employees basic).
Education (1,420 employees of the two colleges considered basic).

(a) Using the information given above, compute the ratio of service (nonbasic) employment to basic employment in Denton County in 1950.
(b) Estimate the increase in population that should eventually result from the location of a manufacturing plant employing 300 persons and exporting all its product. (Show all your calculations.)

APPENDIXES

appendix A ♦ Glossary of Symbols

appendix B ♦ Statistical Formulas

appendix C ♦ Table of Squares, Square Roots, and Reciprocals

appendix D ♦ Table of Values of ΣX^2 and ΣX^4 for Computing Least Squares Trend Lines

appendix E ♦ Use of Logarithms

appendix F ♦ Table of Four-Place Logarithms

appendix G ♦ Table of Ordinates of the Normal Curve and Areas Under the Normal Curve

appendix H ♦ Selected Reading References

APPENDIX A

GLOSSARY OF SYMBOLS

PART II — STATISTICAL DESCRIPTION

X	A variable quantity.
$\bar{X}$	Arithmetic mean of the X variable.
Y	A variable quantity.
$\bar{Y}$	Arithmetic mean of the Y variable.
N	Number of values of a variable.
Σ	The Greek letter *Sigma*, which means "the sum of."
f	Number of values of a variable falling in a particular class in a frequency distribution.
m	Value of the mid-point of a class of a frequency distribution.
A	Value of an arbitrary origin, used in the computation of the mean.
d	The deviation of a value of the variable or the mid-point of a class from an arbitrary origin. ($d = X - A$, or $d = m - A$)
d'	The deviation of the mid-point of a class from an arbitrary origin, expressed in class interval units.
c	Difference between the arbitrary origin and the mean. ($c = \bar{X} - A$)
i	Class interval of a frequency distribution, when all the classes are the same size.
w	Weight assigned each value of a variable.
G	Geometric mean.
log	Logarithm.
f_1	Frequency of the class preceding the modal class.
f_2	Frequency of the class following the modal class.
f_m	Frequency of the modal class.
L	Lower limit of the modal class.
d_1	Difference between frequency of modal class and frequency of class immediately preceding modal class.
d_2	Difference between frequency of modal class and frequency of class immediately following modal class.
Q	Quartile deviation.
Q_1	First quartile.
Q_3	Third quartile.
x	Deviation of a value of the X variable from the mean. ($x = X - \bar{X}$)
σ	Standard deviation.
s	Root-mean-square deviation from an arbitrary origin.
V	Coefficient of variation.
Y_o	Maximum ordinate of the normal curve.
π	Ratio of the circumference of a circle to its diameter. The value of π to five decimal places is 3.14159.
e	Base of the Naperian or natural logarithms. The value of e to five decimal places is 2.71828.

PART III — STATISTICAL INDUCTION

N Size of the universe.
n Number of values of a variable in a sample from a universe of size N.
σ Standard deviation of a sample.
$\hat{\sigma}$ Standard deviation of a universe.
$\sigma_{\bar{x}}$ Standard deviation of a distribution of the means of samples, known as the standard error of the mean.
σ' Estimate of the standard deviation of a universe derived from a sample.
σ_p Standard deviation of a distribution of percentages, known as the standard error of a percentage.
P Percentage.
Q Percentage, where $Q = 100 - P$.
M Mean of the universe.

PART IV — ANALYSIS OF BUSINESS CHANGE

N Number of years to which a trend line is fitted; also used to refer to the number of different series in an index.
X Time scale for data classified as time series.
Y Values of the variable in a time series.
a The value of the Y variable when $X = 0$; the Y intercept.
b The amount of change in the Y variable that is associated with a change of one unit in the X variable; b represents the slope of a straight line.
Y_c Computed values of variable Y; in time series, this symbol represents trend values.
S_1, S_2 Sums of the two subperiods of the Y values of a time series; used in fitting a straight line by the method of semiaverages.
c A constant that determines the curvature in the second-degree parabola equation.
S_1, S_2, S_3 Sums of the three subperiods of the Y values of a time series; used in fitting a Gompertz curve.
n One-third of the number of years to which a Gompertz curve is fitted $\left(n = \frac{N}{3}\right)$.
P_0 Prices in the base year or period.
P_1 Prices in the current period for which the index is being computed.
Q_0 Quantities in the base year or period.
Q_1 Quantities in the current period for which the index is being computed.

PART V — ANALYSIS OF RELATIONSHIP

X Values of the independent variable in simple correlation.
Y Values of the dependent variable in simple correlation.
Y_c Computed values of variable Y; based on the regression equation.
X_1 Values of the dependent variable in multiple and partial correlation.
X_{1c} Computed values of variable X_1; based on the multiple regression equation.
$X_2, X_3 \ldots X_n$ Values of the independent variables in multiple and partial correlation.
a, b Constants of a regression equation in simple correlation.

σ_y Standard deviation of the Y variable.
S_y Standard error of estimate of the Y variable.
k^2 Coefficient of nondetermination.
r^2 Coefficient of determination.
r Coefficient of correlation.
k Coefficient of alienation.
ρ_{yx} Index of correlation, based on a nonlinear regression equation.
$\bar{Y}_k$ Means of the columns in a correlation cross-classification table.
σ_m^2 Variance of the means of columns of a correlation cross-classification table about the mean of the Y variable.
η Correlation ratio.
$b_{12.3}$ Coefficient of net regression between dependent variable X_1 and independent variable X_2 when variable X_3 is also taken into account.
$b_{13.2}$ Coefficient of net regression between dependent variable X_1 and independent variable X_3 when variable X_2 is also taken into account.
$S_{1.23}$ Standard error of estimate for dependent variable X_1 when estimates are based on independent variables X_2 and X_3.
$R_{1.23}$ Coefficient of multiple correlation between dependent variable X_1 and a combination of independent variables X_2 and X_3.
$r_{12.3}$ Coefficient of partial correlation between variables X_1 and X_2 when variable X_3 is held constant.
$r_{13.2}$ Coefficient of partial correlation between variables X_1 and X_3 when variable X_2 is held constant.

PART VI — APPLICATIONS IN BUSINESS

σ Standard deviation.
σ^2 Sample variance.
$\bar{\sigma}^2$ Average sample variance.
σ' Estimate of the universe standard deviation.
$\sigma_{\bar{x}}$ Standard error of the mean.

$\mathrm{UCL}_{\bar{x}}$ Upper control limit of an $\bar{X}$ chart.
$\mathrm{LCL}_{\bar{x}}$ Lower control limit of an $\bar{X}$ chart.

$\bar{X}$ Mean of the sample means.
$\bar{\sigma}$ The average of a set of standard deviation values.
c_2 The ratio between the expected value of $\bar{s}$ for a series of samples from a normal universe and the standard deviation of that universe.
$!$ The factorial of a number.
A_1 The multiplier of $\bar{s}$ used to compute the distance from the central line to the three-sigma control limits of a chart for $\bar{X}$.
A_2 The multiplier of $\bar{R}$ used to compute the distance from the central line to the three-sigma control limits of a chart for $\bar{X}$.
R The difference between the largest and smallest values of a set of numbers; the range.
d_2 The ratio between the expected value of $\bar{R}$ for a series of samples from a normal universe and the standard deviation of that universe.
UCL_R Upper control limit of a chart for R.
LCL_R Lower control limit of a chart for R.
σ_R Standard error of the range.
$\bar{R}$ The average of a set of values of the range.

D_3 The multiplier of $\bar{R}$ used to compute the three-sigma lower control limit for a chart for R.

D_4 The multiplier of $\bar{R}$ used to compute the three-sigma upper control limit for a chart for R.

$\mathbf{UCL}_p$ Upper control limit for a p chart.

$\mathbf{LCL}_p$ Lower control limit for a p chart.

P (also p) The fraction defective (expressed as a decimal fraction).

Q $1 - P$

p' Estimate of the universe fraction defective.

n Number of items in a sample.

APPENDIX B

STATISTICAL FORMULAS

PART II — STATISTICAL DESCRIPTION

Averages

Formula Number			Page
1.	$\bar{X} = \frac{\Sigma X}{N}$	*arithmetic mean of the X variable*	164
2.	$\bar{X} = \frac{\Sigma fm}{N}$	*arithmetic mean of a frequency distribution.*	166
3.	$\bar{X} = A + \frac{\Sigma(fd)}{N}$	*arithmetic mean of a frequency distribution, using an arbitrary origin*	167
4.	$c = \frac{\Sigma(fd)}{N}$	*difference between mean ($\bar{X}$) and arbitrary origin (A)*	167
5.	$\bar{\bar{X}} = A + c$	*arithmetic mean of a frequency distribution, using an arbitrary origin*	167
6.	$c = \frac{\Sigma(fd')}{N} i$	*difference between mean ($\bar{X}$) and arbitrary origin (A)*	167
7.	$\bar{\bar{X}} = \frac{\Sigma(wX)}{\Sigma(w)}$	*weighted arithmetic mean, where w = the weight assigned to each item, or X value.*	169
8.	$G = (X_1 \cdot X_2 \cdot X_3 \ldots X_N)^{\frac{1}{N}}$	*geometric mean*	174
8a.	$\log G = \frac{\Sigma(\log X)}{N}$	*geometric mean, in terms of logarithms*	174
9.	$\text{Mode} = L + \frac{d_1}{d_1 + d_2} i$	$d_1 = f_m - f_1$; $d_2 = f_m - f_2$	18
10.	$\text{Mode} = \text{mean} - 3(\text{mean} - \text{median})$		1

Dispersion

Formula Number			Page
11.	$Q = \frac{Q_3 - Q_1}{2}$	*quartile deviation, or semi-interquartile range*	195
12.	$\sigma^2 = \frac{\Sigma x^2}{N}$	*square of standard deviation, or root-mean-square deviation from the mean*	198
13.	$s^2 = \frac{\Sigma d^2}{N}$	*mean-square deviation from an arbitrary origin*	199
14.	$\sigma^2 = s^2 - c^2$	*square of standard deviation using an arbitrary origin*	200
15.	$s^2 = \frac{\Sigma X^2}{N}$	*mean-square deviation from arbitrary origin zero*	200
16.	$c^2 = \left(\frac{\Sigma X}{N}\right)^2$	*square of difference between mean ($\bar{X}$) and arbitrary origin (A) when $A = 0$*	200
17.	$s^2 = \frac{\Sigma f(d')^2}{N}$ *in class interval units*	*mean-square deviation from an arbitrary origin*	203
18.	$V = \frac{\sigma}{\bar{X}} \cdot 100$	*coefficient of variation*	207
19.	$\text{Skewness} = \frac{\text{arithmetic mean} - \text{mode}}{\text{standard deviation}}$		208
19a.	$\text{Skewness} = \frac{3(\text{arithmetic mean} - \text{median})}{\text{standard deviation}}$		209

Normal Distribution

20.	$Y_o = \frac{N}{\sigma\sqrt{2\pi}}$ *or* $Y_o = \frac{N}{\sigma}.39894$	*maximum ordinate of the normal curve*	211
21.	$Y = Y_o e^{\frac{-x^2}{2\sigma^2}}$	*other ordinates of the normal curve*	211

PART III — STATISTICAL INDUCTION

22.	$\sigma_{\bar{x}} = \hat{\sigma}\sqrt{\frac{N-n}{n(N-1)}}$	*standard deviation of all possible sample means, or standard error of the mean*	223
23.	$\sigma_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$		

Formula Number		Page
24. $\sigma' = \sqrt{\sigma^2 \frac{n}{n-1}}$	.....	*estimate of the standard deviation of a universe from a sample standard deviation* ... 224
25. $\sigma' = \sigma\sqrt{\frac{n}{n-1}}$		
26. $\sigma' = \sqrt{\frac{\Sigma x^2}{n-1}}$	.....	*estimate of the standard deviation of a universe from a sample* 225
27. $\sigma_{\bar{x}} = \frac{\sigma'}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	.....	*standard error of the mean (finite universe).* 225
28. $\sigma_p = \sqrt{\frac{PQ}{n}\frac{N-n}{N-1}}$	.....	*standard error of a percentage (finite universe)* 231
29. $\sigma_p = \sqrt{\frac{PQ}{n}}$	.....	*standard error of a percentage* 232
30. $n = \frac{PQ}{\sigma_p^2}$	.....	*estimating the size of sample needed* 232
31. $n = \frac{\sigma'^2}{\sigma_{\bar{x}}^2}$	.....	*estimating the size of sample needed* 233

PART IV — ANALYSIS OF BUSINESS CHANGE

Secular Trend — Straight Line

32. $Y_c = a + bX$	.....	*straight-line trend* 279
33. $a = \frac{\Sigma Y}{N}$	.....	*y - intercept of straight line (origin at midrange)* 287
34. $b = \frac{\Sigma XY}{\Sigma X^2}$	.....	*slope of straight line* .. 288

Secular Trend — Nonlinear

35. $a = \frac{\Sigma(\log Y)}{N}$	.....	*y - intercept of logarithmic straight line (origin at midrange)* .. 296
36. $b = \frac{\Sigma(X \cdot \log Y)}{\Sigma X^2}$	.....	*slope of logarithmic straight line* 296
37. $Y_c = a + bX + cX^2$	.....	*second-degree parabola* 299

Formula Number			Page
38. $c = \frac{N \cdot \Sigma X^2 Y - \Sigma X^2 \Sigma Y}{N \cdot \Sigma X^4 - (\Sigma X^2)^2}$		constants of second-degree parabola (origin at midrange)	299
39. $a = \frac{\Sigma Y - c\Sigma X^2}{N}$			
40. $\log Y_c = \log a + (\log b)c^t$		Gompertz curve	303

Index Numbers

41. Quantity Index $= \frac{\Sigma\left(\frac{Q_1}{Q_0}100\right)}{N}$		unweighted arithmetic mean	390
42. Price Index $= \frac{\Sigma\left(\frac{P_1}{P_0}100\right)}{N}$		unweighted arithmetic mean	391
43. Log of Quantity Index $= \frac{\Sigma \log\left(\frac{Q_1}{Q_0}100\right)}{N}$		unweighted geometric mean	391
44. Log of Price Index $= \frac{\Sigma \log\left(\frac{P_1}{P_0}100\right)}{N}$		unweighted geometric mean	392
45. Quantity Index $= \frac{\Sigma\left(P_0Q_0\frac{Q_1}{Q_0}100\right)}{\Sigma(P_0Q_0)}$		arithmetic mean weighted with base-year values	396
46. Price Index $= \frac{\Sigma\left(P_0Q_0\frac{P_1}{P_0}100\right)}{\Sigma(P_0Q_0)}$			
47. Log of Quantity Index $= \frac{\left[(wt.)\ (\log\frac{Q_1}{Q_0}100)\right]}{\Sigma wt.}$		weighted geometric mean	398
48. Log of Price Index $= \frac{\left[(wt.)\ (\log\frac{P_1}{P_0}100)\right]}{\Sigma wt.}$		weighted geometric mean	399
49. Quantity Index $= \frac{\Sigma(P_0Q_1)}{\Sigma(P_0Q_0)}100$		method of aggregates	403
50. Price Index $= \frac{\Sigma(P_1Q_0)}{\Sigma(P_0Q_0)}\ 100$		method of aggregates	404

PART V — ANALYSIS OF RELATIONSHIP

Linear Correlation

Formula Number		Page
51. $a = \dfrac{\Sigma X^2 \Sigma Y - \Sigma X \cdot \Sigma XY}{N \cdot \Sigma X^2 - (\Sigma X)^2}$ 52. $b = \dfrac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{N \cdot \Sigma X^2 - (\Sigma X)^2}$	*constants of linear regression equation*	431
53. $S_y = \sqrt{\dfrac{\Sigma (Y - Y_c)^2}{N}}$	*standard error of estimate*	435
54. $S_y = \sqrt{\dfrac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}$	*standard error of estimate*	436
55. $k^2 = \dfrac{S_y^2}{\sigma_y^2}$	*coefficient of nondetermination*	439
56. $r^2 = 1 - \dfrac{S_y^2}{\sigma_y^2}$	*coefficient of determination*	439
57. $r = \dfrac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{[N \cdot \Sigma X^2 - (\Sigma X)^2]\,[N \cdot \Sigma Y^2 - (\Sigma Y)^2]}}$	*coefficient of correlation*	440
58. $a = \bar{Y} - b\bar{X}$	*constant of the regression equation*	441
59. $S_y = \sigma_y k$ 60. $S_y = \sigma_y \sqrt{1 - r^2}$	*standard error of estimate*	441
61. $\sigma_x = \dfrac{\sqrt{N\Sigma X^2 - (\Sigma X)^2}}{N}$	*standard deviation of x variable*	442
62. $r = \dfrac{N \cdot \Sigma f d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{\sqrt{[N \cdot \Sigma f_x (d'_x)^2 - (\Sigma f_x d'_x)^2]\,[N \cdot \Sigma f_y (d'_y)^2 - (\Sigma f_y d'_y)^2]}}$	*coefficient of correlation (grouped data)*	444
63. $b = \dfrac{N \cdot \Sigma f d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{N \cdot \Sigma f_x (d'_x)^2 - (\Sigma f_x d'_x)^2} \cdot \dfrac{i_y}{i_x}$	*constant of the regression equation (grouped data)*	445
64. $\sigma_y = i_y \dfrac{\sqrt{N \cdot \Sigma f_y (d'_y)^2 - (\Sigma f_y d'_y)^2}}{N}$	*standard deviation of the Y variable*	446

Nonlinear Correlation

Formula Number			Page
65.	$\rho^2_{yx} = 1 - \frac{S_y^2}{\sigma_y^2}$	*index of correlation*	456
66.	$\sigma^2_m = \frac{\Sigma f(\overline{Y}_k - \overline{Y})^2}{N}$	*explained variance*	459
67.	$\eta^2 = \frac{\sigma^2_m}{\sigma^2_y}$	*correlation ratio*	459

Multiple Correlation

68.	$S_{1.23} = \sqrt{\frac{\Sigma(X_1 - X_{1c})^2}{N}}$	*standard error of estimate*	466
69.	$R^2_{1.23} = 1 - \frac{S^2_{1.23}}{\sigma_1^2}$	*coefficient of determination*	467

Partial Correlation

70.	$r^2_{13.2} = 1 - \frac{S^2_{1.23}}{S^2_{12}}$	*coefficient of partial determination*	468

PART VI — APPLICATIONS IN BUSINESS

Statistical Quality Control

71.	$c_2 = \sqrt{\frac{2}{n}} \frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-3}{2}\right)!}$	*computing σ' from σ*	545
72.	$A_1 = \frac{3}{c_2\sqrt{n}}$	*computing control limits from $\bar{\sigma}$*	546
73.	$A_2 = \frac{3}{d_2\sqrt{n}}$	*computing control limits from $\overline{R}$*	548

APPENDIX C

TABLE OF SQUARES, SQUARE ROOTS, AND RECIPROCALS

The following table gives the values of the squares, square roots, and reciprocals for all sequences of three digits. The values of the squares of the numbers are exact, but the square roots and reciprocals are rounded to six digits.

Finding the square of a number consists merely of finding the value of n in the first column and reading the value of n^2 in the second column. The square of 1.30 is 1.6900, and the square of 13.0 is 169.00. The reciprocal of the number is given in the fifth column. The reciprocal of 1.30 is .769231, and the reciprocal of 13.0 is .0769231. The position of the decimal in the number determines the location of the decimal in the square or reciprocal, but the digits in the square or reciprocal are the same for a given sequence of numbers in n, regardless of the location of the decimal.

The square root of a given sequence, however, is affected by the position of the decimal point. The square root of 1.44 is 1.2, but the square root of 14.4 is 3.79473. Thus, two columns are required for square roots: one for the value of the numbers from 1 to 10, and another for the numbers from 10 to 100, even though the sequence of digits is the same. The square roots of the numbers from 1 to 10 are given in the third column (headed $\sqrt{n}$) and the square roots of the numbers from 10 to 100 are given in the fourth column (headed $\sqrt{10n}$). The column headed $\sqrt{n}$ is used for numbers with an odd number of digits to the left of the decimal, and the column headed $\sqrt{10n}$ is used for numbers with an even number of digits to the left of the decimal.

SQUARES — SQUARE ROOTS — RECIPROCALS

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
1.00	1.0000	1.00000	3.16228	1.000000
1.01	1.0201	1.00499	3.17805	.990099
1.02	1.0404	1.00995	3.19374	.980392
1.03	1.0609	1.01489	3.20936	.970874
1.04	1.0816	1.01980	3.22490	.961538
1.05	1.1025	1.02470	3.24037	.952381
1.06	1.1236	1.02956	3.25576	.943396
1.07	1.1449	1.03441	3.27109	.934579
1.08	1.1664	1.03923	3.28634	.925926
1.09	1.1881	1.04403	3.30151	.917431
1.10	1.2100	1.04881	3.31662	.909091
1.11	1.2321	1.05357	3.33167	.900901
1.12	1.2544	1.05830	3.34664	.892857
1.13	1.2769	1.06301	3.36155	.884956
1.14	1.2996	1.06771	3.37639	.877193
1.15	1.3225	1.07238	3.39116	.869565
1.16	1.3456	1.07703	3.40588	.862069
1.17	1.3689	1.08167	3.42053	.854701
1.18	1.3924	1.08628	3.43511	.847458
1.19	1.4161	1.09087	3.44964	.840336
1.20	1.4400	1.09545	3.46410	.833333
1.21	1.4641	1.10000	3.47851	.826446
1.22	1.4884	1.10454	3.49285	.819672
1.23	1.5129	1.10905	3.50714	.813008
1.24	1.5376	1.11355	3.52136	.806452
1.25	1.5625	1.11803	3.53553	.800000
1.26	1.5876	1.12250	3.54965	.793651
1.27	1.6129	1.12694	3.56371	.787402
1.28	1.6384	1.13137	3.57771	.781250
1.29	1.6641	1.13578	3.59166	.775194
1.30	1.6900	1.14018	3.60555	.769231
1.31	1.7161	1.14455	3.61939	.763359
1.32	1.7424	1.14891	3.63318	.757576
1.33	1.7689	1.15326	3.64692	.751880
1.34	1.7956	1.15758	3.66060	.746269
1.35	1.8225	1.16190	3.67423	.740741
1.36	1.8496	1.16619	3.68782	.735294
1.37	1.8769	1.17047	3.70135	.729927
1.38	1.9044	1.17473	3.71484	.724638
1.39	1.9321	1.17898	3.72827	.719424
1.40	1.9600	1.18322	3.74166	.714286
1.41	1.9881	1.18743	3.75500	.709220
1.42	2.0164	1.19164	3.76829	.704225
1.43	2.0449	1.19583	3.78153	.699301
1.44	2.0736	1.20000	3.79473	.694444
1.45	2.1025	1.20416	3.80789	.689655
1.46	2.1316	1.20830	3.82099	.684932
1.47	2.1609	1.21244	3.83406	.680272
1.48	2.1904	1.21655	3.84708	.675676
1.49	2.2201	1.22066	3.86005	.671141
1.50	2.2500	1.22474	3.87298	.666667

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
1.50	2.2500	1.22474	3.87298	.666667
1.51	2.2801	1.22882	3.88587	.662252
1.52	2.3104	1.23288	3.89872	.657895
1.53	2.3409	1.23693	3.91152	.653595
1.54	2.3716	1.24097	3.92428	.649351
1.55	2.4025	1.24499	3.93700	.645161
1.56	2.4336	1.24900	3.94968	.641026
1.57	2.4649	1.25300	3.96232	.636943
1.58	2.4964	1.25698	3.97492	.632911
1.59	2.5281	1.26095	3.98748	.628931
1.60	2.5600	1.26491	4.00000	.625000
1.61	2.5921	1.26886	4.01248	.621118
1.62	2.6244	1.27279	4.02492	.617284
1.63	2.6569	1.27671	4.03733	.613497
1.64	2.6896	1.28062	4.04969	.609756
1.65	2.7225	1.28452	4.06202	.606061
1.66	2.7556	1.28841	4.07431	.602410
1.67	2.7889	1.29228	4.08656	.598802
1.68	2.8224	1.29615	4.09878	.595238
1.69	2.8561	1.30000	4.11096	.591716
1.70	2.8900	1.30384	4.12311	.588235
1.71	2.9241	1.30767	4.13521	.584795
1.72	2.9584	1.31149	4.14729	.581395
1.73	2.9929	1.31529	4.15933	.578035
1.74	3.0276	1.31909	4.17133	.574713
1.75	3.0625	1.32288	4.18330	.571429
1.76	3.0976	1.32665	4.19524	.568182
1.77	3.1329	1.33041	4.20714	.564972
1.78	3.1684	1.33417	4.21900	.561798
1.79	3.2041	1.33791	4.23084	.558659
1.80	3.2400	1.34164	4.24264	.555556
1.81	3.2761	1.34536	4.25441	.552486
1.82	3.3124	1.34907	4.26615	.549451
1.83	3.3489	1.35277	4.27785	.546448
1.84	3.3856	1.35647	4.28952	.543478
1.85	3.4225	1.36015	4.30116	.540541
1.86	3.4596	1.36382	4.31277	.537634
1.87	3.4969	1.36748	4.32435	.534759
1.88	3.5344	1.37113	4.33590	.531915
1.89	3.5721	1.37477	4.34741	.529101
1.90	3.6100	1.37840	4.35890	.526316
1.91	3.6481	1.38203	4.37035	.523560
1.92	3.6864	1.38564	4.38178	.520833
1.93	3.7249	1.38924	4.39318	.518135
1.94	3.7636	1.39284	4.40454	.515464
1.95	3.8025	1.39642	4.41588	.512821
1.96	3.8416	1.40000	4.42719	.510204
1.97	3.8809	1.40357	4.43847	.507614
1.98	3.9204	1.40712	4.44972	.505051
1.99	3.9601	1.41067	4.46094	.502513
2.00	4.0000	1.41421	4.47214	.500000

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Continued*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
2.00	4.0000	1.41421	4.47214	.500000
2.01	4.0401	1.41774	4.48330	.497512
2.02	4.0804	1.42127	4.49444	.495050
2.03	4.1209	1.42478	4.50555	.492611
2.04	4.1616	1.42829	4.51664	.490196
2.05	4.2025	1.43178	4.52769	.487805
2.06	4.2436	1.43527	4.53872	.485437
2.07	4.2849	1.43875	4.54973	.483092
2.08	4.3264	1.44222	4.56070	.480769
2.09	4.3681	1.44568	4.57165	.478469
2.10	4.4100	1.44914	4.58258	.476190
2.11	4.4521	1.45258	4.59347	.473934
2.12	4.4944	1.45602	4.60435	.471698
2.13	4.5369	1.45945	4.61519	.469484
2.14	4.5796	1.46287	4.62601	.467290
2.15	4.6225	1.46629	4.63681	.465116
2.16	4.6656	1.46969	4.64758	.462963
2.17	4.7089	1.47309	4.65833	.460829
2.18	4.7524	1.47648	4.66905	.458716
2.19	4.7961	1.47986	4.67974	.456621
2.20	4.8400	1.48324	4.69042	.454545
2.21	4.8841	1.48661	4.70106	.452489
2.22	4.9284	1.48997	4.71169	.450450
2.23	4.9729	1.49332	4.72229	.448430
2.24	5.0176	1.49666	4.73286	.446429
2.25	5.0625	1.50000	4.74342	.444444
2.26	5.1076	1.50333	4.75395	.442478
2.27	5.1529	1.50665	4.76445	.440529
2.28	5.1984	1.50997	4.77493	.438596
2.29	5.2441	1.51327	4.78539	.436681
2.30	5.2900	1.51658	4.79583	.434783
2.31	5.3361	1.51987	4.80625	.432900
2.32	5.3824	1.52315	4.81664	.431034
2.33	5.4289	1.52643	4.82701	.429185
2.34	5.4756	1.52971	4.83735	.427350
2.35	5.5225	1.53297	4.84768	.425532
2.36	5.5696	1.53623	4.85798	.423729
2.37	5.6169	1.53948	4.86826	.421941
2.38	5.6644	1.54272	4.87852	.420168
2.39	5.7121	1.54596	4.88876	.418410
2.40	5.7600	1.54919	4.89898	.416667
2.41	5.8081	1.55242	4.90918	.414938
2.42	5.8564	1.55563	4.91935	.413223
2.43	5.9049	1.55885	4.92950	.411523
2.44	5.9536	1.56205	4.93964	.409836
2.45	6.0025	1.56525	4.94975	.408163
2.46	6.0516	1.56844	4.95984	.406504
2.47	6.1009	1.57162	4.96991	.404858
2.48	6.1504	1.57480	4.97996	.403226
2.49	6.2001	1.57797	4.98999	.401606
2.50	6.2500	1.58114	5.00000	.400000
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
2.50	6.2500	1.58114	5.00000	.400000
2.51	6.3001	1.58430	5.00999	.398406
2.52	6.3504	1.58745	5.01996	.396825
2.53	6.4009	1.59060	5.02991	.395257
2.54	6.4516	1.59374	5.03984	.393701
2.55	6.5025	1.59687	5.04975	.392157
2.56	6.5536	1.60000	5.05964	.390625
2.57	6.6049	1.60312	5.06952	.389105
2.58	6.6564	1.60624	5.07937	.387597
2.59	6.7081	1.60935	5.08920	.386100
2.60	6.7600	1.61245	5.09902	.384615
2.61	6.8121	1.61555	5.10882	.383142
2.62	6.8644	1.61864	5.11859	.381679
2.63	6.9169	1.62173	5.12835	.380228
2.64	6.9696	1.62481	5.13809	.378788
2.65	7.0225	1.62788	5.14782	.377358
2.66	7.0756	1.63095	5.15752	.375940
2.67	7.1289	1.63401	5.16720	.374532
2.68	7.1824	1.63707	5.17687	.373134
2.69	7.2361	1.64012	5.18652	.371747
2.70	7.2900	1.64317	5.19615	.370370
2.71	7.3441	1.64621	5.20577	.369004
2.72	7.3984	1.64924	5.21536	.367647
2.73	7.4529	1.65227	5.22494	.366300
2.74	7.5076	1.65529	5.23450	.364964
2.75	7.5625	1.65831	5.24404	.363636
2.76	7.6176	1.66132	5.25357	.362319
2.77	7.6729	1.66433	5.26308	.361011
2.78	7.7284	1.66733	5.27257	.359712
2.79	7.7841	1.67033	5.28205	.358423
2.80	7.8400	1.67332	5.29150	.357143
2.81	7.8961	1.67631	5.30094	.355872
2.82	7.9524	1.67929	5.31037	.354610
2.83	8.0089	1.68226	5.31977	.353357
2.84	8.0656	1.68523	5.32917	.352113
2.85	8.1225	1.68819	5.33854	.350877
2.86	8.1796	1.69115	5.34790	.349650
2.87	8.2369	1.69411	5.35724	.348432
2.88	8.2944	1.69706	5.36656	.347222
2.89	8.3521	1.70000	5.37587	.346021
2.90	8.4100	1.70294	5.38516	.344828
2.91	8.4681	1.70587	5.39444	.343643
2.92	8.5264	1.70880	5.40370	.342466
2.93	8.5849	1.71172	5.41295	.341297
2.94	8.6436	1.71464	5.42218	.340136
2.95	8.7025	1.71756	5.43139	.338983
2.96	8.7616	1.72047	5.44059	.337838
2.97	8.8209	1.72337	5.44977	.336700
2.98	8.8804	1.72627	5.45894	.335570
2.99	8.9401	1.72916	5.46809	.334448
3.00	9.0000	1.73205	5.47723	.333333
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Continued*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
3.00	9.0000	1.73205	5.47723	.333333
3.01	9.0601	1.73494	5.48635	.332226
3.02	9.1204	1.73781	5.49545	.331126
3.03	9.1809	1.74069	5.50454	.330033
3.04	9.2416	1.74356	5.51362	.328947
3.05	9.3025	1.74642	5.52268	.327869
3.06	9.3636	1.74929	5.53173	.326797
3.07	9.4249	1.75214	5.54076	.325733
3.08	9.4864	1.75499	5.54977	.324675
3.09	9.5481	1.75784	5.55878	.323625
3.10	9.6100	1.76068	5.56776	.322581
3.11	9.6721	1.76352	5.57674	.321543
3.12	9.7344	1.76635	5.58570	.320513
3.13	9.7969	1.76918	5.59464	.319489
3.14	9.8596	1.77200	5.60357	.318471
3.15	9.9225	1.77482	5.61249	.317460
3.16	9.9856	1.77764	5.62139	.316456
3.17	10.0489	1.78045	5.63028	.315457
3.18	10.1124	1.78326	5.63915	.314465
3.19	10.1761	1.78606	5.64801	.313480
3.20	10.2400	1.78885	5.65685	.312500
3.21	10.3041	1.79165	5.66569	.311526
3.22	10.3684	1.79444	5.67450	.310559
3.23	10.4329	1.79722	5.68331	.309598
3.24	10.4976	1.80000	5.69210	.308642
3.25	10.5625	1.80278	5.70088	.307692
3.26	10.6276	1.80555	5.70964	.306748
3.27	10.6929	1.80831	5.71839	.305810
3.28	10.7584	1.81108	5.72713	.304878
3.29	10.8241	1.81384	5.73585	.303951
3.30	10.8900	1.81659	5.74456	.303030
3.31	10.9561	1.81934	5.75326	.302115
3.32	11.0224	1.82209	5.76194	.301205
3.33	11.0889	1.82483	5.77062	.300300
3.34	11.1556	1.82757	5.77927	.299401
3.35	11.2225	1.83030	5.78792	.298507
3.36	11.2896	1.83303	5.79655	.297619
3.37	11.3569	1.83576	5.80517	.296736
3.38	11.4244	1.83848	5.81378	.295858
3.39	11.4921	1.84120	5.82237	.294985
3.40	11.5600	1.84391	5.83095	.294118
3.41	11.6281	1.84662	5.83952	.293255
3.42	11.6964	1.84932	5.84808	.292398
3.43	11.7649	1.85203	5.85662	.291545
3.44	11.8336	1.85472	5.86515	.290698
3.45	11.9025	1.85742	5.87367	.289855
3.46	11.9716	1.86011	5.88218	.289017
3.47	12.0409	1.86279	5.89067	.288184
3.48	12.1104	1.86548	5.89915	.287356
3.49	12.1801	1.86815	5.90762	.286533
3.50	12.2500	1.87083	5.91608	.285714
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
3.50	12.2500	1.87083	5.91608	.285714
3.51	12.3201	1.87350	5.92453	.284900
3.52	12.3904	1.87617	5.93296	.284091
3.53	12.4609	1.87883	5.94138	.283286
3.54	12.5316	1.88149	5.94979	.282486
3.55	12.6025	1.88414	5.95819	.281690
3.56	12.6736	1.88680	5.96657	.280899
3.57	12.7449	1.88944	5.97495	.280112
3.58	12.8164	1.89209	5.98331	.279330
3.59	12.8881	1.89473	5.99166	.278552
3.60	12.9600	1.89737	6.00000	.277778
3.61	13.0321	1.90000	6.00833	.277008
3.62	13.1044	1.90263	6.01664	.276243
3.63	13.1769	1.90526	6.02495	.275482
3.64	13.2496	1.90788	6.03324	.274725
3.65	13.3225	1.91050	6.04152	.273973
3.66	13.3956	1.91311	6.04979	.273224
3.67	13.4689	1.91572	6.05805	.272480
3.68	13.5424	1.91833	6.06630	.271739
3.69	13.6161	1.92094	6.07454	.271003
3.70	13.6900	1.92354	6.08276	.270270
3.71	13.7641	1.92614	6.09098	.269542
3.72	13.8384	1.92873	6.09918	.268817
3.73	13.9129	1.93132	6.10737	.268097
3.74	13.9876	1.93391	6.11555	.267380
3.75	14.0625	1.93649	6.12372	.266667
3.76	14.1376	1.93907	6.13188	.265957
3.77	14.2129	1.94165	6.14003	.265252
3.78	14.2884	1.94422	6.14817	.264550
3.79	14.3641	1.94679	6.15630	.263852
3.80	14.4400	1.94936	6.16441	.263158
3.81	14.5161	1.95192	6.17252	.262467
3.82	14.5924	1.95448	6.18061	.261780
3.83	14.6689	1.95704	6.18870	.261097
3.84	14.7456	1.95959	6.19677	.260417
3.85	14.8225	1.96214	6.20484	.259740
3.86	14.8996	1.96469	6.21289	.259067
3.87	14.9769	1.96723	6.22093	.258398
3.88	15.0544	1.96977	6.22896	.257732
3.89	15.1321	1.97231	6.23699	.257069
3.90	15.2100	1.97484	6.24500	.256410
3.91	15.2881	1.97737	6.25300	.255754
3.92	15.3664	1.97990	6.26099	.255102
3.93	15.4449	1.98242	6.26897	.254453
3.94	15.5236	1.98494	6.27694	.253807
3.95	15.6025	1.98746	6.28490	.253165
3.96	15.6816	1.98997	6.29285	.252525
3.97	15.7609	1.99249	6.30079	.251889
3.98	15.8408	1.99499	6.30872	.251256
3.99	15.9201	1.99750	6.31664	.250627
4.00	16.0000	2.00000	6.32456	.250000
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS *(Continued)*

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
4.00	16.0000	2.00000	6.32456	.250000
4.01	16.0801	2.00250	6.33246	.249377
4.02	16.1604	2.00499	6.34035	.248756
4.03	16.2409	2.00749	6.34823	.248139
4.04	16.3216	2.00998	6.35610	.247525
4.05	16.4025	2.01246	6.36396	.246914
4.06	16.4836	2.01494	6.37181	.246305
4.07	16.5649	2.01742	6.37966	.245700
4.08	16.6464	2.01990	6.38749	.245098
4.09	16.7281	2.02237	6.39531	.244499
4.10	16.8100	2.02485	6.40312	.243902
4.11	16.8921	2.02731	6.41093	.243309
4.12	16.9744	2.02978	6.41872	.242718
4.13	17.0569	2.03224	6.42651	.242131
4.14	17.1396	2.03470	6.43428	.241546
4.15	17.2225	2.03715	6.44205	.240964
4.16	17.3056	2.03961	6.44981	.240385
4.17	17.3889	2.04206	6.45755	.239808
4.18	17.4724	2.04450	6.46529	.239234
4.19	17.5561	2.04695	6.47302	.238663
4.20	17.6400	2.04939	6.48074	.238095
4.21	17.7241	2.05183	6.48845	.237530
4.22	17.8084	2.05426	6.49615	.236967
4.23	17.8929	2.05670	6.50384	.236407
4.24	17.9776	2.05913	6.51153	.235849
4.25	18.0625	2.06155	6.51920	.235294
4.26	18.1476	2.06398	6.52687	.234742
4.27	18.2329	2.06640	6.53452	.234192
4.28	18.3184	2.06882	6.54217	.233645
4.29	18.4041	2.07123	6.54981	.233100
4.30	18.4900	2.07364	6.55744	.232558
4.31	18.5761	2.07605	6.56506	.232019
4.32	18.6624	2.07846	6.57267	.231481
4.33	18.7489	2.08087	6.58027	.230947
4.34	18.8356	2.08327	6.58787	.230415
4.35	18.9225	2.08567	6.59545	.229885
4.36	19.0096	2.08806	6.60303	.229358
4.37	19.0969	2.09045	6.61060	.228833
4.38	19.1844	2.09284	6.61816	.228311
4.39	19.2721	2.09523	6.62571	.227790
4.40	19.3600	2.09762	6.63325	.227273
4.41	19.4481	2.10000	6.64078	.226757
4.42	19.5364	2.10238	6.64831	.226244
4.43	19.6249	2.10476	6.65582	.225734
4.44	19.7136	2.10713	6.66333	.225225
4.45	19.8025	2.10950	6.67083	.224719
4.46	19.8916	2.11187	6.67832	.224215
4.47	19.9809	2.11424	6.68581	.223714
4.48	20.0704	2.11660	6.69328	.223214
4.49	20.1601	2.11896	6.70075	.222717
4.50	20.2500	2.12132	6.70820	.222222
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
4.50	20.2500	2.12132	6.70820	.222222
4.51	20.3401	2.12368	6.71565	.221729
4.52	20.4304	2.12603	6.72309	.221239
4.53	20.5209	2.12838	6.73053	.220751
4.54	20.6116	2.13073	6.73795	.220264
4.55	20.7025	2.13307	6.74537	.219780
4.56	20.7936	2.13542	6.75278	.219298
4.57	20.8849	2.13776	6.76018	.218818
4.58	20.9764	2.14009	6.76757	.218341
4.59	21.0681	2.14243	6.77495	.217865
4.60	21.1600	2.14476	6.78233	.217391
4.61	21.2521	2.14709	6.78970	.216920
4.62	21.3444	2.14942	6.79706	.216450
4.63	21.4369	2.15174	6.80441	.215983
4.64	21.5296	2.15407	6.81175	.215517
4.65	21.6225	2.15639	6.81909	.215054
4.66	21.7156	2.15870	6.82642	.214592
4.67	21.8089	2.16102	6.83374	.214133
4.68	21.9024	2.16333	6.84105	.213675
4.69	21.9961	2.16564	6.84836	.213220
4.70	22.0900	2.16795	6.85565	.212766
4.71	22.1841	2.17025	6.86294	.212314
4.72	22.2784	2.17256	6.87023	.211864
4.73	22.3729	2.17486	6.87750	.211416
4.74	22.4676	2.17715	6.88477	.210970
4.75	22.5625	2.17945	6.89202	.210526
4.76	22.6576	2.18174	6.89928	.210084
4.77	22.7529	2.18403	6.90652	.209644
4.78	22.8484	2.18632	6.91375	.209205
4.79	22.9441	2.18861	6.92098	.208768
4.80	23.0400	2.19089	6.92820	.208333
4.81	23.1361	2.19317	6.93542	.207900
4.82	23.2324	2.19545	6.94262	.207469
4.83	23.3289	2.19773	6.94982	.207039
4.84	23.4256	2.20000	6.95701	.206612
4.85	23.5225	2.20227	6.96419	.206186
4.86	23.6196	2.20454	6.97137	.205761
4.87	23.7169	2.20681	6.97854	.205339
4.88	23.8144	2.20907	6.98570	.204918
4.89	23.9121	2.21133	6.99285	.204499
4.90	24.0100	2.21359	7.00000	.204082
4.91	24.1081	2.21585	7.00714	.203666
4.92	24.2064	2.21811	7.01427	.203252
4.93	24.3049	2.22036	7.02140	.202840
4.94	24.4036	2.22261	7.02851	.202429
4.95	24.5025	2.22486	7.03562	.202020
4.96	24.6016	2.22711	7.04273	.201613
4.97	24.7009	2.22935	7.04982	.201207
4.98	24.8004	2.23159	7.05691	.200803
4.99	24.9001	2.23383	7.06399	.200401
5.00	25.0000	2.23607	7.07107	.200000
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Continued*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
5.00	25.0000	2.23607	7.07107	.200000
5.01	25.1001	2.23830	7.07814	.199601
5.02	25.2004	2.24054	7.08520	.199203
5.03	25.3009	2.24277	7.09225	.198807
5.04	25.4016	2.24499	7.09930	.198413
5.05	25.5025	2.24722	7.10634	.198020
5.06	25.6036	2.24944	7.11337	.197628
5.07	25.7049	2.25167	7.12039	.197239
5.08	25.8064	2.25389	7.12741	.196850
5.09	25.9081	2.25610	7.13442	.196464
5.10	26.0100	2.25832	7.14143	.196078
5.11	26.1121	2.26053	7.14843	.195695
5.12	26.2144	2.26274	7.15542	.195312
5.13	26.3169	2.26495	7.16240	.194932
5.14	26.4196	2.26716	7.16938	.194553
5.15	26.5225	2.26936	7.17635	.194175
5.16	26.6256	2.27156	7.18331	.193798
5.17	26.7289	2.27376	7.19027	.193424
5.18	26.8324	2.27596	7.19722	.193050
5.19	26.9361	2.27816	7.20417	.192678
5.20	27.0400	2.28035	7.21110	.192308
5.21	27.1441	2.28254	7.21803	.191939
5.22	27.2484	2.28473	7.22496	.191571
5.23	27.3529	2.28692	7.23187	.191205
5.24	27.4576	2.28910	7.23878	.190840
5.25	27.5625	2.29129	7.24569	.190476
5.26	27.6676	2.29347	7.25259	.190114
5.27	27.7729	2.29565	7.25948	.189753
5.28	27.8784	2.29783	7.26636	.189394
5.29	27.9841	2.30000	7.27324	.189036
5.30	28.0900	2.30217	7.28011	.188679
5.31	28.1961	2.30434	7.28697	.188324
5.32	28.3024	2.30651	7.29383	.187970
5.33	28.4089	2.30868	7.30068	.187617
5.34	28.5156	2.31084	7.30753	.187266
5.35	28.6225	2.31301	7.31437	.186916
5.36	28.7296	2.31517	7.32120	.186567
5.37	28.8369	2.31733	7.32803	.186220
5.38	28.9444	2.31948	7.33485	.185874
5.39	29.0521	2.32164	7.34166	.185529
5.40	29.1600	2.32379	7.34847	.185185
5.41	29.2681	2.32594	7.35527	.184843
5.42	29.3764	2.32809	7.36206	.184502
5.43	29.4849	2.33024	7.36885	.184162
5.44	29.5936	2.33238	7.37564	.183824
5.45	29.7025	2.33452	7.38241	.183486
5.46	29.8116	2.33666	7.38918	.183150
5.47	29.9209	2.33880	7.39594	.182815
5.48	30.0304	2.34094	7.40270	.182482
5.49	30.1401	2.34307	7.40945	.182149
5.50	30.2500	2.34521	7.41620	.181818
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
5.50	30.2500	2.34521	7.41620	.181818
5.51	30.3601	2.34734	7.42294	.181488
5.52	30.4704	2.34947	7.42967	.181159
5.53	30.5809	2.35160	7.43640	.180832
5.54	30.6916	2.35372	7.44312	.180505
5.55	30.8025	2.35584	7.44983	.180180
5.56	30.9136	2.35797	7.45654	.179856
5.57	31.0249	2.36008	7.46324	.179533
5.58	31.1364	2.36220	7.46994	.179211
5.59	31.2481	2.36432	7.47663	.178891
5.60	31.3600	2.36643	7.48331	.178571
5.61	31.4721	2.36854	7.48999	.178253
5.62	31.5844	2.37065	7.49667	.177936
5.63	31.6969	2.37276	7.50333	.177620
5.64	31.8096	2.37487	7.50999	.177305
5.65	31.9225	2.37697	7.51665	.176991
5.66	32.0356	2.37908	7.52330	.176678
5.67	32.1489	2.38118	7.52994	.176367
5.68	32.2624	2.38328	7.53658	.176056
5.69	32.3761	2.38537	7.54321	.175747
5.70	32.4900	2.38747	7.54983	.175439
5.71	32.6041	2.38956	7.55645	.175131
5.72	32.7184	2.39165	7.56307	.174825
5.73	32.8329	2.39374	7.56968	.174520
5.74	32.9476	2.39583	7.57628	.174216
5.75	33.0625	2.39792	7.58288	.173913
5.76	33.1776	2.40000	7.58947	.173611
5.77	33.2929	2.40208	7.59605	.173310
5.78	33.4084	2.40416	7.60263	.173010
5.79	33.5241	2.40624	7.60920	.172712
5.80	33.6400	2.40832	7.61577	.172414
5.81	33.7561	2.41039	7.62234	.172117
5.82	33.8724	2.41247	7.62889	.171821
5.83	33.9889	2.41454	7.63544	.171527
5.84	34.1056	2.41661	7.64199	.171233
5.85	34.2225	2.41868	7.64853	.170940
5.86	34.3396	2.42074	7.65506	.170649
5.87	34.4569	2.42281	7.66159	.170358
5.88	34.5744	2.42487	7.66812	.170068
5.89	34.6921	2.42693	7.67463	.169779
5.90	34.8100	2.42899	7.68115	.169492
5.91	34.9281	2.43105	7.68765	.169205
5.92	35.0464	2.43311	7.69415	.168919
5.93	35.1649	2.43516	7.70065	.168634
5.94	35.2836	2.43721	7.70714	.168350
5.95	35.4025	2.43926	7.71362	.168067
5.96	35.5216	2.44131	7.72010	.167785
5.97	35.6409	2.44336	7.72658	.167504
5.98	35.7604	2.44540	7.73305	.167224
5.99	35.8801	2.44745	7.73951	.166945
6.00	36.0000	2.44949	7.74597	.166667
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Continued*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
6.00	36.0000	2.44949	7.74597	.166667
6.01	36.1201	2.45153	7.75242	.166389
6.02	36.2404	2.45357	7.75887	.166113
6.03	36.3609	2.45561	7.76531	.165837
6.04	36.4816	2.45764	7.77174	.165563
6.05	36.6025	2.45967	7.77817	.165289
6.06	36.7236	2.46171	7.78460	.165017
6.07	36.8449	2.46374	7.79102	.164745
6.08	36.9664	2.46577	7.79744	.164474
6.09	37.0881	2.46779	7.80385	.164204
6.10	37.2100	2.46982	7.81025	.163934
6.11	37.3321	2.47184	7.81665	.163666
6.12	37.4544	2.47386	7.82304	.163399
6.13	37.5769	2.47588	7.82943	.163132
6.14	37.6996	2.47790	7.83582	.162866
6.15	37.8225	2.47992	7.84219	.162602
6.16	37.9456	2.48193	7.84857	.162338
6.17	38.0689	2.48395	7.85493	.162075
6.18	38.1924	2.48596	7.86130	.161812
6.19	38.3161	2.48797	7.86766	.161551
6.20	38.4400	2.48998	7.87401	.161290
6.21	38.5641	2.49199	7.88036	.161031
6.22	38.6884	2.49399	7.88670	.160772
6.23	38.8129	2.49600	7.89303	.160514
6.24	38.9376	2.49800	7.89937	.160256
6.25	39.0625	2.50000	7.90569	.160000
6.26	39.1876	2.50200	7.91202	.159744
6.27	39.3129	2.50400	7.91833	.159490
6.28	39.4384	2.50599	7.92465	.159236
6.29	39.5641	2.50799	7.93095	.158983
6.30	39.6900	2.50998	7.93725	.158730
6.31	39.8161	2.51197	7.94355	.158479
6.32	39.9424	2.51396	7.94984	.158228
6.33	40.0689	2.51595	7.95613	.157978
6.34	40.1956	2.51794	7.96241	.157729
6.35	40.3225	2.51992	7.96869	.157480
6.36	40.4496	2.52190	7.97496	.157233
6.37	40.5769	2.52389	7.98123	.156986
6.38	40.7044	2.52587	7.98749	.156740
6.39	40.8321	2.52784	7.99375	.156495
6.40	40.9600	2.52982	8.00000	.156250
6.41	41.0881	2.53180	8.00625	.156006
6.42	41.2164	2.53377	8.01249	.155763
6.43	41.3449	2.53574	8.01873	.155521
6.44	41.4736	2.53772	8.02496	.155280
6.45	41.6025	2.53969	8.03119	.155039
6.46	41.7316	2.54165	8.03741	.154799
6.47	41.8609	2.54362	8.04363	.154560
6.48	41.9904	2.54558	8.04984	.154321
6.49	42.1201	2.54755	8.05605	.154083
6.50	42.2500	2.54951	8.06226	.153846
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
6.50	42.2500	2.54951	8.06226	.153846
6.51	42.3801	2.55147	8.06846	.153610
6.52	42.5104	2.55343	8.07465	.153374
6.53	42.6409	2.55539	8.08084	.153139
6.54	42.7716	2.55734	8.08703	.152905
6.55	42.9025	2.55930	8.09321	.152672
6.56	43.0336	2.56125	8.09938	.152439
6.57	43.1649	2.56320	8.10555	.152207
6.58	43.2964	2.56515	8.11172	.151976
6.59	43.4281	2.56710	8.11788	.151745
6.60	43.5600	2.56905	8.12404	.151515
6.61	43.6921	2.57099	8.13019	.151286
6.62	43.8244	2.57294	8.13634	.151057
6.63	43.9569	2.57488	8.14248	.150830
6.64	44.0896	2.57682	8.14862	.150602
6.65	44.2225	2.57876	8.15475	.150376
6.66	44.3556	2.58070	8.16088	.150150
6.67	44.4889	2.58263	8.16701	.149925
6.68	44.6224	2.58457	8.17313	.149701
6.69	44.7561	2.58650	8.17924	.149477
6.70	44.8900	2.58844	8.18535	.149254
6.71	45.0241	2.59037	8.19146	.149031
6.72	45.1584	2.59230	8.19756	.148810
6.73	45.2929	2.59422	8.20366	.148588
6.74	45.4276	2.59615	8.20975	.148368
6.75	45.5625	2.59808	8.21584	.148148
6.76	45.6976	2.60000	8.22192	.147929
6.77	45.8329	2.60192	8.22800	.147710
6.78	45.9684	2.60384	8.23408	.147493
6.79	46.1041	2.60576	8.24015	.147275
6.80	46.2400	2.60768	8.24621	.147059
6.81	46.3761	2.60960	8.25227	.146843
6.82	46.5124	2.61151	8.25833	.146628
6.83	46.6489	2.61343	8.26438	.146413
6.84	46.7856	2.61534	8.27043	.146199
6.85	46.9225	2.61725	8.27647	.145985
6.86	47.0596	2.61916	8.28251	.145773
6.87	47.1969	2.62107	8.28855	.145560
6.88	47.3344	2.62298	8.29458	.145349
6.89	47.4721	2.62488	8 30060	.145138
6.90	47.6100	2.62679	8.30662	.144928
6.91	47.7481	2.62869	8.31264	.144718
6.92	47.8864	2.63059	8.31865	.144509
6.93	48.0249	2.63249	8.32466	.144300
6.94	48.1636	2.63439	8.33067	.144092
6.95	48.3025	2.63629	8.33667	.143885
6.96	48.4416	2.63818	8.34266	.143678
6.97	48.5809	2.64008	8.34865	.143472
6.98	48.7204	2.64197	8.35464	.143266
6.99	48.8601	2.64386	8.36062	.143062
7.00	49.0000	2.64575	8.36660	.142857
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Continued*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
7.00	49.0000	2.64575	8.36660	.142857
7.01	49.1401	2.64764	8.37257	.142653
7.02	49.2804	2.64953	8.37854	.142450
7.03	49.4209	2.65141	8.38451	.142248
7.04	49.5616	2.65330	8.39047	.142045
7.05	49.7025	2.65518	8.39643	.141844
7.06	49.8436	2.65707	8.40238	.141643
7.07	49.9849	2.65895	8.40833	.141443
7.08	50.1264	2.66083	8.41427	.141243
7.09	50.2681	2.66271	8.42021	.141044
7.10	50.4100	2.66458	8.42615	.140845
7.11	50.5521	2.66646	8.43208	.140647
7.12	50.6944	2.66833	8.43801	.140449
7.13	50.8369	2.67021	8.44393	.140252
7.14	50.9796	2.67208	8.44985	.140056
7.15	51.1225	2.67395	8.45577	.139860
7.16	51.2656	2.67582	8.46168	.139665
7.17	51.4089	2.67769	8.46759	.139470
7.18	51.5524	2.67955	8.47349	.139276
7.19	51.6961	2.68142	8.47939	.139082
7.20	51.8400	2.68328	8.48528	.138889
7.21	51.9841	2.68514	8.49117	.138696
7.22	52.1284	2.68701	8.49706	.138504
7.23	52.2729	2.68887	8.50294	.138313
7.24	52.4176	2.69072	8.50882	.138122
7.25	52.5625	2.69258	8.51469	.137931
7.26	52.7076	2.69444	8.52056	.137741
7.27	52.8529	2.69629	8.52643	.137552
7.28	52.9984	2.69815	8.53229	.137363
7.29	53.1441	2.70000	8.53815	.137174
7.30	53.2900	2.70185	8.54400	.136986
7.31	53.4361	2.70370	8.54985	.136799
7.32	53.5824	2.70555	8.55570	.136612
7.33	53.7289	2.70740	8.56154	.136426
7.34	53.8756	2.70924	8.56738	.136240
7.35	54.0225	2.71109	8.57321	.136054
7.36	54.1696	2.71293	8.57904	.135870
7.37	54.3169	2.71477	8.58487	.135685
7.38	54.4644	2.71662	8.59069	.135501
7.39	54.6121	2.71846	8.59651	.135318
7.40	54.7600	2.72029	8.60233	.135135
7.41	54.9081	2.72213	8.60814	.134953
7.42	55.0564	2.72397	8.61394	.134771
7.43	55.2049	2.72580	8.61974	.134590
7.44	55.3536	2.72764	8.62554	.134409
7.45	55.5025	2.72947	8.63134	.134228
7.46	55.6516	2.73130	8.63713	.134048
7.47	55.8009	2.73313	8.64292	.133869
7.48	55.9504	2.73496	8.64870	.133690
7.49	56.1001	2.73679	8.65448	.133511
7.50	56.2500	2.73861	8.66025	.133333
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
7.50	56.2500	2.73861	8.66025	.133333
7.51	56.4001	2.74044	8.66603	.133156
7.52	56.5504	2.74226	8.67179	.132979
7.53	56.7009	2.74408	8.67756	.132802
7.54	56.8516	2.74591	8.68332	.132626
7.55	57.0025	2.74773	8.68907	.132450
7.56	57.1536	2.74955	8.69483	.132275
7.57	57.3049	2.75136	8.70057	.132100
7.58	57.4564	2.75318	8.70632	.131926
7.59	57.6081	2.75500	8.71206	.131752
7.60	57.7600	2.75681	8.71780	.131579
7.61	57.9121	2.75862	8.72353	.131406
7.62	58.0644	2.76043	8.72926	.131234
7.63	58.2169	2.76225	8.73499	.131062
7.64	58.3696	2.76405	8.74071	.130890
7.65	58.5225	2.76586	8.74643	.130719
7.66	58.6756	2.76767	8.75214	.130548
7.67	58.8289	2.76948	8.75785	.130378
7.68	58.9824	2.77128	8.76356	.130208
7.69	59.1361	2.77308	8.76926	.130039
7.70	59.2900	2.77489	8.77496	.129870
7.71	59.4441	2.77669	8.78066	.129702
7.72	59.5984	2.77849	8.78635	.129534
7.73	59.7529	2.78029	8.79204	.129366
7.74	59.9076	2.78209	8.79773	.129199
7.75	60.0625	2.78388	8.80341	.129032
7.76	60.2176	2.78568	8.80909	.128866
7.77	60.3729	2.78747	8.81476	.128700
7.78	60.5284	2.78927	8.82043	.128535
7.79	60.6841	2.79106	8.82610	.128370
7.80	60.8400	2.79285	8.83176	.128205
7.81	60.9961	2.79464	8.83742	.128041
7.82	61.1524	2.79643	8.84308	.127877
7.83	61.3089	2.79821	8.84873	.127714
7.84	61.4656	2.80000	8.85438	.127551
7.85	61.6225	2.80179	8.86002	.127389
7.86	61.7796	2.80357	8.86566	.127226
7.87	61 9369	2.80535	8.87130	.127065
7.88	62.0944	2.80713	8.87694	.126904
7.89	62.2521	2.80891	8.88257	.126743
7.90	62.4100	2.81069	8.88819	.126582
7.91	62.5681	2.81247	8.89382	.126422
7.92	62.7264	2.81425	8.89944	.126263
7.93	62.8849	2.81603	8.90505	.126103
7.94	63.0436	2.81780	8.91067	.125945
7.95	63.2025	2.81957	8.91628	.125786
7.96	63.3616	2.82135	8.92188	.125628
7.97	63.5209	2.82312	8.92749	.125471
7.98	63.6804	2.82489	8.93308	.125313
7.99	63.8401	2.82666	8.93868	.125156
8.00	64.0000	2.82843	8.94427	.125000
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Continued*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
8.00	64.0000	2.82843	8.94427	.125000
8.01	64.1601	2.83019	8.94986	.124844
8.02	64.3204	2.83196	8.95545	.124688
8.03	64.4809	2.83373	8.96103	.124533
8.04	64.6416	2.83549	8.96660	.124378
8.05	64.8025	2.83725	8.97218	.124224
8.06	64.9636	2.83901	8.97775	.124069
8.07	65.1249	2.84077	8.98332	.123916
8.08	65.2864	2.84253	8.98888	.123762
8.09	65.4481	2.84429	8.99444	.123609
8.10	65.6100	2.84605	9.00000	.123457
8.11	65.7721	2.84781	9.00555	.123305
8.12	65.9344	2.84956	9.01110	.123153
8.13	66.0969	2.85132	9.01665	.123001
8.14	66.2596	2.85307	9.02219	.122850
8.15	66.4225	2.85482	9.02774	.122699
8.16	66.5856	2.85657	9.03327	.122549
8.17	66.7489	2.85832	9.03881	.122399
8.18	66.9124	2.86007	9.04434	.122249
8.19	67.0761	2.86182	9.04986	.122100
8.20	67.2400	2.86356	9.05539	.121951
8.21	67.4041	2.86531	9.06091	.121803
8.22	67.5684	2.86705	9.06642	.121655
8.23	67.7329	2.86880	9.07193	.121507
8.24	67.8976	2.87054	9.07744	.121359
8.25	68.0625	2.87228	9.08295	.121212
8.26	68.2276	2.87402	9.08845	.121065
8.27	68.3929	2.87576	9.09395	.120919
8.28	68.5584	2.87750	9.09945	.120773
8.29	68.7241	2.87924	9.10494	.120627
8.30	68.8900	2.88097	9.11043	.120482
8.31	69.0561	2.88271	9.11592	.120337
8.32	69.2224	2.88444	9.12140	.120192
8.33	69.3889	2.88617	9.12688	.120048
8.34	69.5556	2.88791	9.13236	.119904
8.35	69.7225	2.88964	9.13783	.119760
8.36	69.8896	2.89137	9.14330	.119617
8.37	70.0569	2.89310	9.14877	.119474
8.38	70.2244	2.89482	9.15423	.119332
8.39	70.3921	2.89655	9.15969	.119190
8.40	70.5600	2.89828	9.16515	.119048
8.41	70.7281	2.90000	9.17061	.118906
8.42	70.8964	2.90172	9.17606	.118765
8.43	71.0649	2.90345	9.18150	.118624
8.44	71.2336	2.90517	9.18695	.118483
8.45	71.4025	2.90689	9.19239	.118343
8.46	71.5716	2.90861	9.19783	.118203
8.47	71.7409	2.91033	9.20326	.118064
8.48	71.9104	2.91204	9.20869	.117925
8.49	72.0801	2.91376	9.21412	.117786
8.50	72.2500	2.91548	9.21954	.117647
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
8.50	72.2500	2.91548	9.21954	.117647
8.51	72.4201	2.91719	9.22497	.117509
8.52	72.5904	2.91890	9.23038	.117371
8.53	72.7609	2.92062	9.23580	.117233
8.54	72.9316	2.92233	9.24121	.117096
8.55	73.1025	2.92404	9.24662	.116959
8.56	73.2736	2.92575	9.25203	.116822
8.57	73.4449	2.92746	9.25743	.116686
5.58	73.6164	2.92916	9.26283	.116550
8.59	73.7881	2.93087	9.26823	.116414
8.60	73.9600	2.93258	9.27362	.116279
8.61	74.1321	2.93428	9.27901	.116144
8.62	74.3044	2.93598	9.28440	.116009
8.63	74.4769	2.93769	9.28978	.115875
8.64	74.6496	2.93939	9.29516	.115741
8.65	74.8225	2.94109	9.30054	.115607
8.66	74.9956	2.94279	9.30591	.115473
8.67	75.1689	2.94449	9.31128	.115340
8.68	75.3424	2.94618	9.31665	.115207
8.69	75.5161	2.94788	9.32202	.115075
8.70	75.6900	2.94958	9.32738	.114943
8.71	75.8641	2.95127	9.33274	.114811
8.72	76.0384	2.95296	9.33809	.114679
8.73	76.2129	2.95466	9.34345	.114548
8.74	76.3876	2.95635	9.34880	.114416
8.75	76.5625	2.95804	9.35414	.114286
8.76	76.7376	2.95973	9.35949	.114155
8.77	76.9129	2.96142	9.36483	.114025
8.78	77.0884	2.96311	9.37017	.113895
8.79	77.2641	2.96479	9.37550	.113766
8.80	77.4400	2.96648	9.38083	.113636
8.81	77.6161	2.96816	9.38616	.113507
8.82	77.7924	2.96985	9.39149	.113379
8.83	77.9689	2.97153	9.39681	.113250
8.84	78.1456	2.97321	9.40213	.113122
8.85	78.3225	2.97489	9.40744	.112994
8.86	78.4996	2.97658	9.41276	.112867
8.87	78.6769	2.97825	9.41807	.112740
8.88	78.8544	2.97993	9.42338	.112613
8.89	79.0321	2.98161	9.42868	.112486
8.90	79.2100	2.98329	9.43398	.112360
8.91	79.3881	2.98496	9.43928	.112233
8.92	79.5664	2.98664	9.44458	.112108
8.93	79.7449	2.98831	9.44987	.111982
8.94	79.9236	2.98998	9.45516	.111857
8.95	80.1025	2.99166	9.46044	.111732
8.96	80.2816	2.99333	9.46573	.111607
8.97	80.4609	2.99500	9.47101	.111483
8.98	80.6404	2.99666	9.47629	.111359
8.99	80.8201	2.99833	9.48156	.111235
9.00	81.0000	3.00000	9.48683	.111111
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

SQUARES — SQUARE ROOTS — RECIPROCALS (*Concluded*)

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
9.00	81.0000	3.00000	9.48683	.111111
9.01	81.1801	3.00167	9.49210	.110988
9.02	81.3604	3.00333	9.49737	.110865
9.03	81.5409	3.00500	9.50263	.110742
9.04	81.7216	3.00666	9.50789	.110619
9.05	81.9025	3.00832	9.51315	.110497
9.06	82.0836	3.00998	9.51840	.110375
9.07	82.2649	3.01164	9.52365	.110254
9.08	82.4464	3.01330	9.52890	.110132
9.09	82.6281	3.01496	9.53415	.110011
9.10	82.8100	3.01662	9.53939	.109890
9.11	82.9921	3.01828	9.54463	.109769
9.12	83.1744	3.01993	9.54987	.109649
9.13	83.3569	3.02159	9.55510	.109529
9.14	83.5396	3.02324	9.56033	.109409
9.15	83.7225	3.02490	9.56556	.109290
9.16	83.9056	3.02655	9.57079	.109170
9.17	84.0889	3.02820	9.57601	.109051
9.18	84.2724	3.02985	9.58123	.108932
9.19	84.4561	3.03150	9.58645	.108814
9.20	84.6400	3.03315	9.59166	.108696
9.21	84.8241	3.03480	9.59687	.108578
9.22	85.0084	3.03645	9.60208	.108460
9.23	85.1929	3.03809	9.60729	.108342
9.24	85.3776	3.03974	9.61249	.108225
9.25	85.5625	3.04138	9.61769	.108108
9.26	85.7476	3.04302	9.62289	.107991
9.27	85.9329	3.04467	9.62808	.107875
9.28	86.1184	3.04631	9.63328	.107759
9.29	86.3041	3.04795	9.63846	.107643
9.30	86.4900	3.04959	9.64365	.107527
9.31	86.6761	3.05123	9.64883	.107411
9.32	86.8624	3.05287	9.65401	.107296
9.33	87.0489	3.05450	9.65919	.107181
9.34	87.2356	3.05614	9.66437	.107066
9.35	87.4225	3.05778	9.66954	.106952
9.36	87.6096	3.05941	9.67471	.106838
9.37	87.7969	3.06105	9.67988	.106724
9.38	87.9844	3.06268	9.68504	.106610
9.39	88.1721	3.06431	9.69020	.106496
9.40	88.3600	3.06594	9.69536	.106383
9.41	88.5481	3.06757	9.70052	.106270
9.42	88.7364	3.06920	9.70567	.106157
9.43	88.9249	3.07083	9.71082	.106045
9.44	89.1136	3.07246	9.71597	.105932
9.45	89.3025	3.07409	9.72111	.105820
9.46	89.4916	3.07571	9.72625	.105708
9.47	89.6809	3.07734	9.73139	.105597
9.48	89.8704	3.07896	9.73653	.105485
9.49	90.0601	3.08058	9.74166	.105374
9.50	90.2500	3.08221	9.74679	.105263
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$
9.50	90.2500	3.08221	9.74679	.105263
9.51	90.4401	3.08383	9.75192	.105152
9.52	90.6304	3.08545	9.75705	.105042
9.53	90.8209	3.08707	9.76217	.104932
9.54	91.0116	3.08869	9.76729	.104822
9.55	91.2025	3.09031	9.77241	.104712
9.56	91.3936	3.09192	9.77753	.104603
9.57	91.5849	3.09354	9.78264	.104493
9.58	91.7764	3.09516	9.78775	.104384
9.59	91.9681	3.09677	9.79285	.104275
9.60	92.1600	3.09839	9.79796	.104167
9.61	92.3521	3.10000	9.80306	.104058
9.62	92.5444	3.10161	9.80816	.103950
9.63	92.7369	3.10322	9.81326	.103842
9.64	92.9296	3.10483	9.81835	.103734
9.65	93.1225	3.10644	9.82344	.103627
9.66	93.3156	3.10805	9.82853	.103520
9.67	93.5089	3.10966	9.83362	.103413
9.68	93.7024	3.11127	9.83870	.103306
9.69	93.8961	3.11288	9.84378	.103199
9.70	94.0900	3.11448	9.84886	.103093
9.71	94.2841	3.11609	9.85393	.102987
9.72	94.4784	3.11769	9.85901	.102881
9.73	94.6729	3.11929	9.86408	.102775
9.74	94.8676	3.12090	9.86914	.102669
9.75	95.0625	3.12250	9.87421	.102564
9.76	95.2576	3.12410	9.87927	.102459
9.77	95.4529	3.12570	9.88433	.102354
9.78	95.6484	3.12730	9.88939	.102249
9.79	95.8441	3.12890	9.89444	.102145
9.80	96.0400	3.13050	9.89949	.102041
9.81	96.2361	3.13209	9.90454	.101937
9.82	96.4324	3.13369	9.90959	.101833
9.83	96.6289	3.13528	9.91464	.101729
9.84	96.8256	3.13688	9.91968	.101626
9.85	97.0225	3.13847	9.92472	.101523
9.86	97.2196	3.14006	9.92975	.101420
9.87	97.4169	3.14166	9.93479	.101317
9.88	97.6144	3.14325	9.93982	.101215
9.89	97.8121	3.14484	9.94485	.101112
9.90	98.0100	3.14643	9.94987	.101010
9.91	98.2081	3.14802	9.95490	.100908
9.92	98.4064	3.14960	9.95992	.100806
9.93	98.6049	3.15119	9.96494	.100705
9.94	98.8036	3.15278	9.96995	.100604
9.95	99.0025	3.15436	9.97497	.100503
9.96	99.2016	3.15595	9.97998	.100402
9.97	99.4009	3.15753	9.98499	.100301
9.98	99.6004	3.15911	9.98999	.100200
9.99	99.8001	3.16070	9.99500	.100100
10.00	100.000	3.16228	10.0000	.100000
n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$1/n$

APPENDIX D

TABLE OF VALUES OF ΣX^2 AND ΣX^4 FOR COMPUTING LEAST SQUARES TREND LINES

The values of ΣX^2 and ΣX^4 may be read from the table opposite the value of N (representing the number of years to which the trend is fitted). For values of N greater than 60, use the formulas below. Origin is at mid-range so that $\Sigma X = 0$. For an odd number of years, the X scale is represented by . . . -2, -1, 0, $+1$, $+2$ . . . For an even number of years, the X scale is represented by . . . -3, -1, $+1$, $+3$ . . .

Odd Number of Years	Even Number of Years
$\Sigma X^2 = \frac{N(N^2-1)}{12}$	$\Sigma X^2 = \frac{N(N^2-1)}{3}$
$\Sigma X^4 = \frac{3N^2-7}{20}\Sigma X^2$	$\Sigma X^4 = \frac{3N^2-7}{5}\Sigma X^2$
or $\Sigma X^4 = \frac{3N^5-10N^3+7N}{240}$	or $\Sigma X^4 = \frac{3N^5-10N^3+7N}{15}$

VALUES OF ΣX^2 AND ΣX^4 FOR COMPUTING LEAST SQUARES TREND LINES

Odd number of years, X unit, 1 year			Even number of years, X unit $\frac{1}{2}$ year		
N	ΣX^2	ΣX^4	N	ΣX^2	ΣX^4
3	2	2	4	20	164
5	10	34	6	70	1,414
7	28	196	8	168	6,216
9	60	708	10	330	19,338
11	110	1,958	12	572	48,620
13	182	4,550	14	910	105,742
15	280	9,352	16	1,360	206,992
17	408	17,544	18	1,938	374,034
19	570	30,666	20	2,660	634,676
21	770	50,666	22	3,542	1,023,638
23	1,012	79,948	24	4,600	1,583,320
25	1,300	121,420	26	5,850	2,364,570
27	1,638	178,542	28	7,308	3,427,452
29	2,030	255,374	30	8,990	4,842,014
31	2,480	356,624	32	10,912	6,689,056
33	2,992	469,696	34	13,090	9,060,898
35	3,570	654,738	36	15,540	12,062,148
37	4,218	864,690	38	18,278	15,810,470
39	4,940	1,125,332	40	21,320	20,537,352
41	5,740	1,115,332	42	24,682	26,088,874
43	6,622	1,831,294	44	28,380	32,926,476
45	7,590	2,302,806	46	32,430	41,127,726
47	8,628	2,862,488	48	36,848	50,887,088
49	9,800	3,526,040	50	41 650	62,416,690
51	11,050	4,307,290	52	46,852	75,947,092
53	12,402	5,221,242	54	52,470	91,728,054
55	13,860	6,284,124	56	58,520	110,029,304
57	15,428	7,513,436	58	65,018	131,141,306
59	17,110	8,927,998	60	71,980	155,376,028

E APPENDIX

USE OF LOGARITHMS

Logarithms have been used for more than 300 years to reduce the labor of arithmetic computations. Calculating machines have rather generally replaced logarithms in statistical computations for performing multiplication and division, but logarithms are still used to extract roots and raise to powers. They are practically a necessity in computing the geometric mean and fitting certain types of curves.

A *logarithm* of any positive number to a given base is the power to which the base must be raised to give the number. The number corresponding to the logarithm is the *antilogarithm*. Because of its many advantages, the system of logarithms whose base is 10, known as "common logarithms," is generally used in numerical computation. $\text{Log}_{10}N$ is the common logarithm of N and is written Log N with the base 10 understood.

The common logarithm of 100 is 2, since $10^2 = 100$. Likewise the common logarithm of 1,000 is 3, since $10^3 = 1{,}000$. Only the exponent "2" is written, since the base is always 10 for common logarithms. For the integral powers of 10, the logarithm is an integer. For all numbers between the integral powers of 10, the logarithm consists of an integer (positive, negative, or zero) plus a positive number less than one, always written as a decimal fraction. The integral portion of the logarithm is called the *characteristic*, and the decimal portion, the *mantissa*. The characteristic may be positive or negative, depending upon the location of the decimal point, but the mantissa is always positive.

Finding the Mantissa

The mantissas of the common logarithms of all numbers having the same sequence of figures are identical regardless of the position of the decimal point. The values of the mantissas for sequences of three figures are given to four decimals in Appendix F. The first two digits are at the left of the page and the third is at the top. The mantissa of 135 is found to be .1303 by reading down the first column to "13" and across the table to the column headed "5."

Determining the Characteristic

The characteristic of the common logarithm of any number may be determined by the following procedure. Locate an imaginary decimal so placed as to make a number between 1 and 10. For all numbers of 10 or more, the imaginary decimal is placed to the left of the real decimal and the characteristic is positive.

For example, in the number 13.5 the imaginary decimal (marked ^) is located between 1 and 3 (1^3.5). The number of integers between the

decimal and the imaginary decimal is the characteristic of the logarithm. In this example, the characteristic is 1, and the logarithm of 13.5 is 1.1303. Other examples of determining the characteristic are given below.

Number	Characteristic	Logarithm
1∧350.	3	3.1303
1∧35.0	2	2.1303
1∧3.5	1	1.1303

For numbers between 1 and 10, the imaginary decimal falls at the same place as the decimal point, making the number of integers between them zero. The logarithm of 1.35 is .1303. For numbers less than 1, the imaginary decimal is to the right of the decimal point and the sign of the characteristic is minus. For example, the characteristic of .001∧35 is -3 and the logarithm of .00135 is .1303 $-$ 3, or -2.8697, or 7.1303 $-$ 10.

Interpolation

The table in Appendix F gives the mantissas for numbers of three digits. For a sequence of more than three numbers, the mantissa may be approximated from this table by interpolating between the logarithms of the numbers immediately preceding and following the number for which the logarithm is sought. The logarithm of 13.54 is between 1.1303 and 1.1335, the logarithms of 13.5 and 13.6. The difference between the two logarithms is .0032, so it is assumed that the logarithm of 13.54 will be larger than 1.1303 by 4/10th of the difference between 1.1303 and 1.1335. This linear interpolation gives the following value for the logarithm, with the last digit approximate.

$$\text{Log } 13.54 = 1.1303 + (.4 \times .0032) = 1.1303 + .0013 = 1.1316$$

Greater accuracy can be obtained by using tables that give the mantissa for a sequence of more than three digits and that carry the values to a greater number of decimal places.

Finding the Antilogarithm

After a calculation has been made with logarithms, it is necessary to find the number represented by the logarithm. This number, called the *antilogarithm*, is determined by a process that is the reverse of finding the logarithm of a number. This may be illustrated by an example in which the problem is to find the value of N when $\log N = 2.7885$. The digits of the number for which 2.7885 is the logarithm are determined from the mantissa (.7885), and the location of the decimal point is determined by the characteristic (2).

The mantissa is located in the table of logarithms, or if the exact value does not appear in the table, the values just larger and smaller are found. In this example, 7885 falls between 7882 and 7889 (omitting decimals as in the table). The antilogarithms of these two logarithms are 614 and 615; thus, the antilogarithm of 7885 may be approximated by linear interpolation between these two numbers. Since the difference between the two logarithms, 7882 and 7889, is 7, the logarithm of 7885 is $\frac{3}{7}$th of the distance from 7882 in the direction of 7889. This means that the antilogarithm is approximately $614\frac{3}{7}$, or 614.4.

Locating the imaginary decimal point to make a number between 1 and 10 gives $6_{\wedge}144$. Since the characteristic of the logarithm is $+2$, the real decimal point is two digits to the right of the imaginary decimal, using the rule given above for finding the characteristic. The antilogarithm of 2.7885 is, therefore, $6_{\wedge}14.4$.

Multiplication

Multiplication is performed by adding the logarithms of the numbers being multiplied and then finding the antilogarithm as illustrated by the following example.

Find the value of 13.54×94.7

$$\begin{aligned} \log 13.54 &= 1.1316 \\ \log 94.7 &= 1.9763 \\ \hline \log N &= 3.1079 \\ N &= 1_{\wedge}282. \end{aligned}$$

The imaginary decimal is located between the first two digits and since the characteristic is $+3$, the real decimal point is placed three digits to the right of the imaginary decimal.

In using the logarithms of numbers less than 1, it is important to remember that while the characteristic is minus, the mantissa is always positive. The logarithm of .00135 was given before as .1303 -3, which is, of course, -2.8697. There are advantages, however, in writing this logarithm as $7.1303 - 10$. Since 7 was both added and subtracted, the value of the logarithm remains unchanged. $(7.1303 - 10 = -2.8697)$. This device is used in the following example.

Find the value of $64.7 \times .00135$

$$\begin{aligned} \log 64.7 &= 1.8109 \\ \log .00135 &= 7.1303 - 10 \\ \hline \log N &= 8.9412 - 10 \\ N &= .08_{\wedge}73 \end{aligned}$$

Division

Division is performed by subtracting the logarithm of the divisor from the logarithm of the dividend and then finding the antilogarithm, as illustrated in the following examples.

Find the value of $947 \div 13.54$

$$\begin{aligned} \log 947 &= 2.9763 \\ \log 13.54 &= 1.1316 \\ \log N &= 1.8447 \\ N &= 69.93 \end{aligned}$$

Find the value of $.00437 \div 64.7$

$$\begin{aligned} \log .00437 &= 7.6405 - 10 \\ \log 64.7 &= 1.8109 \\ \log N &= 5.8296 - 10 \\ N &= .00006755 \end{aligned}$$

Raising to Powers

To raise a number to a given power, multiply the logarithm of the number by the exponent of the power and find the antilogarithm.

Find the value of N if $N = 1.06^{35}$

$$\begin{aligned} \log N &= 35 \times \log 1.06 \\ \log 1.06 &= .0253 \\ \log N &= 35 \times .0253 = .8855 \\ N &= 7.68 \end{aligned}$$

Extracting Roots

To extract any root of a number, divide the logarithm of the number by the index of the root and find the antilogarithm.

Find the value of N if $N = \sqrt[9]{3.86}$

$$\begin{aligned} \log N &= \frac{\log 3.86}{9} \\ \log 3.86 &= .5866 \\ \log N &= \frac{.5866}{9} = .0652 \\ N &= 1.16 \end{aligned}$$

F APPENDIX

FOUR-PLACE LOGARITHMS

(See Appendix E for an explanation of the use of logarithms.)

No.	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
56	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
No.	0	1	2	3	4	5	6	7	8	9

FOUR-PLACE LOGARITHMS (*Continued*)

No.	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
No.	0	1	2	3	4	5	6	7	8	9

G APPENDIX

TABLE OF ORDINATES OF THE NORMAL CURVE AND AREAS UNDER THE NORMAL CURVE

Values of Ordinates

The values of the ordinates of a normal curve may be computed by multiplying the term $\frac{N}{\sigma}$ for the distribution by the values in the second column of the table. (The values in this column represent the ordinates of a normal curve for which $\frac{N}{\sigma} = 1$.) In a distribution for which $N = 1{,}000$, $\bar{X} = \$400$, and $\sigma = \$20$, the maximum ordinate ($Y_o$) is computed as follows, since Y_o is the ordinate when $\frac{x}{\sigma} = 0$.

$$Y_o = \frac{N}{\sigma} .39894 = \frac{1{,}000}{20} .39894 = 19.947$$

The ordinate for any value of X may be computed in a similar manner. When $X = 440$, $\frac{x}{\sigma} = \frac{X - \bar{X}}{\sigma} = \frac{440 - 400}{20} = 2$. The table gives the value of the ordinate when $\frac{x}{\sigma} = 2$ as .05399, which is the ordinate of a curve for which $\frac{N}{\sigma} = 1$. When $\frac{N}{\sigma} = \frac{1{,}000}{20}$, $Y = \frac{1{,}000}{20} .05399 = 2.70$. If the normal curve is to be plotted on the grid with a histogram or frequency polygon, the standard deviation (σ) is expressed in class interval units.

Areas Between Maximum Ordinate and Ordinate at $\frac{x}{\sigma}$

The values in the third column of the table show the fraction of the area of the normal curve that lies between the maximum ordinate (Y_o) and the ordinate at various distances from the maximum ordinate, measured by $\frac{X - \bar{X}}{\sigma}$ or $\frac{x}{\sigma}$. Reading down the table to $\frac{x}{\sigma} = 1.0$, the fraction of the curve is .34134. Since the normal curve is symmetrical, slightly more than 68 per cent of the area of the normal curve lies within the range of $+1\sigma$ and -1σ. This means that 68 per cent of the individual values of a normal distribution fall within this range.

The percentage of items falling within any range expressed in standard deviation units can be computed in the same manner by doubling the fraction in the third column of the table. For example, 95 per cent of the items in a normal distribution fall between $\pm 1.96\sigma$, and 99 per cent fall within the range $\pm 2.576\sigma$.

ORDINATES OF THE NORMAL CURVE AND AREAS UNDER THE NORMAL CURVE

$\frac{x}{\sigma}$	Ordinate at $\frac{x}{\sigma}$	Area between maximum ordinate (Y_o) and ordinate at $\frac{x}{\sigma}$
.0	.39894	.00000
.1	.39695	.03983
.2	.39104	.07926
.3	.38139	.11791
.4	.36827	.15542
.5	.35207	.19146
.6	.33322	.22575
.7	.31225	.25804
.8	.28969	.28814
.9	.26609	.31594
1.0	.24197	.34134
1.1	.21785	.36433
1.2	.19419	.38493
1.3	.17137	.40320
1.4	.14973	.41924
1.5	.12952	.43319
1.6	.11092	.44520
1.7	.09405	.45543
1.8	.07895	.46407
1.9	.06562	.47128
1.96	.05844	.47500
2.0	.05399	.47725
2.1	.04398	.48214
2.2	.03547	.48610
2.3	.02833	.48928
2.4	.02239	.49180
2.5	.01753	.49379
2.5758	.01445	.49500
2.6	.01358	.49534
2.7	.01042	.49653
2.8	.00792	.49744
2.9	.00595	.49813
3.0	.00443	.49865
3.1	.00327	.49903
3.2	.00238	.49931
3.3	.00172	.49952
3.4	.00123	.49966
3.5	.00087	.49977
3.6	.00061	.49984
3.7	.00042	.49989
3.8	.00029	.49993
3.9	.00020	.49995
4.0	.00014	.49997

H APPENDIX

SELECTED READING REFERENCES

Examining the following reading references will greatly benefit the student who wishes to supplement his study in the present text. In some instances a reading reference pertains to several topics discussed in the text, but in order to eliminate duplicate listing, a reference is shown only once under a part division.

INTRODUCTION

HUFF, DARRELL. *How to Lie with Statistics.* New York: W. W. Norton and Company, Inc., 1954.

KENDALL, M. G., and BUCKLAND, W. R. *A Dictionary of Statistical Terms.* New York: Hafner Publishing Co., 1957.

PART I — ASSEMBLING STATISTICAL DATA

HAUSER, PHILIP M., and LEONARD, WILLIAM R. *Government Statistics for Business Use.* New York: John Wiley & Sons, Inc., 1956.

JENKINSON, BRUCE L. *Bureau of the Census Manual of Tabular Presentation.* Washington, D. C.: U. S. Government Printing Office, 1949.

PART II — STATISTICAL DESCRIPTION

LUTZ, R. R. *Graphic Presentation Simplified.* New York: Funk and Wagnalls Co., 1949.

WILKS, S.S. *Elementary Statistical Analysis.* Princeton: Princeton University Press, 1949.

YULE, G. UDNY, and KENDALL, M. G. *An Introduction to the Theory of Statistics.* New York: Hafner Publishing Co., 1950.

PART III — STATISTICAL INDUCTION

COCHRAN, WILLIAM G. *Sampling Techniques.* New York: John Wiley & Sons, Inc., 1953.

DEMING, W. EDWARDS. *Some Theory of Sampling.* New York: John Wiley & Sons, Inc., 1950.

HANSEN, MORRIS H., HURWITZ, WILLIAM N., and MADOW, WILLIAM G. *Sampling Survey Methods and Theory.* (*Methods and Applications*, Vol. I) (*Theory*, Vol. II) New York: John Wiley & Sons, Inc., 1953.

SUKHATME, PANDURANG V. *Sampling Theory of Surveys with Applications.* New Delhi, India: The Indian Society of Agricultural Statistics, and Ames, Iowa: The Iowa State College Press, 1954.

YATES, FRANK. *Sampling Methods for Censuses and Surveys.* New York: Hafner Publishing Co., 1953.

PART IV — ANALYSIS OF BUSINESS CHANGE

BURNS, ARTHUR F. *Production Trends in the United States Since 1870.* New York: National Bureau of Economic Research, 1934.

BURNS, ARTHUR F., and MITCHELL, WESLEY C. *Measuring Business Cycles.* New York: National Bureau of Economic Research, 1946.

FISHER, IRVING. *The Making of Index Numbers.* Boston: Houghton Mifflin Co., 1927.

KUZNETS, SIMON. *Seasonal Variations in Industry and Trade.* New York: National Bureau of Economic Research, 1933.

MITCHELL, WESLEY C. *Business Cycles. (The Problem and Its Setting,* Vol. 1) New York: National Bureau of Economic Research, 1927.

MITCHELL, WESLEY C. *The Making and Using of Index Numbers. ("Publication of the U. S. Bureau of Labor Statistics,"* Bulletin No. 656.) Washington, D. C.: Government Printing Office, 1938.

MUDGETT, BRUCE D. *Index Numbers.* New York: John Wiley & Sons, Inc., 1951.

PERSONS, WARREN M *The Construction of Index Numbers.* Boston: Houghton Mifflin Co., 1928.

SHISKIN, JULIUS, and EISENPRESS, HARRY. "Seasonal Adjustments by Electronic Computer Methods," *Journal of the American Statistical Association,* Vol. 52, No. 280 (December, 1957).

PART V — ANALYSIS OF RELATIONSHIP

DEAN, JOEL. *Statistical Cost Functions of a Hosiery Mill.* Chicago: University of Chicago Press, 1941.

DEMING, W. EDWARDS. *Statistical Adjustment of Data.* New York: John Wiley & Sons, Inc., 1943.

EZEKIEL, MORDECAI. *Methods of Correlation Analysis.* New York: John Wiley & Sons, Inc., 1941.

FERBER, ROBERT. *Statistical Techniques in Market Research.* New York: McGraw-Hill Book Co., 1949.

SCHULTZ, HENRY. *The Theory and Measurement of Demand.* Chicago: University of Chicago Press, 1938.

PART VI — APPLICATIONS IN BUSINESS

Ch. 20 — Business Barometers

COLE, ARTHUR H. *Measures of Business Change.* Homewood, Illinois: Richard D. Irwin, Inc., 1952.

GARVY, GEORGE. *The Development of Bank Debits and Clearings and Their Use in Economic Analysis.* Washington, D. C.: Board of Governors of the Federal Reserve System, 1952.

SNYDER, RICHARD M. *Measuring Business Changes.* New York: John Wiley & Sons, Inc., 1955.

Ch. 21 — Business Forecasting

ABRAMSON, ADOLPH G., and MACK, RUSSELL H. (eds.). *Business Forecasting in Practice: Principles and Cases.* New York: John Wiley & Sons, Inc., 1956.

BRATT, ELMER C. *Business Cycles and Forecasting.* Homewood, Illinois: Richard D. Irwin, Inc., 1953.

DAUTEN, CARL A. *Business Fluctuations and Forecasting.* Cincinnati: South-Western Publishing Company, 1954.

GORDON, ROBERT A. *Business Fluctuations.* New York: Harper & Brothers, 1952.

MAISEL, SHERMAN J. *Fluctuations, Growth, and Forecasting.* New York: John Wiley & Sons, Inc., 1957.

National Bureau of Economic Research, Conference on Research in Income and Wealth. *Short-Term Economic Forecasting, Studies in Income and Wealth,* Vol. 17. Princeton: Princeton University Press, 1955.

National Bureau of Economic Research, Conference on Research in Income and Wealth. *Long-Range Economic Projection, Studies in Income and Wealth,* Vol. 16. Princeton: Princeton University Press, 1954.

NEWBURY, FRANK D. *Business Forecasting: Principles and Practices.* New York: McGraw-Hill Book Co., 1952.

Ch. 22 — Marketing Research

BOYD, HARPER W., JR., and WESTFALL, RALPH. *Marketing Research: Text and Cases.* Homewood, Illinois: Richard D. Irwin, Inc., 1956.

FERBER, ROBERT, and WALES, HUGH G. (eds.). *Motivation and Market Behavior.* Homewood, Illinois: Richard D. Irwin, Inc., 1958.

LUCK, DAVID J., and WALES, HUGH G. *Marketing Research.* New York: Prentice-Hall, Inc., 1952.

Ch. 23 — Statistical Quality Control

DUNCAN, ACHESON J. *Quality Control and Industrial Statistics.* Homewood, Illinois: Richard D. Irwin, Inc., 1952.

GRANT, EUGENE L. *Statistical Quality Control.* New York: McGraw-Hill Book Co., 1952.

Ch. 24 — Community Economic Surveys

DALE, ALFRED G. *An Economic Survey Method for Small Areas.* Austin: Bureau of Business Research, The University of Texas, 1955.

National Bureau of Economic Research, Conference on Research in Income and Wealth. *Input-Output Analysis: An Appraisal, Studies in Income and Wealth,* Vol. 18. Princeton: Princeton University Press, 1955.

INDEX

A

Acceptance number, 554
Acceptance-rectification sampling plan, 554
Acceptance-rejection sampling plan, 554
Acceptance sampling, 554
Advertising media and agencies, as agents of market research, 525
Advertising research, 531; copy testing, 533; printed media, 531; radio and television, 532
Aggregative method, of computing index numbers, 401
Agricultural Statistics, 75
Alienation, coefficient of, 439
Analysis, importance of time series, 255; of data, 8
Annual Statistical Report of the American Iron and Steel Institute, 78
Antilogarithm, finding the, 599
Area sample, 237
Arithmetic mean, 164; characteristics of, 183–188; formulas for, 164, 166, 167, 169; long method of computing, 165; of relatives, 389, 395; of relatives, weighted with base-year values, 395; of the logarithms, 174; short method of computing, 167; used to average specific seasonals, 335; weighted, 169
Array, defined, 177
Array of ratios to twelve-month moving average, 334; illustrated, 335
Assignable causes, 539
Automobile Facts and Figures, 78
Average deviation, 196; characteristics of, 205; computation of, 197, 198
Averages, 160; arithmetic mean, 164; chief characteristics of, 183; definition and computation of, 163; geometric mean, 174; median, 177; mode, 180; of relatives, 386; use of, 160

B

Bank debits, adjusting for seasonal variation, 362
Bank debits and clearings, as a business barometer, 474
Banking and Monetary Statistics, 76
Bar chart, 128, 147
Base period, comparison with an abnormal, 16
Basic industries in a community, and community economic surveys, 569
Basic industry, 560
Bias, 394; correcting the, 224; defined, 13; downward, 394; in statistical analysis, 13; in the sample, 224; upward, 394
Bureau of Labor Statistics, material published by, 76
Bureau of the Census, 73; current publications of, 81
Business barometers, 471; agricultural prices, 489; bank debits and clearings, 474 business sales, 485; commodity prices, 488; construction, 486; employment, 475; financial, 489; gross national product, 477, individual business processes, 484; industrial production, 485; inventories, 485; measures of total economic activity, 473; national income, 482; orders, 485; personal income, 482; plant and equipment expenditures, 488; retail prices, 489; the need for, 471; wholesale prices, 488
Business cycle, 249; **see also** Cyclical fluctuations; importance of business forecasting, 358; importance of, in business management, 356
Business facilities in a community, and community economic surveys, 567
Business forecasting, **see** Forecasting; importance of, 358
Business sales, as a business barometer, 485

C

Calendar variation, adjustment for, 256
Cartesian coordinates, 136
Catalog of United States Census Publications, 74
Causation, improper assumptions regarding, 15
Cause-and-effect relationships, and correlation, 422; understanding, 15
Censuses, recent economic, 73
Chain relatives, 119
Changes in seasonal pattern, abrupt, 340; gradual, 341
Characteristic, determining the, of logarithms, 598
Chart construction, mechanical features of, 153
Charts, 128; bar, 128, 147; circle and sector, 147; component parts, 147; construction of, 128; control, 540; line, 136; pictograms, 133; pie, 147; *R*, 548; semilogarithmic, **see** Semilogarithmic charts; statistical maps, 149; two- and three-dimension, 132; $\overline{X}$, 540; *Y* axis on, 145
Circle and sector chart, 147
Civic information, and community economic surveys, 566
Classes, defining the, 25; for a frequency distribution, 27
Classification, by differences of degree, 20, 40; by differences of kind, 19, 40, 129; by geographical division, 23; by time intervals, 23, 41; cumulative, 31; defined, 7; of data, 6; quantitative, 21; types of, 19
Cluster sampling, 236

Coefficient of alienation, 439
Coefficient of correlation, 439; formulas for, 440, 444
Coefficient of determination, 439, 455; formula for, 439
Coefficient of multiple correlation, computation of, 465
Coefficient of nondetermination, 439; formulas for, 439
Coefficient of partial correlation, 467; formula for, 468
Coefficient of variation, 207; computing, 207; formula for, 207
Commodity prices, as a business barometer, 488
Commodity Yearbook, 78
Community economic surveys, 559; basic data for, 563; basic industries, 569; basic industry and service industry, 560; business facilities, 567; civic information, 566; construction, 566; economic base of an area, 560; economic potential, 571; housing, 566; labor force, 565; nature of, 559; natural features, 563; population, 565; regional history, 563; services, 568; trade, 568; users of, 562
Component-parts chart, 147
Confidence interval, 221
Constant dollars, 268
Construction, as a business barometer, 486
Construction in a community, and community economic surveys, 566
Construction of charts, mechanical features of, 153
Construction Review, 81
Consultants, market research, 523
Consumer Markets, 78
Consumer research, 87
Consumer surveys, 526
Continuous variables, 29
Control, see Quality control
Control chart, computing limits from $\bar{R}$, 546; R, 548; $\bar{X}$, 540
Control chart for fraction defective, 550; illustrated, 552
Control chart for means, 540; illustrated, 543
Control chart for ranges, 548; illustrated, 549
Control chart for variables, 540
Control limits, computing from $\bar{\sigma}$, 545; computing from $\bar{\sigma}^2$, 542; lower, 543; upper, 543
Correlation, 411; abstract measures of relationship, 437; and cause-and-effect relationships, 422; coefficient of, 439; direct, 415; index of, 455; inverse, 415; linear, see Linear correlation; multiple, see Multiple correlation; negative, 415; nonlinear, see Nonlinear correlation; partial, 467; positive, 415; relationships involving more than two variables, 415
Correlation analysis, cross-classification tables, 420; estimates based on a scatter diagram, 418; line charts, 411; methods of, 411; nonlinear relationships, 419; scatter diagrams, 411; some examples of the use of, 446
Correlation ratio, 456
Correlation table, illustrated, 443
County and City Data Book of 1950, 74
County Business Patterns, 75
Cross-classification of data, 34
Cross-classification table, 420; illustrated, 421, 457
Crosshatched map, 151
Crude mode, 181
Cumulative classifications, 32
Current Population Reports: Labor Force, 476
Curve, growth, 301; normal, 210
Cyclical fluctuations, 248, 253, 356; forecasting, 379, 503; importance of business cycles in business management, 356; isolating, 358; lag series, 504; lead series, 504; problem of measuring, 358

D

Data, adjusting for seasonal variation, 352, 362; adjusting for secular trend, 359; analysis of, 8; classification of, 6, 19; collecting internal, 49; collection and tabulation of external, 85; comprehensive sources of statistical, 72, 79; external sources of, 9; for community economic surveys, 563; guides to published, 70; internal sources of, 9; noncomparable, 14; presentation of, 7; presenting statistical, 25; primary, 43; secondary, 43; source of, 43; use of statistical, in management decisions, 2
Data processing, electronic, 65, 102
Deflation, of dollar series, 268; of time series, 267
Department of Agriculture, material published by, 75
Dependent variable, 423
Depreciation at a uniform rate, 262, 263
Determination, coefficient of, 439, 455
Deviation, average, see Average deviation; quartile, see Quartile deviation; root-mean-square, 198; standard, see Standard deviation
Discontinuous variables, 29
Discrete variables, 29
Dispersion, 192; absolute, 194; chief characteristics of measures of, 204; importance of, 192; measures of, 194; measures of skewness, 207; normal distribution, 209; relative, 206
Disposable personal income, 484
Distribution, normal, 209, 240
Division, logarithms in, 601
Downward bias, 394

E

Economic activity, measures of, 473
Economic Almanac, The, 77
Economic base of an area, 560
Economic Indicators, 80
Economic potential, and community economic surveys, 571
Economic Report of the President, 76
Economic surveys, community, see Community economic surveys
Editor and Publisher Market Guide, 78
Electronic data processing, 65, 102
Empirical mode, 182

Employment, as a business barometer, 475
Erratic fluctuations, 248, 253
Estimates, 433; based on a scatter diagram, 418; reliability of, 434, 465; standard error of, 435, 436, 454
Exhaust method, of classifying data, 54, 101
External data, collecting, 85; examples of classifications of, 20, 22; tabulation of, 97
External sources of statistical data, 9
Extrapolation, defined, 314

F

Factor reversal test, 400
Factors test, 400
Federal Reserve Bulletin, 80, 476, 490
Federal Reserve System, material published by, 76
Financial barometers, 489; bank loans, 490; consumer credit, 490; consumer finances, 491; real estate credit, 491; security issues, 491; security prices, 492
Fluctuations, cyclical, see Cyclical fluctuations; erratic, 248, 253; random, 248, 253; seasonal variation, 248, 251; secular trend, 248, 249; types of, 248
Footnotes for tables, 41
Forecasting, 495; components vs. forecasting the total, 502; cyclical fluctuations, 503, 504; for an industry, 517; for individual business, 517; gross national product, 498, 506; seasonal variation, 502; secular trend, 496; using a measure of seasonal variation in, 349
Forecasts, importance of business, 358; of cyclical fluctuations in time series, 379; of secular trend, 314
Fraction defective, control chart for, 550; illustrated, 552
Frequency distribution, charts of, 131; defined, 22; principles for setting up classes for, 27
Frequency polygon, 136

G

Gallop election poll, 534
Gas Facts, 78
Geographical divisions, and charts, 129; and tables, 23
Geometric mean, 174; characteristics of, 188; formulas for, 174
Geometric mean of relatives, 391; weighted, 396
Gompertz curve, 303; fitting the, 303–306; formula for fitting the, 303
Government, an agent of market research, 522
Governmental sources of statistical data, 73, 79
Government purchases of goods and services, described, 481; forecasting, 513
Graphical interpolation, median, 179; mode, 181
Graphic method, of measuring secular trend, 281; trend values computed by, 282
Gross national product, analysis of, 506; as a business barometer, 477; components of, 479; defined, 477; forecasting, 498; 1946 to 1956, 478
Gross private domestic investment, described, 480; forecasting, 508
Growth curves, 301; Gompertz, 303; Pearl-Reed, 303

H

Handbook of Labor Statistics, 76
Harmonic mean, 173
Histogram, defined, 132
Historical Statistics of the United States, 1789–1945, 75
History of a community, and community economic surveys, 563
Housing in a community, and community economic surveys, 566
Hypotheses, testing, 238

I

Income, real, 268
Independent variable, 423
Index, computing an, from relatives, 367; of correlation, 455; of seasonal variation, 318; parity, 489; price, 269; using a changing, 370
Index numbers, 384; see also Relatives; as averages of relatives, 386; as simple relatives, 386; as unweighted arithmetic mean of relatives, 389; as unweighted geometric mean of relatives, 391; as weighted arithmetic mean, 395; as weighted geometric mean, 398; computed by method of aggregates, 401; construction of, 384; defined, 117, 385; general comments on construction of, 405; nature of, 384; problem of averages, 392; problem of weights, 393
Industrial Arts Guide, 71
Industrial marketing research, 527
Industrial production, as a business barometer, 485
Industry, basic, 560; nonbasic, 561; service, 560
Information, assembling relevant, 5
Internal data, collecting, 49; examples of classifications of, 20, 22
Internal records, basic types of, 50
Internal sources of statistical data, 9
Interpolation, 599; defined, 142; graphical, 179, 181; on point data and period data charts, 142; straight-line, 143
Interquartile range, 195
Inventories, as a business barometer, 485
Investigation, original, 5

J

Judgment sample, 218

L

Labor force of a community, and community economic surveys, 565
Lag series, 504
Lead series, 504

Least squares method, straight line fitted to logarithms by, 296
Least squares method of measuring secular trend, 286; for even number of years, 290; for odd number of years, 287
Ledger sheet, 52, 100
Life Insurance Fact Book, 78
Linear correlation, 430; coefficient of alienation, 439; coefficient of correlation, 439; coefficient of determination, 439; coefficient of nondetermination, 439; estimates, 433; linear regression equation, 430; measurement of, 430; product-moment method of computation, 440, 442; reliability of estimates, 433; standard error of estimate, 435; variance, 438
Linear regression equation, 430; computation of, 432
Linear trend, see Straight-line trend
Line chart, 136; frequency polygon, 136; in correlation analysis, 411; ogive, 139; time series, 140
Link relatives, 119
Literary Digest election poll, 534
Logarithmic trend lines, comparison of two, 309
Logarithms, arithmetic mean of the, 174; determining the characteristic, 598; extracting roots, 601; finding the mantissa, 598; in division, 601; in multiplication, 600; interpolation, 599; raising to powers, 601; straight line fitted to, by method of least squares, 296; use of, 598
Lower control limit, 543

M

Machine methods, of recording and summarizing statistical data, 54
Machines, accounting, 64; card punch, 64; multiple-register, 56; needle-sorting, 58; posting, 57; sorting, 64
Machine tabulation of unit records, 102
Management decisions, use of statistical data in, 2
Management of a statistical office, 11
Mantissa, finding the, 598
Map, crosshatched, 151; large dot, 150; small dot, 150; statistical, 149
Market factors, 529
Market indicators, 529
Marketing research, 521; see also Advertising research; agencies and functions of, 521; comments on some statistical techniques used in, 533; consultants, 523; departments of individual concerns, 524; industrial, 527; quantitative methods of, in sales control, 528; sales analysis, 528; types of, 526
Market potential, and market research, 529
Market Research: A Guide to Information on Domestic Marketing, 71
Maximum ordinate, 210; formula for, 211
Mean, arithmetic, see Arithmetic mean; geometric, 174, 188; harmonic, 173; modified, 337; standard error of, 221–225; weighted arithmetic, 169
Measurement of cyclical fluctuations, 358
Measurement of seasonal variation, 321; evaluation of different methods of, 339
Measurement of secular trend, graphic method, 281; method of least squares, 286; method of semiaverages, 283
Measures of total economic activity, 473; bank debits and clearings, 474; employment, 475; gross national product, 477
Median, 177, 195; characteristics of, 183–188; locating, 177–179; used to average specific seasonals, 336
Minerals Yearbook, 75
Mode, 180; characteristics of, 189; crude, 181; empirical, 182; formulas for, 181, 182; graphic interpolation of, 181
Modified mean, used to measure specific seasonals, 337
Monthly Labor Review, 76, 80, 476
Moody's Manuals of Investment, 77
Motivation research, 535
Multiple-column record, 50, 98; illustrated, 51
Multiple correlation, 460; abstract measures of, 466; computation of coefficient of, 465; measurement of, 451; net correlation, 467; partial correlation, 467; regression equation, 461; reliability of estimates, 465
Multiple-register machines, 56, 101; illustrated, 57
Multiplication, logarithms in, 600

N

National income, 482; defined, 482
Needle-sorting devices, 58, 102
Needle-sorting ledger card, illustrated, 60
Net correlation, 467
Net foreign investment, described, 481; forecasting, 515
Net national product, 482
Newspaper almanacs, 78; *Texas Almanac*, 78; *World Almanac, The*, 78
Nonbasic industry, 561
Noncomparability of data, 14
Nondetermination, coefficient of, 439
Nonlinear correlation, 451; correlation ratio, 456; index of correlation, 455; measurement of, 451; regression line, 454; standard error of estimate, 454
Nonlinear relationships, 419
Nonlinear trend, 295; choice of trend line, 307; division of, into its components, 311; growth curves, 301; measurement of, 281; nature of, 276; second-degree parabola, 299; straight line fitted to logarithms by method of least squares, 296
Normal curve, 210; area under, 212; illustrated, 210
Normal distribution, 209; computing, 211; illustrated, 210, 240

O

Ogive, 139
Open-end class, 28
Ordinates, areas between maximum ordinate and ordinate at $\frac{x}{\sigma}$, 604; maximum, 210; values of, 604

P

Panel, used for sampling, 535
Parabola, second-degree, 299
Parity index, 489
Parity ratio, 489
Partial correlation, 467; coefficient of, 467
p chart, 550
Pearl-Reed curve, 303
Pegboard, 55, 101; illustrated, 55
Percentage, 107; computation of, 107; forcing to total one hundred, 113; standard deviation of a, 230; standard error of a, 228
Percentage change, 115; on semilogarithmic charts, 259, 260, 262
Percentage relatives, defined, 117
Period data, defined, 142; locating on time scale, 142
Period data charts, interpolating on, 142
Personal consumption expenditures, described, 480; forecasting, 516
Personal income, 482; defined, 483; disposable, 484
Personal saving, 484
Petroleum Facts and Figures, 78
Pictogram, 133
Pie chart, 147
Plant and equipment expenditures, as a business barometer, 488
Point data, defined, 142; locating on time scale, 142
Point data chart, interpolating on a, 142
Polygon, frequency, 137
Population, defined, 215
Population of a community, and community economic surveys, 565
Posting machine, 57, 101
Presentation, of data, 7; paragraph form of, 7; tabular form of, 7
Price index, 269
Primary data, 43
Primary source of data, 70
Primary subscript, 463
Private sources of data, 77
Probability sample, 218
Product, variability in manufactured, 538
Product-moment method, 440, 442
Projection of trends, 14
Published data, guides to, 70
Punched-card, illustrated, 61; tabulation, 60, 102
Purposive sample, 218

Q

Qualitative description, defined, 2
Quality control, 538; acceptance sampling, 554; assignable causes, 539; computing control limits from $\overline{R}$, 546; control chart for fraction defective, 550; control charts for variables, 540; defined, 538; random variations, 539; *R* chart, 548; statistical, 538; variability in the manufactured product, 538; $\overline{X}$ chart, 540
Quantitative classifications, 21
Quantitative method of expressing facts, 2
Quartile, first, 195; second, 195; third, 195
Quartile deviation, 194; characteristics of, 205; defined, 195; formula for, 195
Questionnaire, defined, 88; design, 91; illustrated, 95, 96, 97; pretesting a, 94; principles of construction of a, 92
Quota sample, 534

R

R, computing control chart limits from, 546
Random fluctuations, 248, 253
Random numbers, 235
Random sample, 218
Random variations, 539
Range, 194, 204; characteristics of, 204; interquartile, 195; semi-interquartile, 195
Ratio, correlation, 456
Ratios, 106; averaging, 169; between unlike quantities, 122; computation of, 107; determination of base for, 109; of parts to whole, 110; parity, 489
Ratio to twelve-month moving average, method of measuring seasonals, 325
Ratio to two-item average of twelve-month moving average, method of measuring seasonals, 331
R chart, 548
Readers' Guide to Periodical Literature, 71
Real income, 268
Real wages, 268
Reciprocals, defined, 111
Recording statistical data, machine methods of, 54
Records, basic types of internal, 50; multiple-column, 50; unit, 53
Regression equation, 461; computation of, 462; linear, 430
Regression line, 453
Relatives, 117; **see also** Index numbers; averages of, 386; chain, 119; computing an index from, 367; fixed-base percentage, of time series, 117; link, 119; simple, 386; unweighted arithmetic mean of, 389; unweighted geometric mean of, 391; weighted with base-year values, 395
Reliability of estimates, 434, 465
Research, **see** Advertising research, Marketing research; motivation, 535
Residuals, 434
Root-mean-square deviation, 198

S

Sales analysis, in marketing research, 528
Sample, area, 237; bias in the, 224; defined, 215; design of, 235, 238; judgment, 218; precision of an estimate from a, 216; probability, 218; purposive, 218; quota, 534; random, 218; selection of a, 217; systematic, 236
Sample design, 235, 238
Sampling, 94, 215; acceptance, 554; and marketing research, 534; cluster, 236; computation of standard deviation of sample means, 223; confidence intervals, 221; double, 554; importance of, 215; importance of size of universe, 225; improper, 16; panel, 535; sequential, 554; single, 554

Sampling error, 216, 237
Sampling plan, acceptance, 554; acceptance-rectification, 554; acceptance-rejection, 554
Scales, for semilogarithmic charts, 264; using two, 264
Scatter, 192
Scatter diagram, 411; estimates based on a, 418
Seasonal variation, 248, 251, 318; adjusting bank debits for, 362; adjusting data for, 352; adjusting monthly data for, 362, 371; averaging specific seasonals, 333; changing seasonal pattern, 340; evaluation of different methods of measuring, 339; forecasting, 502; indexes of, 318; making use of a measure of, 344; measuring, 321; measuring differences in seasonality, 347; showing the seasonal pattern, 370; specific seasonals, 324; using a measure of, in forecasting, 349
Secondary data, 43
Secondary source of data, 70
Secondary subscript, 463
Second-degree parabola, 299; computation of trend, 300; formula for, 299
Secular trend, 248, 249, 276–317; see also Straight-line trend, Nonlinear trend: adjusting annual data for, 359; adjusting monthly data for, 371; comparison of two logarithmic trend lines, 309; division of into its components, 311; forecasting, 314, 496; general discussion of, 276; making use of a measure of, 308; measuring, 281; nonlinear, see Nonlinear trend
Semiaverages, method of measuring secular trend, 283; trend values computed by method of, 285
Semi-interquartile range, 195; formula for, 195
Semilogarithmic charts, 258; percentage change on, 259, 260, 262; setting up the scale for, 264; using two scales, 264
Semilogarithmic straight line, computation of, 297
Service industry, 560; defined, 560, 561
Services in a community, and community economic surveys, 568
Shingle strip, 55; illustrated, 55
Sigma, defined, 164
Simple relatives, 386
Skewness, 193; coefficient of, 208; formulas for, 208, 209; measures of, 207
Sources of data, 71; comprehensive, 72, 79; external, 9; governmental, 73, 79; internal, 9; magazines, 82; miscellaneous, 82; newspapers, 82; primary, 70; private, 77; private organizations, 83; secondary, 70; statistical services, 82; technical periodicals, 83; trade associations, 83
Specific seasonals, 324; averaging, 333; measuring, 325-333; using arithmetic mean to average, 335; using median to average, 336; using modified mean to average, 337
Standard and Poor's Trade and Securities Statistics, 77
Standard deviation, 197; characteristics of, 206; computed from a frequency distribution, 203, 204; computed from a normal distribution, 202; computed from arbitrary origin, 201; computed from the mean, 199; computing, 223; formulas for, 198–200, 203; importance of, 202; of the means of all possible samples, 221; of the universe, see Standard deviation of the universe
Standard deviation of the universe, 223; estimating the, 224
Standard error of a percentage, 228; computing, 230; formulas for, 232
Standard error of estimate, 435, 454; computation of, 435; formulas for, 435, 436, 441, 446; shorter method of computing, 436
Standard error of the mean, 221; computing, 223; formula for, 225
Standard metropolitan area, 26
Statistical Abstract of the United States, 74
Statistical analysis, 105; purpose of, 105
Statistical data, defined, 2; machine methods of recording and summarizing, 54; presenting, 25; use of in management decisions, 2
Statistical map, 149; crosshatched, 151; large-dot, 150; small-dot, 150
Statistical methods, a tool of management, 2; basic steps in, 4; defined, 2
Statistical office, management of, 11
Statistical organizations, an agent of market research, 523
Statistical pitfalls to avoid, 12
Statistical services, 82
Statistical Services of the United States Government, 71
Statistical tables, 19
Statistical units, 3
Stock market averages, 492
Straight line, characteristics of a, 279; computation of semilogarithmic, 297; formula for a, 279
Straight-line trend, characteristics of a straight line, 279; computation of, by even number of years, 290; computation of, by odd number of years, 288; formula for a straight line, 279; measurement of, 281; nature of, 276; straight line fitted to logarithms by method of least squares, 296; use of, 292
Summarizing statistical data, machine methods of, 54
Summary board, 55; illustrated, 55
Summary strips, 101
Survey of Buying Power, 78
Survey of Current Business, 79, 476
Surveys, community economic, see Community economic surveys; consumer, 526
Systematic sample, 236

T

Table, analytical, 34, 36; construction of a, 34; cross-classification, 420; general-purpose, 34; illustration of correlation, 443; major parts of a, 37; mechanical features of a, 43; repository, 34; special-purpose, 34, 36; statistical, 19; typewritten, 45
Tabulation, of unit records by machine, 57, 102; punched-card, 60
Tally sheet, 51, 98, 420; illustrated, 52, 421

Time scales, examples of, 141; interpolating on point data and period data charts, 142; locating point data and period data on, 142
Time series, 140, 245; **see also** Cyclical fluctuations, Erratic fluctuations, Random fluctuations, Seasonal variation, Secular trend; adjustment for calendar variation, 256; cyclical fluctuations in, 356; deflation of, 267; importance of analysis, 255; in charts, 130; in tables, 23, 41; locating point data and period data on time scale, 142; on semilogarithmic charts, 258; types of fluctuation, 248
Title of a table, 37
Trade associations, an agent of market research, 522; statistical publications of, 78
Trade in a community, and community economic surveys, 568
Trend, nonlinear, **see** Nonlinear trend; secular, **see** Secular trend; straight-line, **see** Straight-line trend; uncritical projection of, 14
Trend lines, choice of, 307; comparison of two logarithmic, 309
Trend values, monthly, 375
Twelve-month moving average, ratio to, 325
Two-item average of twelve-month moving average, ratio to, 331

U

Unit records, 53, 101; machine tabulation of, 57
Universe, defined, 215; importance of size of, 225; standard deviation of, 223
Unlike quantities, ratios between, 122
Unweighted arithmetic mean of relatives, 389; formulas for, 390, 391
Unweighted geometric mean of relatives, 391; formulas for, 43, 44
Upper control limit, 543
Upward bias, 394
U. S. Census of Population, 73

V

Variability in the manufactured product, 538; **see also** Quality control; assignable variations, 539; random variations, 539
Variable, defined, 3; dependent, 423; independent, 423; obtaining values of a, 3
Variables, continuous, 29; control charts for, 540; discontinuous, 29; discrete, 29
Variance, 438
Variation, adjustment for calendar, 256; coefficient of, 207; seasonal, **see** Seasonal variation
Variations, random, 539

W

Wages, real, 268
Weighted arithmetic mean, 169; formula for, 169
Weighted arithmetic mean of relatives, 395; formulas for, 395, 396
Weighted geometric mean of relatives, 396; formulas for, 398, 399

X

$\bar{X}$ chart, 540

Y

Y axis, on charts, 145
Yearbook of Agriculture, 75