

**TEXT FLY WITHIN
THE BOOK ONLY**

UNIVERSAL
LIBRARY

OU_154614

UNIVERSAL
LIBRARY

OUP—552—7-7-66—10,000

OSMANIA UNIVERSITY LIBRARY

Call No. 519.9 Accession No. 31
Author S77E. & p. 50ml-3
Title Elementary statistics

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

ELEMENTARY STATISTICS

For Students of Social Science and Business

ELEMENTARY STATISTICS

For Students of Social Science and Business

R. CLAY SPROWLS

Assistant Professor of Business Statistics
University of California, Los Angeles

McGRAW-HILL BOOK COMPANY, INC.

New York Toronto London

1955

ELEMENTARY STATISTICS

Copyright © 1955 by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

Library of Congress Catalog Card Number 55-6867

II

THE MAPLE PRESS COMPANY, YORK, PA.

To AILEEN, ANN, AND LAWRENCE

All branches of statistics, not just acceptance inspection and sampling surveys, deal with the same basic problem, namely, the problem of decision making in the face of uncertainty. All decision rules, not just acceptance inspection plans, must be evaluated by their consequences. These consequences are expressible in terms of risks, or more intrinsically, in terms of the probabilities of taking the various permissible actions which are induced by the experiment, decision rule, and the possible states of the system. In brief . . . not facts from figures but rather decisions from observations should become the main emphasis in elementary statistical instruction.

M. A. Girshick, on page 645 of the September, 1953, issue of the *Journal of the American Statistical Association*, in a review of "Facts from Figures" by M. J. Moroney.

PREFACE

This book on statistics will deal primarily with the formulation of decisions based upon incomplete information. The book is so oriented for two reasons: (1) many statisticians of this era view the central problem of statistics as drawing conclusions from experimental data; (2) business managers must continually make decisions the consequences of which are uncertain. Since I believe the study of statistics can and should be of great value to management in this task, I have attempted to develop all the chapters on methodology around a single unifying theme, which may be stated as follows:

1. Statistics is the science of making decisions based upon incomplete knowledge.

2. Implicit in making a decision is the acceptance of one of two alternative hypotheses.

3. Associated with the acceptance of either hypothesis is the risk that the one selected is incorrect.

4. The function of statistics is to aid in the formulation of the alternative hypotheses and the evaluation of the associated risks of being wrong. This, I believe, is the main burden of statistics. It makes a course in statistics primarily a course in scientific method and secondarily a course in statistical techniques.

Any author must be selective in his choice of topics if his book is to be kept to a reasonable size. I have not attempted to compile an encyclopedia of techniques, but rather to bring together materials that will be useful in teaching the fundamental ideas of statistics according to the theme stated above. To reach this objective, I have had either to eliminate entirely or to present only briefly some of the topics traditionally included in elementary statistics. In my judgment, no great loss is thereby sustained since many of these topics, especially in books oriented toward business, are outmoded by recent advances. The real business of statistics is inference, not charting or tabling or diagramming. Basic statistical ideas have much to contribute to a general as well as to a specialized education because statistics is a way of thinking about problems. I believe that it is an intellectual discipline of which more people should be aware, and the college or university student who does not meet

these ideas in his first course in statistics may never meet them or else may meet them too late in his intellectual development.

This book has been criticized in earlier mimeographed versions for its emphasis on the controlled sampling experiment which, in the opinion of some, applies more to statistical quality control than to the majority of (business) decision problems. I do not accept this viewpoint. The controlled sampling experiment, instead of being used only in statistical quality control, is currently found applied in market research, sampling for auditing, measuring the accuracy of clerical work, and, in its broader aspects, in government sampling surveys. More widespread knowledge of the methodology would lead to even wider application. Businessmen need the best information that is available, and statistical methods can help them get it. What kind of statistics is "good" for the student in business (or in any other field) is a matter of judgment. My own opinion is that he must know above all how to think and how to solve problems. Not all problems are statistical, or even quantitative in nature, but the controlled sampling experiment is nevertheless a valuable model to study. It emphasizes probabilities and the errors in decision, both of which frequently enter into uncontrolled experiments. The controlled experiment is an ideal model for judging decision problems in general because it sets forth the right questions to ask. In this sense, a statistics book should not be a manual of techniques. The study of statistics should go beyond technical exercises to the level of a cultural and educational experience.

I have based the organization and presentation of the materials in this book on the belief that effective teaching and learning processes depend upon (1) motivation and personal involvement in the subject matter, (2) repetition, and (3) actual experience. To satisfy the first criterion, I have perhaps been somewhat unconventional in two respects. In the first place, I have adopted an informal style of writing, including extensive use of the second person, in the earlier portions of the book. My purpose in so doing is to facilitate the reception of new ideas by involving the student as deeply as possible in the learning process. I want him to feel that each exercise is his own personal problem and that he is responsible for giving an answer. It is my judgment that the effectiveness of this informality as a teaching device far outweighs the prejudices of stylistic purists against the use of the second person. In the second place, I have deliberately selected gambling on horse races as the first problem in the book. I choose this problem because it is a motivating one. Most students approach the study of statistics for the first time with disinterest, or even apprehension; what could be more dull or dry than statistics? The horse-racing problem has the stimulating effect of shock. Taken by surprise, students are motivated to express their own ideas.

The value of this problem in arousing the student's interest in statistics cancels out any personal objections I might have to the subject matter. I can only hope that others will find it as effective as I have.

The need for repetition accounts for the unusual arrangement of materials in some portions of the book and for continual reference back and forth to sections with the same topics. Ideas and concepts are often introduced earlier than necessary so that they will be familiar to the student when they are developed in detail. Thus, the summation sign and the standard deviation are introduced in Chap. 2, together with general notes on symbols and calculation, although the actual computation of the standard deviation is not presented until Chap. 7. The basic ideas in decision making are repeated frequently so that they will be a part of the student's very existence by the end of the course.

Actual experience is a most necessary part of the learning process. To this end, the book is built around four fundamental exercises: (1) an experiment in sampling a bead box (Sec. 1.2); (2) a random sampling experiment (Sec. 1.7); (3) a personal decision problem (Exercise 3.14); (4) a term project—a small research problem (Exercise 6.9). Continual reference is made to the two experiments which are an integral part of the course, and data are provided for the individual reader who lacks the equipment to do these experiments. The value of the personal decision problem lies in its ability to bring home to the student the basic methods of statistics and their relationship to everyday decisions, even though such decisions may not require the use of statistical data. Although most students will never be statisticians or conduct statistical research, they learn more about the use of statistics by conducting a small research project of their own. In this process they must make all the minor decisions which are so important in actual research but which are so often unnoticed by those who use statistics for lack of firsthand experience. Students thus realize the complexity of apparently simple problems. At approximately the middle of the course, I have the students read Chap. 6 without benefit of lectures, and then have them start a term project on testing a hypothesis about or making an estimate of an unknown percentage.

These four exercises are supplemented by other exercises and problems in all sections of the book. Although practice on these is initially very essential, the point where practice no longer contributes to principles is soon reached. I have therefore included only enough exercises to suggest applications and to provide basic training; the individual instructor may wish to add others. Computations are kept at a minimum for courses without a laboratory or such computing equipment as desk calculators or adding machines. Optional sections exploring certain subjects more deeply than seems necessary to some instructors may be omitted without

PREFACE

destroying the continuity of thought. Except in the chapters on time series and index numbers, the business and economic aspects of statistics are mostly relegated to the problems. Statistics is not a discipline uniquely tied to any particular specialized field of knowledge. I believe, if other names were substituted for the variables, this book would be easily adapted to another field.

This book is intended both to simplify the statistics and to raise the standards of student instruction. I believe that there has lately been too much writing down to the level of the student. Colleges and universities should seek to raise the student's educational level instead of trying to adapt courses to his everyday English. Since business school students (and perhaps those in the social sciences in general) are notoriously lacking in mathematical experience of the most elementary kind, this text includes technical statistical terms and symbols, and attempts to teach some abstract reasoning. These should enhance the student's ability to analyze and think clearly and should help him to avoid fuzzy thoughts and statements. The student who finishes this course should be no longer a statistical layman, but should have a mature understanding of statistical principles. I am firmly convinced that these materials can be effectively presented even to students who dislike and resist any subject resembling mathematics.

This book is primarily written, not for those who intend to become practicing statisticians, but for ordinary citizens who will be exposed to statistical statements, arguments, and reports. The student is trained to evaluate these reports and to ask the questions that will prevent his being misled. He is taught to ask questions about the hypotheses, the random sample, the alpha or the beta, and the operating characteristics. He knows that the researcher who cannot answer these questions may not have done a thorough or an adequate statistical job. If a parallel may be drawn, the objective here is to give students a course in first aid or mental hygiene rather than to train them to be doctors.

TO THE STUDENT

Like other courses dealing with some form of mathematical knowledge, the ideas and methods of statistics are sometimes difficult for students in the beginning. A word of advice may be helpful. To understand the materials, you must sit down at a desk with pencil and paper, work through problems, write out definitions, and put ideas into your own words. This is a tedious task, but I think it is the only successful method of learning the course in statistics. Now a word of caution. The course develops in a logical way, building upon foundations laid in the first few sections. Each topic adds to what has gone before. You must start the

learning process at the beginning of the book. You cannot wait until the middle or the end of the course to read the textbook to "get caught up." As the course develops, you will learn a body of statistical techniques for handling data so that valid conclusions may be drawn from them. These techniques are important, for they are the tools which (with some imagination) you will be able to apply both in your other courses in school and later on in your work. Equally important, you will learn how to think about decision problems—how to analyze a problem, how to weigh alternatives, how to adopt consistent rules of behavior. This knowledge can be applied to everyday problems throughout the rest of your life.

ACKNOWLEDGMENTS

I wish to express my appreciation to Prof. Armen Alchian who has read the entire manuscript and has made valuable contributions to the organization and content of the book; to Anne Cook for typing the manuscript; to Grace Stimson for reading and editing the manuscript; to numerous students for helpful criticisms and suggestions while using the text in preliminary drafts; and to reviews of unknown origin from several publishers.

I am indebted to Prof. Sir Ronald A. Fisher, Cambridge, to Dr. Frank Yates, Rothamsted, and to Messrs. Oliver and Boyd, Ltd., Edinburgh, for permission to reprint Tables A.4, A.9, and A.10, from their book "Statistical Tables for Biological, Agricultural, and Medical Research." The generosity of other authors and publishers in supplying materials is acknowledged where they appear.

R. CLAY SPROWLS

CONTENTS

<i>Preface</i>	vii
1. Statistics—The Science of Making Decisions from Observations	1
2. Tools—Notation, Probability, and the Binomial Distribution	25
3. Statistical Inference in the Discrete Case: Binomial Distribution (I)	51
4. Statistical Inference in the Discrete Case: Binomial Distribution (II)	73
5. Statistical Inference: Estimation of a Proportion	98
6. Statistical-research Methodology	108
7. Tools—Frequency Distributions, Averages, and Measures of Dispersion	133
8. Statistical Inference: Frequency Distributions	167
9. Statistical Inference: The Arithmetic Mean and Related Problems	183
10. Simple Linear Regression and Correlation	222
11. The Analysis of Time Series: Examples, Point of View, and Tools	262
12. The Analysis of Time Series: Seasonal Variation	289
13. The Analysis of Time Series: Trends and Cycles	306
14. Index Numbers	332
15. Miscellaneous Topics	357
 <i>Appendix A: Statistical Tables</i>	 369
B: Glossary of Symbols	383
 <i>Index</i>	 387

CHAPTER 1

STATISTICS—THE SCIENCE OF MAKING DECISIONS FROM OBSERVATIONS

This study of statistics will begin with consideration of several typical, but necessarily oversimplified, decision problems. As the course proceeds, more examples will be given from each of several fields. In thinking about each of these problems, try to put yourself in the position of the man in the organization who has to make the decision and to come to a definite conclusion.

1.1 Simple Decision Problems

Example 1.1. This first problem is not a usual business problem and the subject matter may not be within the realm of your knowledge, but it is included because past experience has shown that it is motivating.

A friend (?) of yours has a “system” for picking winners in horse races. He claims, in fact, that his system will pick winners 50 per cent of the time during several weeks of racing. He argues to your satisfaction that with this high a percentage of winners, the system will almost surely win money in the long run. Since your friend has no money to finance bets at the local track or bookmaker, he is trying to interest you in supplying the necessary capital for a venture in which the system is applied to a group of races.

You are a bit skeptical of his claim and decide to test it on a few races before you give him any money. You select a random sample of 10 races from all the races to be run during the next week at the several tracks operating currently. You tell your friend to apply his system to each of these 10 races, to tell you his selection in each race, and to check with you on the outcome.

Faced with a real-life decision problem of this kind, you may take one of two courses of action. You may wait until you observe how many of the 10 races he wins and then make your decision by judging these results or you may plan in advance how many races he must win before you will invest your money. Both of these procedures are used in everyday decision problems. At times it is better to wait and see what happens before making a decision; at other times it pays to plan your behavior beforehand. Some individuals who have come into contact with the field of

statistics believe that the statistician is primarily a judge; i.e., he is concerned with drawing valid conclusions from experimental data or with making the proper decision from observed facts. Most statisticians believe that experiments should be planned so that the data are gathered with the method of analysis incorporated into the experimental procedure. *They view the proper function of statistics as the relating of conclusions or decisions to the possible outcomes of experiments before the empirical data have been gathered.* This latter view is the one which will be adopted in the first part of this course.

EXERCISE

1.1. *a.* Suppose you decide to act like a judge in this problem. Will you personally invest your money if 6 or more races are won? 5? 4? 3? 2? 1? 0?

b. Suppose you decide to act like a planner in this problem. What number of races must be won before you will invest in the scheme? 6 or more? 5? 4? 3? 2? 1? 0?

c. Are your answers in (a) and (b) the same? That is to say, will the same number of wins lead you to invest in the scheme whether you are acting as a judge or a planner?

d. Do you think that your answers in (a) and (b) should be the same?

e. Suppose you are to face this same decision problem every day for 30 days. Which role do you prefer to adopt: that of the judge or the planner?

Remark 1. Many students refuse to reach a decision in this problem because there are only 10 races in the sample. They claim they need more races before they can decide. It may be that 100 races are more helpful than 16, but since you are allowed only 10, you must reach a decision based upon this sample. The number of observations is not crucial here.

Remark 2. Some students will want to confuse the issue still further by trying to anticipate the odds. If the average odds are, say, 5 to 1, they will invest their money if fewer horses win than are necessary for investment at odds of 2 to 1 or even money. Their decision rule varies with the anticipated odds. The odds are certainly not irrelevant to the decision, nor would the author lead you to believe that they are. For the moment, the odds are not crucial; the problem is to decide if the claim (50 per cent winners) is true.

Remark 3. The statement that the system picks winners 50 per cent of the time does not guarantee that there will be 5 winners in your selection of 10 races. It does not guarantee that you will make a profit on any set of 10 races on which you may happen to bet. You may not be able to see the reason for these two statements now, but you should be able to do so shortly.

Remark 4. Some students will say that they lack sufficient experience to make a decision in this problem. The author agrees, but he also thinks it is good practice to make such a decision. This course is not intended to teach you how to bet on horse races; it is intended to give you experience in solving similar decision problems.

Example 1.2. A small electronic-equipment plant manufactures, among other things, a special tube for television sets. At the end of the fiscal year, the auditor must value the inventory of tubes which are on hand. He realizes that some of

these tubes may be unsalable because they are broken or otherwise defective, and he desires to value the inventory only on the basis of tubes in good condition. The manager of the plant tells the auditor that, from past experience, he believes 20 per cent of the tubes will be broken or bad and accordingly suggests that the inventory be valued at 80 per cent of full value. The auditor is skeptical of this claim. He decides to test it by selecting 10 tubes at random from the inventory, checking each one for defects, and counting the number which are unsalable.

EXERCISE

1.2. a. Suppose that you are the auditor and that you are acting like a judge. Will you decide that the inventory is 20 per cent defective if 5 tubes are found to be defective? 4? 3? 2? 1? 0?

b. Suppose you are a planner. What number of defective tubes will you allow and still accept the inventory as 20 per cent defective?

Remark. Past experience with statistics classes has shown that the numbers in the following examples are selected more often by students than any other numbers:

Example 1.1 5

Example 1.2 2

In other words, most students will invest in the horse-racing scheme if 5 or more races are won; they will accept the inventory as 20 per cent defective if 2 or fewer tubes are defective.

You can see how these numbers are chosen. Example 1.1: There are 10 races in the sample, and the claim is made that 50 per cent of the selections will win if the system is followed. $10 \cdot 50\% = 5$. Example 1.2: There are 10 electron tubes in the sample, and the manager claims that 20 per cent are defective. $10 \cdot 20\% = 2$.

Example 1.3. Suppose that you, as the manager of the china department of a large department store, are responsible for the ordering of inventory stock. Two different patterns of china, pattern *A* and pattern *B*, are ordered from the same manufacturer. The two patterns are ordered in equal quantities on the assumption that they sell equally well. From time to time a random sample of 256 sales from the total actual sales of the two patterns is examined, usually just before more sets are ordered for stock.

EXERCISE

1.3. a. Will you accept the assumption that 50 per cent of all the sales are pattern *A* if of the 256 sales 129 are pattern *A*? 130? 131? 132? Will you accept the assumption if 127 are pattern *A*? 126? 125? 124?

b. What number of pattern *A* sales will you allow and still agree that the total sales of pattern *A* are not greater than those of pattern *B*? What number of pattern *A* sales will you allow and still agree that the total sales of pattern *A* are not less than those of pattern *B*?

There are two answers to Exercise 1.3 because the sales of pattern *A* may be more or less than the sales of pattern *B*. Answers consistent

with the analysis given by most students for Examples 1.1 and 1.2 are as follows:

Sales of A Exceed Sales of B. There are 256 sales and the standard is 50 per cent pattern A. Since 50 per cent of 256 is 128, there must be 128 or fewer A sales to conclude that sales of A do not exceed sales of B.

Sales of A Are Less than Sales of B. There are 256 sales and the standard is 50 per cent pattern A. Since 50 per cent of 256 is 128, there must be 128 or more sales of pattern A to conclude that sales of A are not less than sales of B.

The usual analysis leads to the same number for both sales situations. This does not make sense. Any number of pattern A sales other than 128 will lead to a rejection of the claim that the sales are equally divided. The method of reasoning used in Examples 1.1 and 1.2 is not applicable in Example 1.3. (Do you expect the number of pattern A sales to be exactly 128?) It may be that a new method is needed for this problem, or that the analysis of the previous examples is incorrect.

1.2 Experiment in the Operation of a Statistical Decision Rule

It is not possible in this course to gain firsthand knowledge of the operation of the decision-making process in any one of these three problems. It is possible, however, to undertake an experiment which simulates the actual conditions of, say, the horse-racing example.

The laboratory is equipped with a sampling box which may be filled with beads of different colors in varying proportions.¹ For the present experiment it will be filled with 500 beads, some red and some white. These 500 beads will represent all the horse races at the currently operating tracks. Each red bead will designate a race that is won, each white bead a race that is lost. The percentage of red beads in the box corresponds to the percentage of winners which the friend's system will select. His claim is that this percentage is 50, but it is unknown in the real-life situation and it will remain unknown to you in this experiment.

A good approximation to a random sample of 10 beads is obtained by thoroughly mixing the beads in the box, tilting it, and allowing 10 beads to run out into the wooden tube. On the basis of the sample results, you must decide what to do: to invest or not to invest in the horse-racing scheme. Possible decisions and actions are set forth in the table shown on page 5. The rule to follow in making the decision is based upon the most common answer given by students in Example 1.1.

Statistical-decision Rule. Accept the box as filled with 50 per cent

¹ The author uses a sampling box described in the following article: Theodore H. Brown and Dickson H. Leavens, A Sampling Device, *American Statistician*, vol. 5, p. 16, 1951. Any large box of beads and a paddle with holes into which a given number of beads may roll are an acceptable substitute for this sampling box.

Bead-box decision	Horse-race decision	Action
Accept the box as containing 50 per cent red beads	Accept the system as selecting 50 per cent winners	Invest in the horse-racing scheme
Reject the box as containing 50 per cent red beads	Reject the system as selecting 50 per cent winners	Do not invest in the horse-racing scheme

red beads (invest in the system) if 5 or more red beads occur in the sample; otherwise, reject the box as filled with 50 per cent red beads (do not invest in the system).

EXERCISES

1.4. *a.* Take a sample from the bead box, and decide whether to invest in accordance with the above rule. Repeat the process until a total of 100 samples has been drawn by the whole class. [Readers without access to a sampling box should consult part (e) below.]

b. Record on a worksheet the number of red beads in each sample and the corresponding decision, plus those of other members of the class:

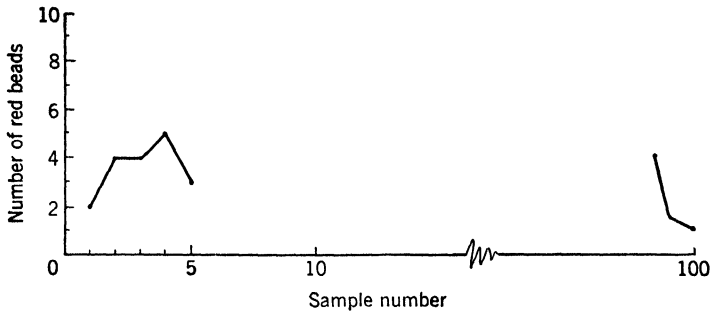
WORK SHEET

Sample number	Number of red beads	Decision
1	5	Invest
2	3	Not invest
3	6	Invest
.	.	.
.	.	.
100	8	Invest

Be sure to record the samples *in the order taken*.

c. Make a distribution of the number of samples which contain different numbers of red beads.

<i>Number of red beads in sample of 10</i>	<i>Number of samples</i>
0	1
1	4
2	8
.	.
.	.
.	.
10	0
Total.....	 100



d. Plot the number of red beads in the samples in the order of their appearance

e. Readers without access to a sampling box may use the data given in Table 1.1, obtained by the author from 100 trials with the apparatus described above.

1.5. a. How many of the total number of decisions in the whole class are to invest? How many are not to invest?

b. Suppose the rule is changed so that the decision is to invest if 4 or more red beads are observed in the sample. Now how many decisions are there to invest?

c. Repeat questions (a) and (b), using a rule with 3 or more; 2 or more; 1 or more; 0 or more.

d. Why would you not adopt this rule: Invest in the system if 0 or more red beads occur in the sample?

1.3 Elements Common to the Decision Problems

It is well to point out now certain elements which are common to each of the Examples 1.1 to 1.3. In each problem there is a statement or a claim to be tested. This claim will be called a *hypothesis*. These hypotheses are listed after the example numbers.

Example 1.1: The system will pick 50 per cent winners in horse races.

Example 1.2: Twenty per cent or more of all tubes in the inventory are defective.

Example 1.3: Fifty per cent of the sales in the china department are sales of pattern A.

In each problem a *sample* of objects—races, electron tubes, sales—has been chosen from a larger group of these objects. Moreover, if you will read the problem statements carefully, you will notice that in each case the sample was said to be chosen *randomly*. Randomness applies to the method of selecting the units which make up the sample from the larger group of units. It is a method of sampling which will be explored in greater detail in the future. In each problem a number (called a *statistic*) was computed from the sample observations. These numbers were the number of races which were won (Example 1.1); the number of electron

TABLE 1.1. OUTCOMES OF 100 TRIALS OF THE BEAD-BOX EXPERIMENT

Sample	Number of red beads	Sample	Number of red beads	Sample	Number of red beads
1	2	35	2	69	1
2	4	36	2	70	4
3	4	37	2	71	1
4	5	38	2	72	3
5	4	39	1	73	1
6	3	40	1	74	1
7	3	41	2	75	2
8	2	42	3	76	2
9	1	43	5	77	2
10	1	44	6	78	1
11	2	45	3	79	3
12	4	46	2	80	1
13	3	47	2	81	3
14	3	48	1	82	3
15	0	49	1	83	2
16	1	50	2	84	1
17	2	51	3	85	1
18	2	52	0	86	4
19	0	53	4	87	3
20	1	54	2	88	4
21	0	55	2	89	3
22	1	56	3	90	4
23	2	57	2	91	3
24	1	58	2	92	3
25	2	59	1	93	1
26	0	60	1	94	2
27	1	61	3	95	1
28	1	62	0	96	1
29	0	63	2	97	3
30	2	64	2	98	4
31	0	65	0	99	2
32	4	66	1	100	1
33	2	67	1		
34	2	68	2		

tubes which were defective (Example 1.2); the number of sales which were of pattern *A* (Example 1.3).

You were asked to reach a decision in two ways:

1. *Judge*: You could judge the sample results (the statistic) and decide what to do. For example, you observe that 6 races are won. You judge

this to be a sufficient number of winners for you to invest and you therefore invest in the scheme. Again, you observe only 2 winners; you judge this an insufficient number of winners for you to invest and you do not invest in the scheme.

2. *Planner*: You plan in advance what the sample results must be (i.e., a value of the statistic) for you to reach a decision one way or the other. For example, you decide that 5 races must be won before you will invest in the scheme. Therefore, you make a rule which says that if 5 or more races are won, you will invest; if 4 or fewer races are won, you will not invest. If you observe 6 winners in the sample of 10 races, you automatically invest; if you observe only 2 winners, you automatically refuse to invest. Acting as a planner, the problem is to decide what number will be used in the decision-making rule. No definite numbers were given in the statements of the problems; you were left to decide what you would do in each case.

Finally, the decision which was made in each problem could be wrong. You could refuse to put your money into the horse-racing system when, in fact, you should; i.e., when the system actually picks 50 per cent winners. You could reject the hypothesis that the inventory was 20 per cent defective when, in fact, 20 per cent of the electron tubes are defective and unsalable. You could reject the hypothesis that 50 per cent of the sales are of pattern *A* when, in fact, 50 per cent of the sales are of pattern *A*. The decisions are uncertain because they are based on samples from a larger group of observations and not upon all the units in the group.

1.4 Some Definitions

Some of the notions about sampling previously alluded to will be elaborated upon and definitions for some of the important concepts will be given in this section. Key words and phrases are in italics and you should be sure that you understand their meaning.

Statistical analysis studies the numerical characteristics of what will be called *elementary units*. Examples of elementary units are:

1. A family
2. A household
3. An individual
4. An industrial product
5. A horse race (Example 1.1)
6. An electron tube (Example 1.2)
7. A sale (Example 1.3)

Each of these elementary units has one or more characteristics which it is desirable to study. One characteristic of each of the above elementary units is:

1. The number of children under sixteen years of age
2. The number of wage earners
3. Net taxable income
4. Length of the product
5. Winner or loser (Example 1.1)
6. Defective or good (Example 1.2)
7. Pattern *A* or pattern *B* (Example 1.3)

These characteristics may be expressed in numbers by one of the two usual methods of generating numbers in statistics:

1. Measuring
2. Categorizing

Each elementary unit will be assigned a numerical value by one of these processes of generating numbers. These values will not be the same (at least not in any interesting situation), but will vary from one unit to the next. The numerical values of the characteristics will be called *observations*. An observation of each of the characteristics listed above could be as follows:

1. 4 (children)
2. 2 (wage earners)
3. 2500 (dollars)
4. 50 (feet in length)
5. Score 1 for a winner; 0 for a loser
6. Score 1 for a defective tube; 0 for a good tube
7. Score 1 for pattern *A*; 0 for pattern *B*

The collection of all the observations of some characteristic is called a *universe* or a *population*. (These words will be used interchangeably.) The universe is an abstraction which defines all the possible observations (i.e., the numerical values of the elementary units), although the universe is often thought of and is, in fact, loosely spoken of as the aggregate of the elementary units. This may lead to confusion, since the same elementary unit can be the source of several populations. This happens when (1) different characteristics are defined for the same elementary unit, or (2) even with the same variable, the method of measurement yields several populations. An example of (1) is the population of the length of life (in hours) of a certain group of electron tubes and the population of the diameters (in inches) of the same tubes. As an example of (2), suppose the elementary unit is an individual and the variable is height. A universe may be defined by the observations in each of the following experiments:

1. The repeated measurement of height by one person using only one ruler
2. The repeated measurement of height by one person using several rulers
3. The repeated measurement of height by several persons using one ruler

4. The repeated measurement of height by several persons using several rulers

This illustrates the care with which the observations that comprise the population must be defined and also the necessity for specifying the procedure to be used in measuring the numerical values of a characteristic.

In common usage, the word "population" is associated with collections of human beings. But elementary units of statistical populations need not be human; in fact, when human beings are the elementary units, the term "human population" is very often used. Thus the problems of sampling human populations come up in public opinion polling, consumer surveys, etc.

The aim of the statistical methods developed in this course will be to study a population by analyzing a well-selected portion of that population. Such a portion is called a *sample*. Samples are studied for several reasons:

1. A population may be so large that it is impossible or impractical (too costly) to study the aggregate of the observations.

2. The characteristic may be such that the elementary unit is completely or partially destroyed when the observation is determined.

3. Even in the absence of (1) or (2), a well-chosen sample can give the necessary information about the population at a cost far less than that incurred by a study of the whole universe.

4. Samples sometimes yield more accurate and more reliable measures than whole populations.

1.5 The Method of Selecting the Sample

If, as stated in Sec. 1.4, a population is to be studied by the analysis of a sample from that population, the question of how to select the sample naturally arises. How should the 10 races or 10 electron tubes or 256 china sales in the sample be selected? Up to this point the question has been answered merely by saying that the sample should be chosen randomly, but the meaning of randomness must be examined in detail.

The concept of randomness is an important one in statistics. It implies haphazardness, but a haphazard selection is not often a random one. You may pick a series of numbers as they come to you—a haphazard selection. Experience has shown that this is not a random selection because an unconscious bias in favor of certain numbers makes them more likely to be chosen. Such a bias may favor large numbers, small numbers, even numbers, odd numbers, etc. Random samples are haphazard in the sense of lack of predictability. Randomness precludes the ability to predict, for example, a new-born infant's sex. An objective definition of randomness has not been given. You will be expected to

get an intuitive feeling for it and to learn how to draw random samples in practice.

A sample is said to be random when every elementary unit (and, therefore, every observation) has a definite and known probability of being drawn into the sample. A special kind of probability sample is one where the known probabilities are all equal. Each observation then has the same probability of being selected for the sample as every other observation. The selection of a probability sample is an automatic statistical procedure in which neither the sampler nor the sampling unit has any personal choice as to which unit comprises the sample.

The idea of a random sample is fairly simple in principle. The real problem is a practical one: how to select a random sample or how to ensure that the probabilities of the sampling units being drawn will be known and, in a simple case, will all be equal. Ideally, one needs a machine which selects the sample items with equal probability and without knowledge of the population or the variable. Such a device for practical use is a table of random numbers. One page of a larger table of random numbers is reproduced as Table 1.2.

Remark. Several tables of random numbers are available. One obtainable without charge, if requested on letterhead stationery, is the "Table of 105,000 Random Decimal Digits," Statement 4914, File No. 261-A-1, Bureau of Transport and Economic Statistics, Interstate Commerce Commission, Washington, D.C.

1.6 Example of the Use of the Table of Random Numbers

Imagine a large fish bowl filled with 998 poker chips or bingo balls. Each chip carries a serial number, 001, 002, 003, . . . , 997, 998. On the back of each chip is written another number, representing the numerical value of some characteristic called X . One particular chip might look like this:

075

Front

1.6

Back

Since a real fish bowl and the 998 poker chips are not available, Table 1.3 presents in tabular form a distribution of the contents of one such fish bowl, which may be described as follows: There is one chip numbered 001 with value of $X = 0.0$. There is one chip numbered 002 with value of $X = 0.1$. There are two chips which carry X values equal to 0.5; one is numbered 006 and one is numbered 007. There are seven chips which carry the X value 1.1; they are numbered 026 to 032. The rest of the distribution follows according to the same scheme of numbering.

The 998 X values comprise the population for this sampling study. The elementary unit is a chip; an observation is the X value on each

TABLE 1.2. RANDOM NUMBERS

Col. Line	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
351	06433	80674	24520	18222	10610	05794	37515	48619	62866	33963	14045	79451	04934	45576
352	39298	47829	72648	37414	75755	04717	29899	78817	03509	78673	73181	29973	18664	04555
353	89884	59651	67533	68123	17730	95862	08034	19473	63971	37271	31445	49019	49405	46925
354	61512	32155	51906	61662	64130	16888	37275	51262	11569	08697	91120	64156	40365	74297
355	99653	47635	12506	88535	36553	23757	34209	55803	96275	26130	47949	14877	69594	83041
356	95013	11085	13772	76638	48423	25018	99041	77529	81360	18180	97421	55541	90275	18213
357	55864	44004	13122	44115	01601	50541	00147	77685	58788	33016	61173	93049	04694	43534
358	35334	82410	91601	40617	72876	33967	73830	15405	96554	88265	34537	38526	67924	40474
359	57729	88646	76487	11622	96297	24160	09903	14047	22917	60718	66487	46346	30949	03173
360	86648	89317	63677	70119	94739	25875	38829	68377	43918	77653	04127	69930	43283	35766
361	30574	06039	07967	32422	76791	39725	53711	93385	13421	67957	20384	58731	53396	59723
362	81307	13114	83580	79974	45929	85113	72268	09858	52104	32014	53115	03727	98624	84616
363	02410	96385	79007	54939	21410	86980	91772	93307	34116	49516	42148	57740	31198	70336
364	18969	87444	52233	62319	08598	09066	95288	04794	01534	92058	03157	91758	80611	45357
365	87863	80514	66860	62297	80198	19347	73234	86265	49096	97021	92582	61422	75890	86442
366	68397	10538	15438	62311	72844	60203	46412	65943	79232	45702	67055	39024	57383	44424
367	28529	45247	58729	10854	99058	18260	38765	90038	94209	04055	27393	61517	23002	96560
368	44285	09452	15867	70418	57012	72122	36634	97283	95943	78363	36498	40662	94188	18202
369	86299	22510	33571	23309	57040	29285	67870	21913	72958	75637	99936	58715	07943	23748
370	84842	03748	90894	61658	15001	94055	36308	41161	37341	81838	19389	80336	46346	91895
371	56970	10799	52098	04184	54967	72938	56834	23777	98392	31417	98547	92058	02277	50315
372	83125	85077	60490	44369	66130	72936	69848	59773	08144	61070	73094	27059	69181	55623
373	55803	21383	02464	26141	68779	66388	75242	82690	74099	77885	23813	10054	11900	44653
374	47019	06683	33203	29608	54553	25971	69573	83854	24715	48866	65745	31131	47636	45137
375	84828	61152	79526	29554	84580	37859	28504	61980	34997	41825	11623	07320	15003	56774

376	68921	31331	79227	05748	51276	57143	31926	99915	45821	97702	87125	44488	77613	56823
377	36458	28285	30424	98420	72925	40729	22337	48293	86847	43186	42951	37804	85129	28993
378	93752	96065	36847	87729	81679	59126	59437	33225	31280	41232	34750	91097	60752	69783
379	26768	02513	58454	56958	20575	76746	49878	08846	32828	24425	30249	78801	26977	92074
380	42613	72456	43636	58085	06766	60227	96414	32671	45587	79620	84831	38156	74211	82752
381	95457	12176	65482	25596	02678	54592	63607	82096	21913	75544	55228	89796	05694	91552
382	95276	67524	63564	95958	39750	64379	46059	51666	10433	10945	55306	78562	89630	41230
383	66954	53574	64776	92345	95110	59448	77249	54044	67942	24145	42294	27427	84875	37022
384	17457	44151	14113	62462	02798	54977	48349	66738	60184	75679	38120	17640	36242	99357
385	03704	23322	83214	59337	01695	60666	97416	55064	17427	89180	74018	44865	53197	74810
386	21538	16997	33210	60337	27976	70661	08250	69599	60264	84549	78007	88450	06488	72274
387	57178	16739	98310	70348	11317	71623	55510	64756	87759	92354	78694	63638	80939	98644
388	31048	40058	94953	55866	96283	46620	52087	80817	74533	68407	55862	32476	19326	95558
389	69799	83300	16498	80733	96422	58078	99643	39847	96884	84657	33697	39578	90197	80532
390	90595	65017	59231	17772	67831	33317	00320	90401	41700	95510	61166	33757	23279	85523
391	33570	34761	98939	78784	09977	29398	93896	78227	90110	81378	96639	37008	04050	04228
392	15340	82760	57477	13898	48431	72936	78160	87240	52716	87697	79433	16336	52862	69149
393	64079	07733	36512	56186	99098	48850	72527	08486	10951	26832	39763	02485	71688	90936
394	63491	84886	67118	62063	74958	20946	28147	39338	32169	03713	93510	61244	73774	01245
395	92003	76568	41034	28260	79708	00770	88643	21188	01850	69689	49426	49128	14660	14143
396	52360	46658	66511	04172	73085	11795	52594	13287	82531	04388	64693	11934	35051	68576
397	74622	12142	68355	65635	21828	39536	18988	53609	04001	19648	14053	49623	10840	31915
398	04157	50079	61343	64315	70836	82857	35335	87900	36194	31567	53506	34304	39910	79630
399	86003	60070	56241	32836	27573	11479	94114	81641	00496	36058	75899	46620	70024	88753
400	41268	80187	20351	09636	84668	42486	71303	19512	50277	71508	20116	79520	06269	74173

SOURCE: Table of 105,000 Random Decimal Digits, p. 8, Interstate Commerce Commission, Washington, 1949.

chip. The problem of taking a random sample of 10 observations from this population is that of selecting 10 chips from this table (the counterpart of the bowl) so that each chip has the same probability of being selected, in this case $1/998$ since there are 998 chips. Each chip is numbered with a three-digit number for purposes of identification. The table of random numbers contains digits arranged randomly and with approximately equal frequencies. Reading a three-digit number from the random-number table will identify a particular chip and, therefore, its X value.

The physical process of choosing 10 numbers from the random-number table may be carried out in several ways. For example, the Interstate Commerce Commission (ICC) table consists of 30 pages, each page having 14 columns and 50 rows of 5-digit numbers. One acceptable method of selecting a random number is to select one page of the table, select a particular spot on the page, and use this spot as a starting point. Then, 10 numbers may be read from this starting point, up, down, across, diagonally, etc. These 10 numbers will identify the 10 chips and their X values for the sample.

Example 1.4. Suppose the starting page were page 8. A part of this table is reproduced in Table 1.2. The column and line selected to start reading the random numbers may be column 6, line 356. The three random digits of lines 356 to 365 and the X values for which these random numbers are the serial numbers are:

Chip	X value
250	2 3
505	3.0
339	2.6
241	2 3
258	2 4
397	2.7
851	4.0
869	4 1
090	1.7
193	2 1

Now, instead of the chips and their numbers, only the 10 values of the variable X are of interest.

Remark 1. If either of the numbers 000 or 999 is read, it is discarded and another observation taken since none of the chips in the bowl have those numbers. If the same number comes up twice, that observation is recorded twice. This is called *sampling with replacement*; a chip which has just been drawn is thrown back into the bowl and a thorough mixing takes place before the next chip is drawn.

Remark 2. A more refined method of drawing random numbers is sometimes used. The random starting point is determined by first locating a number in the table which will identify the page of the starting point, then a number for the column, and finally, a number for the line.

TABLE 1.3. DISTRIBUTION OF 998 NUMBERS FOR SAMPLING EXPERIMENT

X	Fre- quency of X	Identi- fication number	X	Fre- quency of X	Identi- fication number	X	Fre- quency of X	Identi- fication number
0 0	1	001	2 0	24	147-170	4 0	24	829-852
0 1	1	002	2 1	27	171-197	4 1	22	853-874
0 2	1	003	2 2	29	198-226	4 2	19	875-893
0 3	1	004	2 3	31	227-257	4 3	17	894-910
0 4	1	005	2 4	33	258-290	4 4	15	911-925
0 5	2	006-007	2 5	35	291-325	4 5	13	926-938
0 6	2	008-009	2 6	37	326-362	4 6	11	939-949
0 7	3	010-012	2 7	38	363-400	4 7	9	950-958
0 8	4	013 016	2 8	39	401-439	4 8	8	959-966
0 9	4	017-020	2 9	40	440-479	4 9	7	967-973
1 0	5	021-025	3 0	40	480-519	5 0	5	974-978
1 1	7	026 032	3 1	40	520-559	5 1	4	979-982
1 2	8	033-040	3 2	39	560-598	5 2	4	983-986
1 3	9	041-049	3 3	38	599-636	5 3	3	987-989
1 4	11	050-060	3 4	37	637-673	5 4	2	990-991
1 5	13	061-073	3 5	35	674-708	5 5	2	992-993
1 6	15	074-088	3 6	33	709-741	5 6	1	994
1 7	17	089-105	3 7	31	742-772	5 7	1	995
1 8	19	106-124	3 8	29	773-801	5 8	1	996
1 9	22	125-146	3 9	27	802-828	5 9	1	997
						6 0	1	998

Remark 3. Many students are confused about the applicability of "equal probability" or "equal chance" in this sampling experiment. The source of confusion is this: Values of X are not equally frequent. There are more values of $X = 3.0$ than of $X = 0.0$, for example. How can it be said that there is an equal chance of selection? The equal chance does not apply to the X values. Certainly some values of X are more likely than others because the population was constructed that way. The "equal chance" refers to the elementary unit which is 1 chip in the bowl. Each of the 40 chips which has the value of $X = 3.0$ has the same chance of selection as the 1 chip with $X = 0.0$. There are equal probabilities of selecting a chip; there are unequal probabilities of selecting X values because several chips carry the same X value.

If the table of random numbers is to be used to draw observations with equal probabilities, the arrangement of the numbers in the table must be random and the digits must appear in equal frequencies. The mechanical device used to generate the table ensures that these conditions are approximately met. Tests for randomness and equality of digit frequencies can be and are applied to the arrangement of digits in the table. Not only must the single digits 0, 1, . . . , 9 occur in equal frequencies,

but pairs (00, 01, . . . , 99), triples, etc., must be about equally frequent when read down, up, across, or diagonally in the table. The digits following a particular number (say, 2, or 3) must also be equally frequent. These criteria are never exactly satisfied in practice, but they are approximated well enough for practical use of a table of random numbers in selecting random samples.

To be adaptable to the use of a table of random numbers, the elementary units of the population must be numbered (refer to the population of 998 X values in Table 1.2) for purposes of identification. The table can be used, then, only with populations which have a finite number of sampling units which can be enumerated or listed for identification. Such a listing is called a *frame* by some statisticians.¹ Although frames for some populations are easily obtained, they are not always available, and other sampling methods (to be discussed later) are therefore necessary.

Examples of frames:

1. The registrar's office of a university has a list of all students who are in residence as of a particular date.
2. The telephone company has a list of all subscribers to its service.
3. A business firm has ledgers which carry the names of all accounts receivable, accounts payable, etc.
4. A city has a directory which lists its inhabitants as of a certain date.

1.7 Experiment in the Use of a Table of Random Numbers

The selection of a random sample is so important in statistical theory that each student should have experience in using a table of random numbers. Exercise 1.6 provides this experience and, moreover, supplies a model and data for later use in the development of additional theory.

EXERCISES

1.6. *a.* Each student is to draw one sample of 10 observations from the bowl of 998 chips (Table 1.3), using the table of random numbers (Table 1.2). Record the 10 values of X in a notebook for safekeeping.

b. Total the 10 X values and find the average X value by dividing this total by 10. Round each of these averages to one decimal place according to these rules:

1. If the second decimal digit is greater than 5, round the first decimal place up. Example: 3.78 rounds to 3.8.

2. If the second decimal digit is less than 5, round the first decimal place down. Example: 3.73 rounds to 3.7.

3. If the second decimal digit is equal to 5, round as follows: (*a*) If the first

¹ See William Edwards Deming, "Some Theory of Sampling," p. 27, John Wiley & Sons, Inc., New York, 1950.

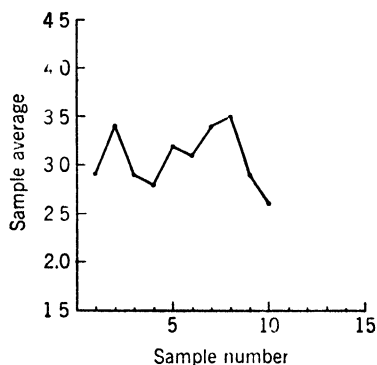
decimal digit is even, round down. Example: 3.25 to 3.2. (b) If the first decimal digit is odd, round up. Example: 3.35 to 3.4.

c. Number each student in the class consecutively, starting with 1, 2, 3, Record the average obtained by each student on a work sheet laid out as follows:

Sample	Total of 10 X values	Average X value
1	42 3	4.2
2	38 3	3.8
3	48 5	4.8
.	.	.
.	.	.
.	.	.

Individual readers without the inclination to draw several samples of 10 observations may use the data in Table 1.4 to complete the exercise.

d. Plot the sample averages in the order of their appearance on a chart, like the one shown below.



e. Make a distribution of the sample averages by counting the number of samples whose averages were between the class limits shown below:

<i>Class limits of the average value of X</i>	<i>Number of samples</i>
1 .95–2 .25	
2 .25–2 .55	
2 .55–2 .85	
2 .85–3 .15	
3 .15–3 .45	
3 .45–3 .75	
3 .75–4 .05	
Total.....	N

If sample averages less than 1.95 or greater than 4.05 occur, add intervals with the same spacing as in the above distribution.

TABLE 1.4. OUTCOMES OF 30 RANDOM SAMPLES DRAWN FROM TABLE 1.3

Sample	Total of 10 X values	Sample	Total of 10 X values
1	29 0	16	31 8
2	34 4	17	31 0
3	29 2	18	33 7
4	28 2	19	35 4
5	31 6	20	31 9
6	31.2	21	25 6
7	34 3	22	33 8
8	35 2	23	33 0
9	29 3	24	31 6
10	25.8	25	34 4
11	30.0	26	31.3
12	30.8	27	36 3
13	26 0	28	35 8
14	25 8	29	28 0
15	33 1	30	29 4

1.7. Answer these questions about Exercise 1.6:

- What is the largest sample mean? The smallest?
- What is the largest value of X in the population (Table 1.3)? The smallest?
- How does the range of values of the means compare with the range of values in the population?
- Can you tell by inspection what is the value of the population average? Do the sample averages vary among themselves and also differ from the population average?
- How does the grand average of all sample averages compare with the population average?
- How does the general shape of the distribution of averages made in part (e) compare with the shape of the distribution of the population?

1.8. How would you number the units of a population in order to use just three columns of the table of random numbers if there are 1000 units in the population?

1.9. Suppose that there are 250 units in a population. How would you number them so as to use three columns of the table of random numbers but not throw away selections from the table for which there are no units in the population with corresponding serial numbers?

1.8 More on the Operation of a Statistical-decision Rule

With the definitions and notation developed in the preceding sections, a solution to the horse-race problem can be started. Here is a very brief review of the problem. A friend of yours claims he has a system which will pick winners in horse races 50 per cent of the time. He wants you

to invest some money with him. You ask him to test the system on 10 races which you select at random from a large number of races. Your problem is to pick a number between 0 and 10 which will help you in making a decision, i.e., you must decide how many races he must win before you will invest your own money in his system.

You have already experimented with the operation of a statistical-decision rule by sampling red and white beads from a box which contained an unknown percentage of each color. For classroom demonstration, the author carried out a similar experiment. The experimental procedure may be put into the statistical terms developed in Sec. 1.4.

Population: 500 beads; some red, some white.

Elementary unit: A bead.

Characteristic: Color of the bead.

Observation: If red, score a 1; if white, score a 0.

Sample: 10 beads.

The color of the bead is a qualitative characteristic, but you may assign a numerical value to the color of the bead by scoring a 1 for the occurrence of a red bead and a 0 for the occurrence of a white bead. A sample of 10 red and white beads recorded in terms of the color characteristic as

R W R R W W R W W R

is now scored as

1 0 1 1 0 0 1 0 0 1

The individual observations are no longer important in themselves, but only in comprising the whole sample of 10 observations. (In the horse-racing problem, the outcome of any particular race is unimportant except as it contributes to your knowledge of how many races in the sample were won or lost.) Interest centers in the sample of 10 observations and *a variable must be defined for the sample*. A very simple operation which defines a variable is the sum--the sum of the observations in the sample. According to the rule for assigning numerical values, this sum is

$$1 + 0 + 1 + 1 + 0 + 0 + 1 + 0 + 0 + 1 = 5$$

Another sequence of 5 beads could be

R W R R R

which would be scored as

1 0 1 1 1

whose sum is

$$1 + 0 + 1 + 1 + 1 = 4$$

In the bead problem, this sum is the *number of red beads* in the sample. Ten numbers, or observations, were generated by turning the crank of the sampling box to mix the beads thoroughly and then tilting the box to allow the 10 beads to run out, an approximation to a random-sampling selection. The particular values which occur in a given sample depend upon the random operation (or chance) selecting the elementary units which comprise the sample. The term *random variable* will therefore be used when speaking of a variable whose observations are generated by a random or probability selection.

The results of the author's experiment with 100 samples selected from the bead box are shown in Table 1.5. The following notation is used:

x_i = value of random variable of sample (number of red beads)

f_i = frequency of occurrence of x_i

p_i = relative frequency of occurrence of x_i (f_i/N)

N = number of samples ($N = 100$)

TABLE 1.5. FREQUENCY AND RELATIVE FREQUENCY OF OCCURRENCE OF RED BEADS IN 100 SAMPLES OF 10 BEADS FROM A POPULATION WITH PERCENTAGE OF RED BEADS UNKNOWN

x_i	f_i	p_i	x_i	f_i	p_i
0	0	0.00	6	22	0.22
1	2	0.02	7	6	0.06
2	8	0.08	8	2	0.02
3	8	0.08	9	2	0.02
4	23	0.23	10	0	0.00
5	27	0.27			

The table is read as follows: Of the 100 random samples of 10 beads selected:

1. No sample contained 0 red beads.
2. 2 samples contained 1 red bead.
3. 8 samples contained 2 red beads.
4. 8 samples contained 3 red beads.
5. 23 samples contained 4 red beads, etc.

Or, the table may be read like this:

1. 0 per cent of the samples contained 0 red beads.
2. 2 per cent (.02 in decimal form is 2 per cent) of the samples contained 1 red bead.
3. 8 per cent of the samples contained 2 red beads.

4. 8 per cent of the samples contained 3 red beads.

5. 23 per cent of the samples contained 4 red beads, etc.

One interpretation is in terms of the *number* of samples containing a given number of red beads; the other interpretation is in terms of the *percentage* of samples containing a given number of red beads.

Suppose that you had to decide whether to invest in the horse-racing scheme in each of the 100 trials of *this* experiment, using the statistical decision rule to accept the box as filled with 50 per cent red beads (invest in the system) if 5 or more red beads occur in the sample; otherwise, to reject the box as filled with 50 per cent red beads (do not invest in the system). How many decisions to invest were there in the total number (100)? How many samples contained 5 or more red beads? There were 59 samples containing 5 or more red beads; 59 times out of the 100 opportunities, or 59 per cent of the time, you would have decided to invest your money; 41 per cent of the time you would have decided not to invest. If the rule were changed so that you would invest if 4 or more red beads occurred in the sample, you would have invested 82 per cent of the time. The percentage of decisions to invest in accordance with other rules can easily be found.

Which of these decision rules is the best one to use? What is a good rule to follow in making this investment decision? The proper rule is the one which will lead you to the proper decision. *The effectiveness of any rule will be measured in terms of the percentage of incorrect decisions.* One rule is preferable to another if it leads you to make an incorrect decision a smaller percentage of the times it is applied. The author claims that the rule to invest in the scheme if 5 or more red beads occur in the sample will lead you to make an incorrect investment decision 41 per cent of the time because he filled the box used in his sampling experiment with 50 per cent red and 50 per cent white beads. Of the 500 beads in the box, 250 were red and 250 were white. The claim that you made the wrong decision 41 per cent of the time (because you refused to invest your money in 41 out of the 100 times you had to make the decision) is based upon the following reasoning, which is crucial to your understanding of statistical decisions. The argument follows.

1. You should agree with this statement: If the horse-racing system will pick 50 per cent winners, you will invest your money. If you disagree, go back to Sec. 1.1 and read the original statement of the problem again.

2. The experiment was set up so that 250 red and 250 white beads filled the box (50 per cent red beads). (In the title to Table 1.5 it was stated that the percentage of red beads was unknown.) The claim about the percentage of red beads (the percentage of winners) was made true in the experiment by the selection of the contents of the box.

3. If you agree with (1), *the correct decision is always to invest* because the box was filled with 50 per cent red beads, the counterpart of 50 per cent winners in the horse-race problem. Each decision to invest as a result of one random sample is a correct one; each decision not to invest as the result of a random sample is an incorrect one. With the rule to invest if 5 or more red beads occur (5 or more winners), 59 out of the 100 random samples led to the decision to invest and 41 led to the decision not to invest. In other words, 59 per cent of the decisions were correct and 41 per cent were incorrect.

Why was a wrong decision reached? Because the element of chance decided what number of red beads or what number of races were won in any one sample of 10 beads or 10 races. Whenever the random selection put 4 or fewer red beads in the sample, the decision was not to invest—an incorrect decision. The relative frequencies of making an incorrect decision with different rules are summarized in Table 1.6. The relative

TABLE 1.6. RELATIVE FREQUENCY OF DECISIONS NOT TO INVEST WHEN BEAD BOX IS FILLED WITH 50 PER CENT RED BEADS

<i>Number of red beads which lead to decision to invest</i>	<i>Percentage of decisions not to invest</i>
5 or more	41
4 or more	18
3 or more	10
2 or more	2
1 or more	0
0 or more	0

frequency of making this kind of error can be reduced by changing the number used in the statistical decision rule. You might think it wise to eliminate this kind of error entirely, but this is another story and will have to be postponed. It suffices here to say that one of the determinants of the decision-making rule is the relative frequency with which this kind of error is committed—*rejecting a claim or hypothesis when, in fact, that claim is actually correct.*

What if the claim is not true? What if the system picks winners less than 50 per cent of the time? How does this possibility influence the statistical decision rule? The box which you used to conduct your experiment in making decisions to invest (Exercise 1.4) was filled with 500 beads of which 100 were red and 400 were white. Thus, 20 per cent were red, and your sampling box was filled with such a proportion of beads that *the friend's claim was not true.* You should now agree with this statement: Since the box was filled with 20 per cent red beads, which corresponds to a system which picks 20 per cent winners, *the correct decision is not to invest.* In every one of the 100 opportunities to decide

whether to invest, the decision should have been not to invest. The incorrect decision in your experiment is different from that of the author's experiment. His error is to refuse to invest when the claim (50 per cent) is true; your error is to invest when the claim (50 per cent) is not true. The first error is rejecting the claim when, in fact, it is correct; the second error is accepting the claim when, in fact, it is false. Another determinant of the best decision rule is the relative frequency with which this kind of error is committed—*accepting a claim when, in fact, that claim is actually false*.

The development has gone far enough for the present. Observed frequency and relative frequency distributions are not too important for this course, though one of the largest businesses in the United States—the insurance business—is built upon just such a foundation. Mortality tables and other experience tables are nothing more than empirical relative frequency distributions based upon the historical experience of a large number of people. Fortunately, it is not necessary to observe the operation of the betting system for a long period of time in order to decide whether to invest in the horse-race scheme. The relative frequencies of making incorrect decisions, which serve as the basis for determining the best statistical decision rule, may be predicted by using simple formulas of probability theory. The next chapter is devoted to a discussion of probability along with some definitions, symbols, and algebra needed to pursue the argument further.

EXERCISES

1.10. *a.* With a rule to invest if 5 or more red beads occur in the sample, what percentage of the time did the class decide to invest—an incorrect decision—in Exercise 1.5? With a rule of 1 or more? 3 or more? 2 or more? 1 or more? 0 or more?

b. Enter these percentages in a table similar to Table 1.6.

c. As the number of red beads in the rule is decreased, what happens to the percentage of incorrect decisions?

d. Compare the results with Table 1.6. How do the percentages of these two different incorrect decisions behave? Does one become larger as the other becomes smaller?

e. What happens when the percentage of one incorrect decision is reduced to zero?

f. Can the percentages of each of the errors be made small simultaneously?

1.11. A survey of certain characteristics of public accounting firms in a given city is to be made. An up-to-date list of certified public accountants from which to select the sample is available. In order to select the accounting firms for the sample, the N accountants on the list were numbered, n random numbers from a table of random numbers were selected, and the accountants identified by the

random numbers were chosen. The accounting firms to which these accountants belonged were then used as the sample. Does this method ensure an equal probability of selection for each accounting firm? Why or why not?

1.12. Give an example of a population that is so large that it must be studied with a sample. Give an example of a population that must be studied with a sample because the elementary unit is destroyed in the measurement procedure.

1.13. Describe a situation in which you would plan your statistical decision rule in advance. Describe a situation in which you would wait until you observed the sample results before making your decision.

Write out the meaning of each of the following terms as it relates to the next two problems: hypothesis, population or universe, elementary unit, characteristic of the elementary unit, frame, sample, random, random variable, and observation.

1.14. The head of a local real-estate board believes that 50 per cent of the owner-occupied single-family dwelling units in his city are painted every 3 years. To check his claim, he selects a random sample of 144 houses and questions the owner about when he last painted and when he intends to paint his house.

1.15. Walnuts are graded according to standards of size, color, shape, etc., which are set by state law. In order to qualify for a grade A label, at least 16 per cent of a grower's crop must meet the specifications. Obviously not all walnuts in a crop can be inspected. A sampling procedure consists of selecting at random (with a mechanical selector) 100 walnuts from the crop, grading each one, and assigning a grade A label to the crop if 13 or more walnuts meet the state specifications.

1.16. Two kinds of errors emerged from the discussion in Sec. 1.8. What are these two errors in the house-painting and the walnut-grading problems (Exercises 1.14 and 1.15)?

CHAPTER 2

TOOLS—NOTATION, PROBABILITY, AND THE BINOMIAL DISTRIBUTION

Before continuing the main line of argument of statistical inference some statistical tools are needed. These include some notation and computational techniques, a discussion of probability, and an introduction to the binomial distribution. This chapter is an interruption to discuss these tools.

2.1 Notation

A part of any discipline or science is mastery of the notation which is used. Symbols are employed to designate mathematical operations and to identify certain variables, constants, and statistics which are the results of computations. Unfortunately, there is no standardized notation for all statistical terms. However, this text will adopt those symbols which are more widely used than others and will be consistent in its use of other symbols. A brief outline of most of the notation to be used during the course follows:

1. Letters will be used to represent variables. For example, h for height, w for weight, x and y for *any* variable, and n for sample size.

2. Greek letters will be used to designate some terms. The most common are:

α (alpha)	σ (sigma)	Δ (delta)
β (beta)	ρ (rho)	π (pi)
μ (mu)	Σ (capital sigma)	χ (chi)

3. Subscripts are often attached to a letter to indicate different values of the same characteristic. For example, n_1, n_2, n_i, n_j . These are read either " n one," " n two," " n i ," " n j ," or " n sub one," " n sub two," " n sub i ," " n sub j ." Subscripts may also attach to the letter X : X_1, X_2, X_i . If X designates the characteristic age for some population, X_1 is the age of individual 1 drawn in the sample; X_2 is the age of elementary

unit 2; X_i is the age of the i th elementary unit, that is to say, it is a *general term*.

4. Signs of mathematical relationships:

$x = y$	(x equals y)
$x > y$	(x greater than y)
$x < y$	(x less than y)
$x \geq y$	(x equal to or greater than y)
$x \leq y$	(x equal to or less than y)
$\sqrt{x}$	(square root of x)
$x \doteq y$	(x is approximately equal to y)
$x \neq y$	(x does not equal y)
$x + y$	(x plus y)
$x - y$	(x minus y)
$x \cdot y$	(x times y , that is, x multiplied by y)

5. Signs of grouping for multiplication and division:

()	(parentheses)
[]	(brackets)

Examples: $(a - b)$ divided by $(c - d)$ is written as

$$(a - b)/(c - d)$$

$(a - b)$ multiplied by $(c - d)$ is written as

$$(a - b)(c - d)$$

6. Summation is one of the most important and most frequently used operations in statistics work. It means the addition of several observations of a variable.

Example 2.1. Take the five numbers 4, 7, 9, 16, 22. Designate each one with a letter X : X_1, X_2, X_3, X_4, X_5 . X_i is the general term, where the subscript i identifies the observation and, therefore, its value. If i equals 1, the observation is X_1 and its value is 4. If i is 5, the observation is X_5 and its value is 22. The subscript i may take on any one of the values 1, 2, 3, 4, or 5. It is customary, then, to write the general term including the limits for i as X_i , where i may take on the values 1, 2, 3, 4, or 5, which may be abbreviated to $X_i, i = 1, 2, 3, 4, 5$ or still further, $X_i, i = 1, \dots, 5$. With n observations, where n is the number in the sample, this is written as $X_i, i = 1, \dots, n$.

Suppose the five numbers are to be added, thus:

$$4 + 7 + 9 + 16 + 22 = 58$$

$$X_1 + X_2 + X_3 + X_4 + X_5 = ?$$

The operation of addition in terms of the X 's will be written

$$\sum_{i=1}^{i=5} X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

In general,

$$\sum_{i=1}^{i=n} X_i = X_1 + X_2 + \cdots + X_n$$

In elementary statistics, the summation is usually over all n observations of the sample and the limits for Σ are often omitted. The capital sigma is the symbol used for the operation of addition. The notation, including the limits, is read: "the sum of the X_i 's as i goes from 1 to 5 or the sum of the X_i 's as i goes from 1 to n ." The limits are important because they indicate just what observations are included in the summation.

7. Very soon, these two notations will appear:

$$\sum_{i=1}^{i=n} X_i^2 \quad \text{and} \quad \left(\sum_{i=1}^{i=n} X_i \right)^2$$

What is the difference between them? They will both be used in the same equation and they are not different ways of writing the same operation.

2.2 Computational Aids

You will need to develop some facility with arithmetic and elementary algebra in this course. A very excellent book which will help you to learn or to review, as the case may be, is Helen M. Walker, "Mathematics Essential for Elementary Statistics," Henry Holt and Company, Inc., New York, 1951.

Simple Problems for Review

1. You should know how to add and subtract *accurately*.
2. You should know how to multiply. Try these problems: (64)(56), (1.2)(2.8), (3.3)(0.2), (0.012)(0.045).
3. Division: 65/56, 1.2/5.8, 3.3/0.2, 0.012/0.015.
4. Percentages: Express both in percentage and in decimal form 75/150, 100/33, 62/28.

Mathematical Tables

"Mathematical Tables from Handbook of Chemistry and Physics," Chemical Rubber Publishing Co., Cleveland, Ohio.

"Barlow's Tables," Chemical Publishing Company, Inc., New York.

Herbert Arkin and Raymond R. Colton, "Tables for Statisticians," College Outline Series, Barnes & Noble, Inc., New York, 1950.

Tables most frequently used in this course are reproduced in Appendix A. These include a table of squares and a table of square roots with an explanation of their use.

Machine Calculation. Many statistics laboratories are equipped with automatic or semiautomatic computing machines. Work the problems for review above on the machine. In addition, make the following computations on the machine:

Add: 17, 18, 25, 62, 01, 100, 54.

Add: $(17)^2$, $(18)^2$, $(25)^2$, $(62)^2$, $(01)^2$, $(100)^2$, $(54)^2$.

Remark. In summing these squares, cumulate the total in the machine. Do not make each multiplication; write the answer on another sheet of paper, and then add. The total may be cumulated by clearing only the multiplier and multiplicand dials after each operation and leaving the answer in its dial.

EXERCISES

2.1. You are given a sample of 10 observations which take on the following numerical values: 6, 4, 3, 1, 2, 8, 10, 7, 9, 5.

a. Identify each of the numbers with a value of x_i (from x_1 to x_{10}) in the order in which they appear above. For example, $x_1 = 6$, $x_2 = 4$, . . . , $x_{10} = 5$.

b. Write out in words the meaning of each of the following operations, and carry out the arithmetic:

$$(1) \sum_{i=1}^{i=10} x_i$$

$$(2) \sum_{i=1}^{i=10} x_i^2$$

$$(3) \left(\sum_{i=1}^{i=10} x_i \right)^2$$

$$(4) \sum_{i=1}^{i=5} x_i$$

$$(5) \sum_{i=1}^{i=5} x_i^2$$

$$(6) \left(\sum_{i=1}^{i=5} x_i \right)^2$$

c. You were given a sample of 10 observations. Let n designate the sample size, that is, $n = 10$. Now compute the following:

$$(1) \sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i \right)^2}{n}$$

$$(2) \sqrt{\sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i \right)^2}{n}}$$

d. The average value of the 10 observations in the sample is the total of the 10 numerical values divided by 10. In general, the average value is the total of the

observations divided by the number of observations (n). Let this average value be symbolized by $\bar{X}$. Then,

$$\bar{X} = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

Find $\bar{X}$ for the 10 numbers.

e. Let the deviation of any observation from its average be symbolized by $(x_i - \bar{X})$. Set up a table showing the values of x_i , $(x_i - \bar{X})$, and $(x_i - \bar{X})^2$ as follows:

i	x_i	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
1	6	5	25
2	4	-1	1
	.	.	
	.	.	
	.	.	
	.	.	
	.	.	
	.	.	
10	5	-5	25
Total		?	?

(The dots indicate that the values between $i = 2$ and $i = 10$ are omitted. You are to fill in the missing values in your table.) The total of the third column is $\sum_{i=1}^{i=10} (x_i - \bar{X})$. The total of the fourth column is $\sum_{i=1}^{i=10} (x_i - \bar{X})^2$. Write out the meaning of each of these operations in words.

f. Compare the total of the fourth column with the answer to part (c) of Exercise 2.1.

2.2. Use the data gathered in part (a) of the experiment of Sec. 1.7 (Exercise 1.6), and compute the following:

$$\frac{\sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n}}{n - 1}$$

Record this for safekeeping. It will be used in Exercise 9.17.

2.3 Probability

The word "probability" or "chance" is used loosely in everyday conversation. Statements like "It will probably rain tonight," or "This team has a good chance of winning the game," or "The chances are that you will recover from the cold," are commonly heard. Probability statements are just as common in the business world. Two of them which appeared recently are:

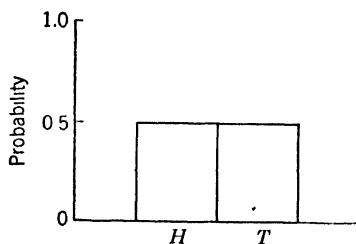
It seems probable in the event of Republican election victory that a return to more classical economic and political concepts would result. This would indicate

less Federal control, more local autonomy; less spending, more budget balancing and less taxation, more investment incentive.¹

Efforts to halt the relentless surge of taxes—Federal, State, and local—doubtless will be one of the first New Year's resolutions to be made. Chances of the resolution lasting very far into the year are another matter.²

These statements are vague and not very useful in statistics. The discussion must be more quantitative, and the aim here will be to present conditions leading to numerical statements of probability. The subject will be introduced by a study of the simplest phenomena of chance—coin and dice tossing.

Classical Notions about Probability. A coin has two sides, one commonly called “head” and one “tail.” It is not unusual to think of the two sides of a symmetrically made coin as being “equally likely” to turn up when the coin is tossed and to say that the probability of a head is one-half and the probability of a tail is one-half. The numerical values of probability are arrived at as follows. Probabilities are measured on a scale which extends from *zero* at one end to *one* at the other end. The total probability of one must be allocated to the two possible alternatives for a coin—head and tail. If it is agreed that these two sides are “equally likely,” the total probability is apportioned equally between them, or one-half to head and one-half to tail. These probabilities will be written $\text{Pr}(\text{head}) = 1/2$ and $\text{Pr}(\text{tail}) = 1/2$. A graph of the probabilities may take the following form, with the scale of probabilities written on the ordinate, or vertical, scale.³



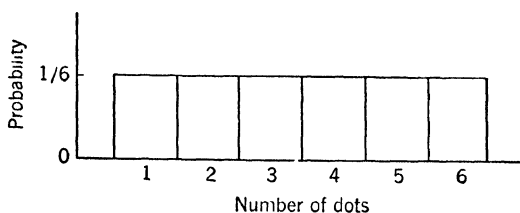
Now consider a die. There are 6 sides to a die, each containing dots. One side has 1 dot, another 2, . . . , and the last side has 6 dots. If

¹ “Security and Industry Survey,” p. 1, Merrill Lynch, Pierce, Fenner & Beane, November, 1952.

² *Los Angeles Times*, Dec. 21, 1952.

³ This graph is called a histogram or a bar chart. It is customary to say that the total area under the histogram of probabilities is one or unity. The area of each bar represents the probability of the associated event. Since each bar has one-half of the total area (unity), the probabilities are one-half for head and one-half for tail. The histogram and its interpretation are trivial in the case of the coin. The representation of probabilities as areas under a graph is introduced now because it will later be useful in more complicated probability problems.

it is agreed that each of these sides is equally likely, one-sixth of the total probability of unity will be allocated to each side. Therefore, $\text{Pr}(1 \text{ dot}) = 1/6$, $\text{Pr}(2 \text{ dots}) = 1/6$, . . . , $\text{Pr}(6 \text{ dots}) = 1/6$. A histogram which represents this probability situation is shown below.



Again the total area of this histogram is unity; the area of each bar is one-sixth of the total area and represents a probability of $1/6$.

This simple discussion leads to the following statement: *If an event can happen in m ways out of a total of N possible ways, each of which is to be considered equally likely, the probability of the event is the ratio of m to N .*

Example 2.2. A head can appear on a coin in only one way; there are two possible ways a coin may appear. $m = 1$ and $N = 2$, and $\text{Pr}(\text{head}) = m/N = 1/2$.

A 5 can appear on a die in only one way; there are 6 possible ways a die may appear. $m = 1$ and $N = 6$, and $\text{Pr}(5) = m/N = 1/6$.

The most serious objection to this classical statement of probability is that “equally likely” means “equally probable,” and probability is defined in terms of itself. A more modern viewpoint of probability avoids this difficulty by specifying a mathematical model based upon certain arbitrary definitions. The model is useful as long as deductions from it describe the results of random sampling in real-life situations.

Mathematical Probability. You may already have grasped the idea that probabilities relate to collections of events, numbers, occurrences, etc., like the sides of a coin or of a die. This collection is called a *fundamental probability set*, usually abbreviated F.P.S. *Subsets* of this F.P.S. will be called *events* and given the notation E_i ; for example, the head of a coin or six spots on a die. These subsets or events may be conveniently represented by *points* on a line. The collection of all the points corresponding to all the events is called an *event space* which will be symbolized by S . To each event E_i there is a number attached, called the *probability* of E_i and written $\text{Pr}(E_i)$. This number must possess certain attributes which are given by

1. $\text{Pr}(E_i) \geq 0$ (i.e., nonnegative).

2. $\sum_{\text{All } E_i \text{ in } S} \text{Pr}(E_i) = 1$ (total probability, or probability over the whole

event space, must equal unity).

The previous paragraph stated that a value of a probability $\Pr(E_i)$ is associated with each event E_i . Another number is also associated with the event E_i . This number is the value of a variable x_i . Then, the probabilities of the events in S , $\Pr(E_i)$, automatically define the probabilities of the values of x_i , $\Pr(x_i)$, and x is called a *random variable*. When only a finite number of events and therefore a finite number of values of x exist, the variable is said to be *discrete*. The probability distribution of x which completely describes the relationship between x and $\Pr(x)$ is written $f(x)$.

These definitions can be illustrated for a die, as shown in Fig. 2.1. A die has six sides which are the E_i events represented by the six points on a line in the middle of Fig. 2.1. These six points taken as a group are

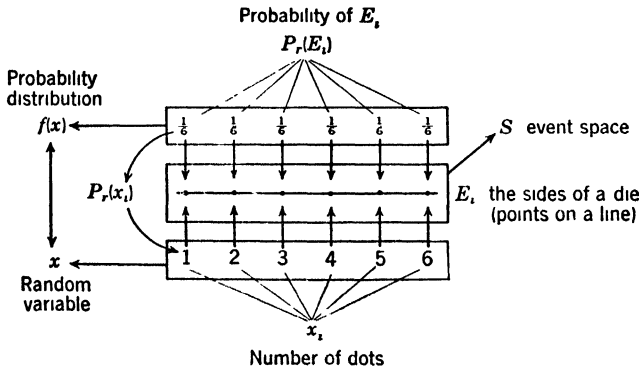


FIG. 2.1. Probability model for a die.

the event space S . To each point E_i there is associated a number called the probability of E_i , written $\Pr(E_i)$. The values given to each of these numbers are conventionally $1/6$. With each of these points E_i it is also natural to associate another number, that is, the number of dots on one face of the die. These are the values of the variable x_i . Through the common association with the events E_i , each of the probabilities $\Pr(E_i)$ is associated with the corresponding value of x_i and defines the probability of x_i , $\Pr(x_i)$. The collection of all the values of $\Pr(x_i)$ is the probability distribution $f(x)$. The collection of values of x_i is the random variable x (the number of spots on the die). The probability distribution may be set out in a table:

x	1	2	3	4	5	6
$f(x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

These probabilities possess the two attributes given at the beginning of this discussion:

1. Each probability ($1/6$) is nonnegative.
2. The total of the probabilities for all 6 events is 1.

$$1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$$

Two Probability Theorems. The *addition theorem* states that the probability of occurrence of one of several *mutually exclusive* events is the sum of their separate probabilities. (Mutually exclusive means that the events cannot occur simultaneously.) For example, if E and F are mutually exclusive,

$$\Pr(E \text{ or } F) = \Pr(E) + \Pr(F)$$

Example 2.3. Let E be the event 1 on a die and F be the event 5 on a die. These events are mutually exclusive. By the addition theorem, the probability of either a 1 or a 5 is $1/6 + 1/6 = 1/3$.

Example 2.4. Let E be the event 5 on two dice and F be the event 2 on two dice. $\Pr(E) = 4/36$; $\Pr(F) = 1/36$. These events are mutually exclusive, and $\Pr(5 \text{ or } 2) = \Pr(5) + \Pr(2) = 4/36 + 1/36 = 5/36$.

The *multiplication theorem* states that if several events are *independent* of each other, the probability of their simultaneous occurrence is the product of their separate probabilities. Independence is not rigorously defined here, but is given the same meaning as in everyday conversation. The outcome of one event is not influenced by the outcome of another event, nor does it influence the outcome of any other event. If E and F are two independent events, or if no inference about the occurrence of one event follows from the occurrence of another event,

$$\Pr(E \text{ and } F) = \Pr(E) \cdot \Pr(F)$$

Example 2.5. Let E be the event 7 on a pair of dice. Let F be the event 7 on another pair of dice. $\Pr(E) = 1/6$; $\Pr(F) = 1/6$. Are these events independent? Does the 7 of event E in any way affect the 7 of event F or is it, in turn, affected by the 7 of event F ? E and F are independent. Then, according to the multiplication theorem,

$$\Pr(E \text{ and } F) = \Pr(7 \text{ and } 7) = \Pr(7) \cdot \Pr(7) = 1/6 \cdot 1/6 = 1/36$$

Example 2.6. Let E, F, G, H , and I be the events "head" on a coin for which the probability is $1/2$. What is the probability that all five of these events will be heads?

$$\Pr(E, F, G, H, I) = \Pr(E) \cdot \Pr(F) \cdot \Pr(G) \cdot \Pr(H) \cdot \Pr(I) = 1/2^5 = 1/32$$

EXERCISE

2.3. a. There are 36 points in the fundamental probability set which consists of pairing faces on two dice, thus: (1,1), (2,1), (1,2), . . . , (6,6). Set out these 36 points on a sheet of paper in the form of a square with 6 points assigned to a side.

b. What probability would you attach to each of these points? Write in these numbers above each of the points.

c. Attach to each point in the F.P.S. another number: the total number of dots on the pair of dice which the point represents. For example, (1,1), total = 2; (1,2), total = 3, etc.

d. Collect together all the points in the F.P.S. which are alike with respect to the number of dots. These are the events E_i of the event space S . How many points are in the event space? (11).

e. What probability would you attach to each of these events?

f. To each of the events E_i associate the value of the variable x which is the sum of the dots on the pair which that subset represents. Now each event will have two numbers attached to it: a value of $\Pr(E_i)$ and a value of x_i .

g. Order the values of x from lowest to highest, and write down the probability distribution of x and $f(x)$. Present this relationship in a table.

Use of Probabilities in Prediction. The probabilities of the previous sections are merely numbers which satisfy a particular set of definitions. They were determined by deductive reasoning, by conventionally assigning them to events, or by applying the addition and multiplication theorems. No coins were tossed; no dice were rolled. Yet, if these probabilities are to be useful, they must bear some relation to experiments in which coins are tossed or dice are rolled. What is this relationship? Of what value are the numbers which are called probabilities?

These probabilities may be used to predict the outcomes of real-life experiences. If a coin is tossed a large number of times, the prediction is that it will turn up heads 50 per cent of the time. If a die is rolled a large number of times, the prediction is that it will turn up 5 in 1/6 of the rolls. The probability predicts the relative frequency of occurrence of some event. It does not predict the outcome of any one trial, or the order of the outcomes of several trials, but only what will happen in the long run or on the average in a large number of trials. What is the justification for the use of probabilities in making predictions? The justification is a very simple one: these probabilities really work when they are assigned to events and the probability laws are applied in actual experimental problems. A great mass of empirical evidence shows that they do work and no serious evidence shows the contrary. In other words, experimental data verify the theorems of probability.

Two ideas are involved in this justification. The first is *randomness*. In reality, long sequences of tosses of a coin or rolls of a die are experiments in practical random sampling, and the probabilities of the mathematical model are numerically close to the observed relative frequencies. Moreover, when the discrepancy between the idealized experiment and actual experience is great, there is usually a good reason for it: the tosser can manipulate the coin, or the cards are dealt off the bottom of the deck, or the die is loaded.

The second idea involved is not only that the relative frequency of occurrence in actual long-run series of trials approximates the probability but also that *the larger the number of observations, the better generally the approximation*. The randomness of a sequence precludes the predictability of the order of outcomes as well as the value of an outcome, but the sequence as a whole behaves regularly and permits prediction. This is a very rough statement of the *law of large numbers*, or, as it is called in everyday conversation, the law of averages.

The Use of Relative Frequencies in Prediction. In simple games of chance (coins, cards, dice, etc.) probabilities may be computed without performing experiments. As demonstrated in the previous sections, numbers which both satisfy the probability theorems and are useful for making predictions may be determined by deduction, enumeration, or convention. But in many real-life situations probabilities cannot be computed. It is not possible, for example, to deduce the probability that an individual will die within a given span of time. All that one can do is observe in a large-scale experiment the relative frequencies with which people do die and use these relative frequencies for predicting what will happen in the future. Insurance companies make their mortality tables in approximately this manner. If of 100,000 individuals alive at the age of 10, 92,637 are alive at age 20, the relative frequency of an individual aged 10 living to be age 20 is .926. Insurance companies use this empirically determined relative frequency to predict the future relative frequency and to sell insurance policies based upon such a determination. In doing this, they rely upon randomness and the law of large numbers for success. The life-death experience of the 100,000 individuals who are examined is assumed to be a *random* sequence; a sample of 100,000 is assumed large enough for the *law of large numbers* to operate and to ensure that the observed relative frequency of death is numerically close to the unknown probability. Without randomness and without a large number of observations, empirical relative frequencies are hazardous predictors.

Remark. The assumption of independence from one observation to the next is crucial in using empirically determined relative frequencies for predicting the future. Death rates, for example, from diphtheria, influenza, fires, etc., are not based upon independent observations

Expected Value. The mean of a random variable is called the expected value and is written $E(x)$ or sometimes simply Ex . The symbol E stands for an operation. In the discrete case,

$$Ex = \sum_{\text{All } x_i \text{ in } S} x_i \cdot \Pr(x_i) \quad (2.1)$$

where x_i is a value of the random variable, $\text{Pr}(x_i)$ is its probability, and the summation is over all the values of x_i .

Example 2.7. What is the expected value of the number of heads on a toss of two fair coins? The probability distribution is:

x	0	1	2
$f(x)$	1/4	1/2	1/4

$$Ex = (0) \cdot 1/4 + (1) \cdot 1/2 + (2) \cdot 1/4 = 1 \text{ head}$$

Example 2.8. Suppose you were to flip the two coins in Example 2.7 with George and suppose further that George will pay you \$1 for every head that appears. If the game is to be "fair," how much should you pay to play with George? (In a fair game, the expectation of the winnings is equal to the entry fee.) The probabilities now attach to the payment for each outcome of the coin toss.

x	0	1	2
$f(x)$	1/4	1/2	1/4
W	\$0	\$1	\$2

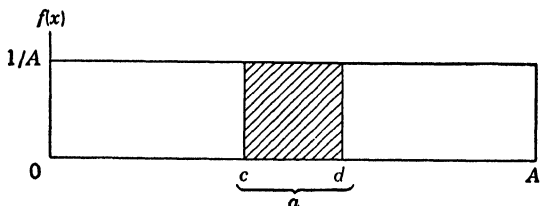
$$E(W) = (\$0) \cdot 1/4 + (\$1) \cdot 1/2 + (\$2) \cdot 1/4 = \$1$$

You should pay neither more nor less than \$1 to play if the game is to be fair.

Elements of Mathematical Probability: Continuous Case. Let the event space of probability theory consist of all the points on the line segment from 0 to A . This is shown in the diagram at the bottom of the page. Since the variable x may take on any value along this line segment, it is said to be *continuous*. Corresponding to the set of attributes in the discrete case is a set of attributes for the continuous case.

1. $f(x) \geq 0$ for every x . (Probability must be nonnegative.)
2. The total probability must equal unity. In the continuous case, there is need for a special kind of addition (which need not concern you) indicated by an integral sign $\int$. Since values of x may range from 0 to A , the sum of $f(x)$ for values of x from 0 to A must equal 1.

If x is considered to be a random point on this line segment of length A , $f(x)$ is conventionally thought to be a rectangular distribution of height $1/A$.



Probability has been represented as an area before in Sec. 2.3. The total area of $f(x)$ is equal to unity. Then, the probability that the random point x will lie in the interval cd is the *area* above this interval. This area may be found by adding according to the operation $\int$. This is the *addition theorem* for continuous variables. In this very special case, this addition need not be carried out because the height of $f(x)$ is always $1/A$ and, therefore, since the area of a rectangle equals base times height, all areas of $f(x)$ are proportional to their bases. The probability of a random point lying in an interval of length a on the line of length A is simply the ratio a/A .

Example 2.9. The pointer on a roulette wheel may stop at any point on the wheel. On many wheels the circumference is divided up into 38 equal segments which are numbered 0, 00, 1, 2, . . . , 36. The whole circumference may be considered to have a length $A = 38$. The length of any one segment is $a = 1$. The probability distribution for such a roulette wheel is:

x	0	00	1	2	...	36
$f(x) = a/A$	1/38	1/38	1/38	1/38	...	1/38

EXERCISES

2.4. a. If you toss two dice, what is the probability that you will roll a 7 on the first roll? Either a 7 or an 11? Either a 3 or a 12?

b. How did you compute these probabilities?

c. What assumptions must you make in order to justify the use of these probabilities for predicting what will happen when two dice are tossed together?

2.5. If 10 coins are tossed, what is the probability that all 10 will turn up heads? Interpret this probability.

2.6. The American Experience Table of Mortality gives the death rate of individuals aged 20 as 7.80 per 1000 individuals. If you were aged 20 and desired to insure your life for a period of 1 year for \$1000, what premium should you pay? Does this premium guarantee the insurance company that it will have enough money to pay the claims of the beneficiaries of the twenty-year-olds who die during the year? If the actual premium charged you is greater than that which you have computed, can you account for the difference?

2.7. How much would you be willing to pay to play a fair game in which you receive \$1 for every dot which appears on the roll of one die?

2.8. Evaluate this very popular bit of advice: When you toss a coin to decide who is going to pay the check, let your companion do the calling. "Heads" is called seven times out of ten. The simple law of averages gives the man who listens a tremendous advantage.

2.9. On June 29, 1952, a drawing of the Loteria Nacional took place in Mexico. Prizes in this lottery were as shown in the table.

<i>Number of prizes</i>	<i>Value</i>
1	\$1,000,000
1	100,000
2	50,000
4	25,000
5	20,000
10	10,000
20	5,000
280	1,000
97	800
21	500
198	400
4179	200

Twenty-two thousand tickets were offered for sale at \$200 each. What was the value of each ticket?

The \$ sign indicates the Mexican peso which was valued at about 8 cents in American money in 1954. Thus, the maximum prize was \$80,000 and the price of a ticket was \$16. One-tenth tickets were sold in the lottery for 20 pesos and were entitled to one-tenth of the stated prize. What was the price of a tenth ticket in American money? What was the maximum that this ticket could win? What was the probability that this ticket would win some prize? What was the worth of this ticket? Would you buy this fractional ticket if offered an opportunity to do so?

2.10. A part of the Commissioner's Standard Ordinary Mortality Table is given below.

Age	Number living	Deaths each year	Deaths/1000
18	955,942	2199	2 30
19	953,743	2260	2 37
20	951,483	2312	2 43
21	949,171	2382	2 51
22	946,789	2452	2 59

Let A and B be two individuals with the same or different ages. For example, A may be 20 and B 21.

a. What is the probability that A will live and B will die within the next year? That A will die and B will live? That both A and B will live? That both A and B will die? What rule is used in computing these probabilities?

b. What is the probability that either A or B but not both will live for 1 year? That A or B or both will live for 1 year? What rule is used in computing these probabilities?

c. What is the probability that at least one of three people of ages 20, 21, and 22 will live for one more year? (*Hint*: Find the probability that all three of them will die.)

d. The death rate of individuals ages 20 is 2.43 per 1000 in the CSO mortality table and 7.80 per 1000 in the American Experience Table (see Exercise 2.6).

Can you account for the difference in these death rates? Which table would you prefer to use?

e. Suppose an individual aged 22 is known to have a fairly serious heart disease. What is the probability that he will die within the next year?

f. What assumptions are necessary for the valid construction and application of the mortality table?

2.11. The probability that an individual will die from a particular operation is 1 in 100. A doctor has performed this operation successfully 99 times. Does this mean that the one hundredth operation will be fatal to the patient? Does the relative frequency interpretation of a probability have any meaning for the doctor? For the patient?

2.4 The Binomial Distribution

The previous development of the bead-box and horse-racing problems (Sec. 1.8) ended with these ideas:

1. There are two kinds of errors in reaching decisions, and the best decision rule is based upon the probabilities of making these errors.

2. The probabilities of error need not be estimated from the actual drawing of a large number of samples but may be computed in advance from formulas. The present section is concerned with the necessary probability computations.

The same relationship between the color of the bead (red) and the outcome of a horse race (winner) is retained. The random variable x is the number of red beads (number of winners) in a sample of 10 beads (races). Following Fig. 2.1, the probability model may be diagrammed as in Fig. 2.2.

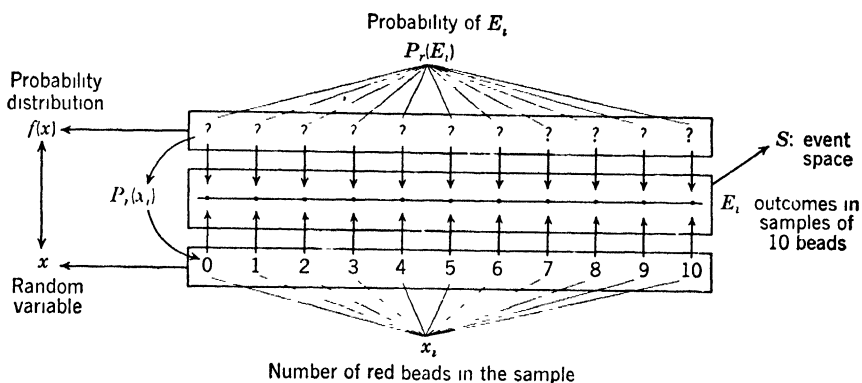


FIG. 2.2. Probability model for the results of sampling 10 beads.

The events E_i are the 11 different possible outcomes in samples of 10 beads shown as points on a line. To each of these events is attached a number which is the value of the variable x_i , the number of red beads in the sample. To each event is also attached a number which is the value of $Pr(E_i)$. Through the common association with the events, $Pr(E_i)$

becomes also $\Pr(x_i)$. The probabilities attached to each x_i make x a random variable, and the x_i and $\Pr(x_i)$ together are the probability distribution expressed by the x and $f(x)$ relationship.

Remark. When the events E_i are the outcomes of a random sampling experiment, the event space of probability theory is called a *sample space* and the probability distribution is called a *sampling distribution*.

It is impossible to enumerate all the points in the fundamental probability set and to collect together those subsets which comprise the events E_i . The process of sampling with replacement (samples of 10) from a box filled with 500 beads will yield 500^{10} different samples and, thus, 500^{10} points in the fundamental probability set. The computation of this number of points is in itself a staggering problem. Fortunately, the probability relationship may be expressed mathematically by the *binomial distribution*:

$$f(x) = \frac{n!}{x!(n-x)!} P^x \cdot Q^{n-x} \quad (2.2)$$

where n = sample size

x = random variable; number of units in the sample having the characteristic studied

P = proportion of units in the population having the characteristic studied

$Q = 1 - P$

$n!$ = n factorial (! is a factorial sign); a factorial is the product of all integers from n (or x) down to unity

$n! = n(n-1)(n-2) \dots (1)$

$x! = x(x-1)(x-2) \dots (1)$

$0! = 1$

$x^0 = 1$

Values of P will be written in decimal form:

$$10\% = .10 \quad 90\% = .90 \quad 100\% = 1.00$$

Example 2.10. Find $\Pr(x = 5)$ in the bead-box experiment.

$$\begin{aligned} n = 10 \quad x = 5 \quad n - x = 5 \quad P = .5 \quad Q = 1 - P = .5 \\ \Pr(x = 5) = f(5) = \frac{10!}{5!5!} (.5)^5 (.5)^5 = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5 \\ = \frac{252}{1024} = .246 \end{aligned}$$

Example 2.11. Find $\Pr(x = 10)$.

$$n = 10 \quad x = 10 \quad n - x = 0 \quad P = .5 \quad Q = .5$$

$$\Pr(x = 10) = f(10) = \frac{10!}{10!0!} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0 = \frac{1}{1024} = .001$$

The numbers computed in Examples 2.10 and 2.11 are probabilities. As such, they may be used to predict the outcomes of a physical experiment in which 10 beads are selected randomly with replacement from a population which contains 50 per cent red beads. If a large number of samples are drawn under these conditions, the relative frequency with which 5 red beads will occur is predicted to be .246; the relative frequency with which 10 red beads will occur is predicted to be .001. As the number of samples increases, the observed relative frequencies should come closer to the probabilities. Interpreted in terms of everyday experience, these predictions mean that the probability is 246 in 1000 that 5 red beads will occur and 1 in 1000 that 10 red beads will occur, under these random sampling conditions.

Remark. The individual terms of the binomial distribution may be derived from the fundamental addition and multiplication theorems. A complete enumeration of the possible results for $n = 2$ and $n = 3$ is presented, together with the probability calculation which is equivalent to the corresponding term of the binomial distribution.

$n = 2$

Sample possibilities	Probability*	Number of red beads	Probability†	Binomial
WW	QQ	0	Q^2	$\frac{2!}{0!2!} P^0Q^2 = Q^2$
WR	QP	1	$2PQ$	$\frac{2!}{1!1!} P^1Q^1 = 2PQ$
RW	PQ	1		
RR	PP	2	P^2	$\frac{2!}{2!0!} P^2Q^0 = P^2$

$n = 3$

WWW	QQQ	0	Q^3	$\frac{3!}{0!3!} P^0Q^3 = Q^3$
RWW	PQQ	1	$3PQ^2$	$\frac{3!}{1!2!} P^1Q^2 = 3PQ^2$
WRW	QPQ			
WWR	QQP			
WRR	QPP	2	$3P^2Q$	$\frac{3!}{2!1!} P^2Q^1 = 3P^2Q$
RWR	PQP			
RRW	PPQ			
RRR	PPP	3	P^3	$\frac{3!}{3!0!} P^3Q^0 = P^3$

* Multiplication theorem.

† Addition theorem.

A shorthand notation will be used for the probabilities computed from the binomial distribution. For example, the probability that x equals 5

if the sample size is 10 and the population percentage is 50 will be written:

$$\Pr(x = 5|n = 10, P = .5)$$

Since the vertical line means "if" or "given that," this notation is read: "the probability that $x = 5$ given that or if $n = 10$ and $P = .5$." Both the sample size and the population percentage P must be specified in order to find the correct probability either by computation or from a table.

Remark. The binomial distribution is the appropriate probability distribution when the sampling is *with replacement*. Does the bead box sample with replacement? No! Watch the beads run out into the tube and you can see that the sampling is without replacement. The bead which runs out first is not put back into the box so that it may be selected again. Technically, probability distributions *with* and *without replacement* differ from each other. However, when the population contains more than roughly twenty times the number of observations in the sample, the two sampling distributions are approximately the same. Since the bead box contains 500 beads and the sample size is only 10, the ratio is 500/10 or 50:1, much greater than the 20:1 rule of thumb.

Computations of the individual terms of the binomial distribution are tedious. There are tables¹ available which give the probabilities for different values of n , x , and P . Several small tables of the binomial distribution are reproduced in Appendix A of this book.

EXERCISES

2.12. Find the value of each of these factorials: $5!$, $4!$, $3!$, $8!$, $0!$, $1!$, $8!/7!$, $8!/5!$.

2.13. Translate into words each of the following mathematical expressions:

- | | | | |
|------------|---------------|----------------------|----------------------|
| a. $x > 5$ | b. $x < 5$ | c. $x \leq 10$ | d. $x \geq 10$ |
| e. $5 > x$ | f. $0 \leq x$ | g. $0 \leq x \leq 5$ | h. $0 \leq P \leq 1$ |

2.14. Write in both percentage and decimal form.

One hundred per cent	Five per cent
Ninety per cent	One per cent
Ten per cent	One-tenth of one per cent

2.15. What is the value of $0!$? Of x^0 ?

2.16. Translate each of the following probability statements.

- $\Pr(x = 4|n = 10, P = .5)$
- $\Pr(x = 2|n = 10, P = .5)$
- $\Pr(x = 7|n = 10, P = .5)$

2.17. For $n = 10$ and $P = .5$, find (a) $\Pr(x = 0)$; (b) $\Pr(x = 1)$; (c) $\Pr(x = 9)$; (d) $\Pr(x = 2)$. Interpret each of these probabilities in terms of the performance

¹ "Tables of the Binomial Probability Distribution," National Bureau of Standards, Applied Mathematics Series, no. 6, Government Printing Office, Washington, 1950.

Harry G. Romig, "50-100 Binomial Tables," John Wiley & Sons, Inc., New York, 1953.

of a physical experiment. Under what conditions may these probabilities be used with validity to predict the outcomes of such an experiment?

2.18. From Table A.2 in the Appendix find the following probabilities:

- a. $\Pr(x = 6|n = 10, P = .3)$
 - b. $\Pr(x = 5|n = 10, P = .25)$
 - c. $\Pr(x = 0|n = 10, P = .05)$
 - d. $\Pr(x = 0 \text{ or } x = 1|n = 10, P = .5)$
 - e. $\Pr(x \geq 8|n = 10, P = .45)$
- (Hint. Use the addition theorem.)

2.19. If two samples are selected in succession and x_1 is the outcome of the first sample and x_2 is the outcome of the second sample, find:

- a. $\Pr(x_1 = 5 \text{ and } x_2 = 5|n = 10, P = .5)$
- b. $\Pr(x_1 = 0 \text{ and } x_2 = 0|n = 10, P = .05)$

What assumptions are made about x_1 and x_2 ? What do each of these probabilities mean?

2.20. From the "Tables of the Binomial Probability Distribution" find the following probabilities:

- | | |
|----------------------------------|-------------------------------------|
| a. $\Pr(x = 10 n = 30, P = .3)$ | f. $\Pr(x \geq 10 n = 30, P = .3)$ |
| b. $\Pr(x = 20 n = 40, P = .5)$ | g. $\Pr(x \geq 20 n = 40, P = .5)$ |
| c. $\Pr(x = 35 n = 45, P = .7)$ | h. $\Pr(x \geq 35 n = 45, P = .7)$ |
| d. $\Pr(x = 0 n = 25, P = .15)$ | i. $\Pr(x = 15 n = 15, P = .15)$ |
| e. $\Pr(x = 25 n = 35, P = .82)$ | j. $\Pr(x \leq 25 n = 25, P = .82)$ |

2.21. From the "50-100 Binomial Tables" find the following probabilities:

- | | |
|----------------------------------|-------------------------------------|
| a. $\Pr(x = 30 n = 80, P = .3)$ | e. $\Pr(x \leq 30 n = 80, P = .3)$ |
| b. $\Pr(x = 40 n = 80, P = .3)$ | f. $\Pr(x \geq 40 n = 80, P = .3)$ |
| c. $\Pr(x = 50 n = 100, P = .5)$ | g. $\Pr(x \leq 50 n = 100, P = .5)$ |
| d. $\Pr(x = 55 n = 100, P = .5)$ | h. $\Pr(x \geq 55 n = 100, P = .5)$ |

2.22. From one table or the other, find the value of x_r which satisfies the following probability statements:

- a. $\Pr(x \leq x_r|n = 100, P = .5) \doteq .044$
- b. $\Pr(x \leq x_r|n = 80, P = .3) \doteq .053$
- c. $\Pr(x \geq x_r|n = 50, P = .2) \doteq .95$
- d. $\Pr(x \geq x_r|n = 100, P = .2) \doteq .975$

2.5 Comparison of Observed Relative Frequencies and Probabilities

The probability distribution of x for $n = 10$ and $P = .5$ may be computed from the individual terms of the binomial distribution or read from Table A.2. How well do these probabilities predict the outcomes of the author's experiment in which 100 samples were drawn from a bead box containing 50 per cent red beads? The results are given in Table 2.1. The observed values differ from the probabilities, but as more samples are selected, the relative frequencies will be numerically closer to the probabilities.

TABLE 2.1. RELATIVE FREQUENCIES AND PROBABILITIES OF THE NUMBER OF OCCURRENCES IN 100 SAMPLES OF 10 FROM A POPULATION WITH $P = .5$

Number of red beads x_i	Observed relative frequency	Probability $\Pr(x_i)$
0	00	0009766
1	02	0097656
2	08	0439453
3	08	1171875
4	23	2050781
5	27	2460938
6	22	2050781
7	06	1171875
8	02	0439453
9	02	0097656
10	00	0009766
Total	1 00	1 0000

From Table 1.5 and Table A.2

2.6 Graphic Presentation of Frequency Distributions

The presentation in tabular form of distributions of the frequency or relative frequency of occurrence of some random variable is generally called a *frequency distribution*. The problem of making such a distribution when the variable can assume only integral values is simple. The integral values are ordered from highest to lowest or lowest to highest and the number of observations of each integer recorded in the proper row. A simple way to construct a frequency distribution is to use a tally sheet. First, order the integral values of the random variable. Then, record the number of occurrences of each value by making a small check opposite it after each occurrence.

Example 2.12. Suppose the sequence of the number of red beads starts out: 1, 0, 2, 3, 1, 0, 1, 2, 0, 1, 1, The tally sheet will look something like this, the fifth tally for $x = 1$ being a bar across the other four tallies.

$x_i(\text{red})$	Tally
0	
1	
2	
3	
.	.
.	.
.	.

Frequency distributions may be graphed in several forms. Two very simple forms are the *histogram*, or *bar chart*, and the *frequency polygon*. The histogram is usually constructed by plotting the random variable on

the horizontal axis (the *abscissa*) and the frequency on the vertical axis (the *ordinate*) of arithmetic graph paper ruled with small squares. The scale chosen depends upon the number of values of the variable, the width of the paper, and the size of the histogram to be constructed. Each abscissa segment is labeled with a value of the variable, with the lowest value to the left and increasing values to the right. This is the conventional method in algebraic scaling. The ordinate is ruled according to frequencies, from 0 to 30 in the present illustration. The choice

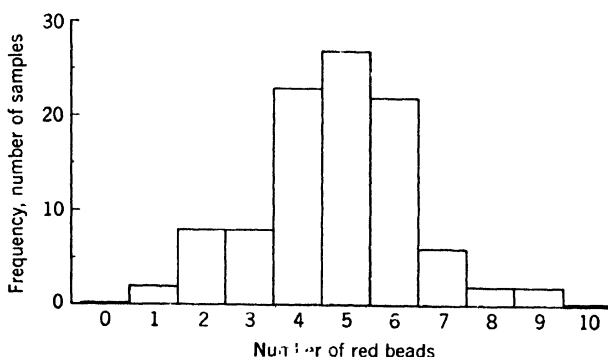


FIG. 2.3. Histogram of the observed frequencies of Table 1.3.

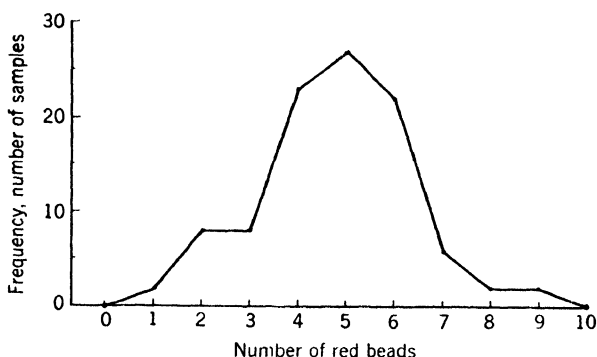


FIG. 2.4. Frequency polygon of the observed frequencies of Table 1.3.

of a scale for the ordinate depends upon much the same factors as the choice of a scale for the abscissa. The histogram is constructed by erecting bars above the values of the random variable. The length of each bar is equal to the frequency of occurrence and the width is equal to the abscissa scale unit (see Fig. 2.3).

A frequency polygon is constructed from the same basic scales as the histogram. For each value of the variable a single point is plotted in the *middle* of the abscissa scale unit at a height corresponding to the frequency. The several points are connected by straight-line segments to complete the polygon (see Fig. 2.4).

Histograms, bar charts, and frequency polygons are valuable devices to represent a frequency distribution graphically. The reader can see quickly which values occur more often, what are the greatest and least values, and whether there is a tendency for the distribution to "pile up" at some values near the middle of the distribution or around another point. He can also gain an idea of the variation in the observations from the spread of the distribution.

Remark 1. In the bar chart (Fig. 2.3) the area of a bar is proportional to the number of observations of the random variable which that bar represents. The total area under the bars represents the total number of observations. Since the widths of the bars are equal and the area of a bar is the product of the width times the height, the length of each bar is proportional to the number of observations.

Remark 2. In the frequency polygon (Fig. 2.4) the distance from the plotted point to the horizontal axis is proportional to the number of observations. The total area under the polygon is proportional to the total number of observations if the polygon is closed off at each end.

2.7 Numerical Characteristics of Distributions

Frequency and probability distributions have two important numerical characteristics: (1) the average value and (2) the variation of the individual observations around this average. A measure of each of these characteristics of the binomial distribution will now be discussed.

Average Number of Red Beads per Sample. The most common average is the arithmetic mean. The basic data for both the observed distribution and the probability distribution are given in Table 2.2.

Remark. The probabilities in Tables 2.1 and 2.2 are reported to 7 decimal places only to show that the total probability is exactly 1 and the mean exactly 5. In practical problems, probabilities given to 2 or 3 decimals are satisfactory, but in this book, more decimal places will often be carried to allow the student to check accurately all computations.

Observed Frequency Distribution. The symbol for the mean of an observed distribution (or of a sample) is $\bar{X}$. Its formula is

$$\bar{X} = \frac{\sum f_i x_i}{N}$$

where $N = 100$ samples

x_i = number of red beads in each sample

The operations are shown in Table 2.2, and the mean number of red beads in the 100 samples is 4.77.

Probability or Sampling Distribution. The symbol for the mean of a probability distribution is the Greek letter mu (μ). Its formula is the same as that for the expected value, given in Sec. 2.3. The mean number of red beads per sample in the sampling distribution is 5.00.

Since both numbers are averages, why distinguish between them? Why use the symbol $\bar{X}$ for one and the symbol μ for the other? This is done because there is one essential difference between them. $\bar{X}$ is computed from *sample* observations and, therefore, *varies* from sample to sample or, as in this problem, from one group of samples to another,

TABLE 2.2. COMPUTATION OF THE MEAN NUMBER OF RED BEADS PER SAMPLE

Observed distribution			Probability distribution	
Number of red beads x_i	Frequency f_i	$f_i x_i$	Probability of x_i $\Pr(x_i)$	$x_i \Pr(x_i)$
0	0	0	0009766	0000000
1	2	2	0097656	0097656
2	8	16	0439453	0878906
3	8	24	1171875	3515625
4	23	92	2050781	8203124
5	27	135	2460938	1 2304690
6	22	132	2050781	1 2304686
7	6	42	1171875	8203125
8	2	16	0439453	3515624
9	2	18	0097656	0878904
10	0	0	0009766	0097660
Total	100	477	1 0000000	5 0000000

$$\bar{X} = 477/100 = 4.77$$

$$Ex = \mu = 5.00$$

because the chance element involved in selecting the observations for the samples will make the $\bar{X}$'s different. The mean of the probability distribution is a *constant*. The symbol $\bar{X}$ will always be used for a sample mean and the symbol μ for the mean of either a sampling distribution (as here) or a population (later in the course).

The mean of the binomial probability distribution may be computed less laboriously than in Table 2.2. If P is the population percentage and n is the sample size, the mean of the binomial distribution is

$$\mu = nP \quad (2.3)$$

For example, with $P = .5$ and $n = 10$, $\mu = 10(.5) = 5$, which agrees with the result of Table 2.2.

Remark. An illustration of the derivation of Eq. (2.3) for the special case when $n = 3$ follows. The probability values are taken from the remark in Sec. 2.4 which discussed the basic derivation of the binomial distribution.

Number of red beads x_i	Probability $\Pr(x_i)$	$x_i \cdot \Pr(x_i)$
0	Q^3	$0 = 0$
1	$3PQ^2$	$3PQ^2 = 3P(Q^2)$
2	$3P^2Q$	$6P^2Q = 3P(2PQ)$
3	P^3	$3P^3 = 3P(P^2)$
Total	$(Q + P)^3$	$3P(Q + P)^2$

From Eq. (2.1), $Ex = \mu = \sum x_i \cdot \Pr(x_i)$. $Ex = \mu = 3P(Q + P)^2 = 3P = nP$, since $(Q + P) = 1$ and $n = 3$.

2.8 Variability in the Number of Red Beads

Not all samples have the same number of red beads, nor are they expected to have. Both the observed and expected numbers vary about the mean number in Table 2.2. This variation is called *dispersion* and one measure of it is the *variance*. The data for the sampling distribution are shown in Table 2.3. Each value of x_i is measured as a deviation from the mean μ . This deviation is then squared and multiplied by the probability of occurrence. The total of these products is the variance. The variance of a sampling distribution (or a population) is given the symbol σ^2 (sigma squared). In Table 2.3, σ^2 is 2.5000.

A related measure of dispersion is the *standard deviation*. The standard deviation is the square root of the variance and is written σ (sigma). In Table 2.3, σ is 1.581.

The meaning of the standard deviation (or variance) is derived from its application and what it tells about the distribution. In general, it measures the dispersion or variation of a set of observations about the mean value of the observations. Definite uses will be given for it later, but for the present certain ideas may be helpful. First, σ equal to zero means that all the observations must equal the mean. Therefore, all the observations must be the same. There is *no variation* or dispersion. Second, the larger the σ for the same random variable, the greater is the dispersion. The emphasis is upon *the same random variable* for good reason. If for one distribution of, say, the heights of students, the standard deviation is 1 inch and for a second distribution it is 1.5 inches, the second distribution is more variable than the first. A third distribution with σ equal to 2 inches would be still more variable. Suppose that σ equals 10 pounds for the distribution of weights of the students when the standard deviation of their heights is 2 inches. Certainly 10 is a larger number than 2. Can it be said that weights are more variable than heights? No, because the units of measure are different. The 10

is in pounds; the 2 is in inches. Only when applied to the same random variable does a larger standard deviation indicate a greater variation.

TABLE 2.3. COMPUTATION OF THE VARIANCE AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION
($n = 10$, $P = .5$, $\mu = 5$)

Number of red beads x_i	Probability of x_i $\Pr(x_i)$	$x_i - \mu$	$(x_i - \mu)^2$	$\Pr(x_i)(x_i - \mu)^2$
0	0009766	-5	25	0244150
1	0097656	-4	16	1562496
2	0439453	-3	9	3955077
3	1171875	-2	4	4687500
4	2050781	-1	1	2050781
5	2160938	0	0	0000000
6	2050781	+1	1	2050781
7	1171875	+2	4	4687500
8	0439453	+3	9	3955077
9	0097656	+4	16	1562496
10	0009766	+5	25	0244150
Total	1 0000000	0		2 5000008

$$\sigma^2 = 2\ 5000 \quad \sigma = \sqrt{2\ 5000} = 1\ 581$$

Third, students are warned against misunderstanding the concepts of variance and standard deviation because they cannot employ any of the usual forms of evaluating measurements; i.e., there are no standards of weight, distance, or volume which are applicable. Yet, these words are a necessary part of a statistical vocabulary and their meaning must be appreciated. Finally, the percentage of the total number of observations within certain limits above and below the mean is related to the standard deviation. Exact relationships are given in Sec. 4.4.

The variance and standard deviation of the binomial distribution may be expressed in relatively simple formulas involving n , P , and Q , where $Q = 1 - P$.

$$\sigma^2 = nPQ \quad (2.4)$$

$$\sigma = \sqrt{nPQ} \quad (2.5)$$

The variance and the standard deviation of the binomial distribution with $n = 10$ and $P = .5$ are

$$\sigma^2 = 10(.5)(.5) = 2.5 \quad \sigma = \sqrt{10(.5)(.5)} = 1.581$$

These values agree with Table 2.3. Equations (2.3) to (2.5) are neat short-cut methods of finding μ , σ , and σ^2 of the binomial distribution. They are simply means of arriving at these important constants without the laborious arithmetic involved if they are computed from a table show-

ing the whole probability distribution. Because they are so neat and simple, they are easy to learn and use.

EXERCISES

2.23. In part (c) of Exercise 1.4, the frequency distribution of the results of the bead-box sampling experiment was made. These exercises pertain to this experiment.

a. Plot the histogram and frequency polygon of this frequency distribution.

b. Construct the relative frequency distribution of the observed number of red beads.

c. Compute the probability values of the binomial distribution for $n = 10$, $P = .2$, $x = 0$, and $x = 1$. Complete a table of probabilities of the binomial distribution for $x = 2$ to $x = 10$, using Table A.2 of Appendix A.

d. Compare the observed relative frequencies of (b) and the probabilities of (c). How well do the probabilities predict the relative frequencies?

e. Compute $\bar{X}$ for the observed frequency distribution. Compute μ , σ^2 , and σ , using Eqs. (2.3) to (2.5). Compare $\bar{X}$ with μ .

f. Plot in polygon form the probabilities for $n = 10$, $P = .5$ (Table 2.1) and $P = .2$ (this exercise) on the same sheet of graph paper, and compare them. Place one polygon on top of the other.

2.24. Record the following information on a sheet of paper:

a. Your age as of your last completed birthday; i.e., record 21 years, 8 months, as 21.

b. Your weight to the nearest pound.

c. Your height to the nearest half inch.

These data should be assembled for the whole class and kept on one large sheet for use in problems in later chapters.

2.25. a. Find μ , σ^2 , and σ for $n = 25$ and $n = 100$ and the following values of P : .1, .2, .3, .4, .5, .6, .7, .8, .9.

b. For the same n , at which value of P are σ^2 and σ the largest?

c. For the same value of P , what happens to σ^2 and σ as n is increased from 25 to 100?

d. For which value of P is the variability in the distribution the largest?

SUPPLEMENTARY READING

Cramer, H.: "Mathematical Methods of Statistics," Princeton University Press, Princeton, N.J., 1946. Chapter 13 is a readable account of statistics and probability in one of the top-level books on mathematical statistics.

Feller, W.: "An Introduction to Probability Theory and Its Applications," vol. 1, John Wiley & Sons, Inc., New York, 1950. The introduction is recommended for all students and Chap. 1 for the student with an interest in symbolic logic.

Levinson, H. C.: "The Science of Chance," Rinehart & Company, Inc., New York, 1950. The whole book is a popular presentation of probabilities in many situations.

Maroney, M. J.: "Facts from Figures," Penguin Books, Inc., Baltimore, 1953. The whole book is recommended reading, but Chaps. 1, 2, and 20 are relevant now.

CHAPTER 3

STATISTICAL INFERENCE IN THE DISCRETE CASE: BINOMIAL DISTRIBUTION (I)

Two experiments have been performed with the bead box: one when the box contained 50 per cent red beads and one when it contained 20 per cent red beads. A third experiment is now to be performed under these conditions: (1) the box is again presented to you and you are told that it contains either 50 per cent or 20 per cent red beads; (2) you may not look into the box but you may select 10 beads at random by thoroughly mixing the contents and allowing the 10 beads to run out into view; and (3) on the basis of the number of red beads observed in the sample, you are to decide whether the box contains 50 per cent or 20 per cent red beads.

3.1 A Decision Rule for Deciding between Two Alternative Hypotheses

A more formal statement of this problem follows:

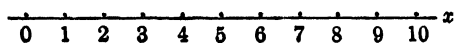
H_0 : The hypothesis to be tested; $P_0 = .50$ (50 per cent of beads are red)

H_1 : The alternative hypothesis; $P_1 = .20$ (20 per cent of beads are red)

Test Procedure. Select 10 beads at random from the box and count the number of red beads.

Decision Rule. The problem is to decide upon a rule which will unambiguously determine whether to accept H_0 or H_1 .

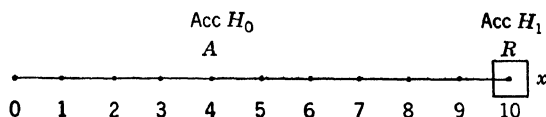
The number of red beads in the sample is a random variable which may take on any one of 11 values, 0, 1, . . . , 10. This set of values is called the *sample space*; each value of the random variable in the sample space is called a *sample point*. The sample space may be diagramed as points on a line:



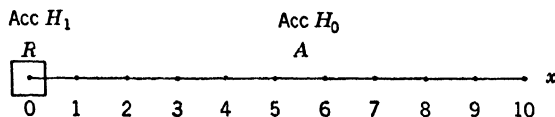
The task now is to decide on the basis of evidence falling in the sample

space whether to accept H_0 or H_1 . The decision rule must then state for which samples H_0 is to be accepted (that is, H_1 is rejected) and for which sample points H_1 is to be accepted (that is, H_0 is rejected). The sample space must be divided into two areas: (1) an area R which accepts H_1 and (2) an area A which accepts H_0 . Thus, the decision rule is one which attaches to each sample point a decision to accept H_0 or H_1 and, by implication, to reject the other alternative. R is called the *region of rejection* and A the *region of acceptance*.

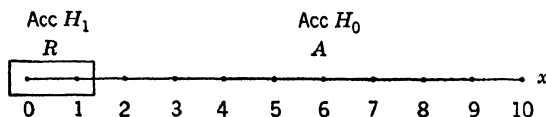
For example, let R consist of the single point $x = 10$. It is blocked off as shown below. If the sample of 10 beads contains 10 red beads, H_0 is rejected and H_1 is accepted; if 0, 1, . . . , 9 red beads are observed, H_0 is accepted and H_1 is rejected.



If R is the point $x = 0$, H_1 is accepted if the sample contains 0 red beads.



If R consists of the two points $x = 1$ and $x = 0$, H_1 is accepted if either 0 or 1 red bead is observed in the sample:



The method of setting up a statistical decision rule is to divide the sample space into two regions: R and A . Sample points in region R lead to the decision to accept H_1 and reject H_0 . Sample points in the region A lead to the decision to accept H_0 and reject H_1 .

The sampling distribution (probability distribution) of the random variable x (the number of red beads in the sample) is the binomial distribution,

$$f(x) = \frac{n!}{x!(n-x)!} P^x Q^{n-x} \quad (3.1)$$

The probability that x will assume the values 0, 1, . . . , 10 for the value

of P under each hypothesis may be computed from this probability law (see Table A.2). The probabilities are reproduced in Table 3.1.

Table 3.1 clearly shows (except for rounding errors) some probability that each sample point may occur under each hypothesis. For example, if the box is filled with 50 per cent red beads, the probability is .044 that 2 red beads will occur in the sample; if it is filled with 20 per cent red beads, the probability is .302 that 2 red beads will occur. Because the occurrence of each value of x is consistent with each hypothesis, a decision to accept H_0 or accept H_1 may be in error. The fact that some

TABLE 3.1. BINOMIAL PROBABILITIES FOR $n = 10$, $P = .5$ AND $P = .2$

x	$Pr(x n = 10, P = .5)$	$Pr(x n = 10, P = .2)$
0	.001	.107
1	.010	.268
2	.044	.302
3	.117	.201
4	.205	.088
5	.246	.026
6	.205	.006
7	.117	.001
8	.044	.000*
9	.010	.000*
10	.001	.000*

* These probabilities are zero to three decimal places.

sample points (for example, $x = 2$) are more likely under one hypothesis than the other suggests that the values of x for which the probability of occurrence of the sample is small under H_0 and large under H_1 should lead to the acceptance of H_1 ; the values of x for which the probability of occurrence of the sample is small under H_1 and large under H_0 should lead to the acceptance of H_0 . That is to say, the statistical-decision rule should be based upon a selection of R and A in the sample space so as to ensure that:

1. The probability of accepting H_1 (rejecting H_0) when, in fact, H_0 is correct is small.
2. The probability of accepting H_0 (rejecting H_1) when, in fact, H_1 is correct is small.

3.2 Errors in Testing Hypotheses

The previous section states that errors in accepting H_0 or H_1 result from the probability that an observed sample point may occur under each hypothesis. If H_0 is correct, the preferable decision is to accept H_0 and reject H_1 ; if H_1 is correct, the preferable decision is to accept H_1

and reject H_0 . The following four situations are then found to be possible:

I: H_0 is true, and the decision is to accept H_1 .

II: H_1 is true, and the decision is to accept H_0 .

III: H_0 is true, and the decision is to accept H_0 .

IV: H_1 is true, and the decision is to accept H_1 .

These four situations may be presented in tabular form:

Decision	Correct hypothesis	
	H_0	H_1
Outcome		
Acc H_0	Acceptable	Error
Acc H_1	Error	Acceptable

Two of the situations represent acceptable decisions; two situations are in error. These two errors differ from each other. The error of accepting H_1 when H_0 is true is qualitatively different from the error of accepting H_0 when H_1 is true, and very rarely are they of equal importance.

Remark. These errors are well illustrated by the following oversimplified decision problem. Suppose a man is on trial for murder and a jury has to decide upon his guilt.

The jury decision may be	The man may be	
	Guilty	Innocent
Guilty	O.K.	Error
Innocent	Error	O.K.

Are these errors the same? Which one has Western law deemed most important to avoid? Are all systems of law like this? Is there any noticeable change in the Western law attitude toward these errors recently?

In order to distinguish between these two errors and to provide a neat shorthand method of referring to them in the future, they are given numbers. The error of accepting H_1 when, in fact, H_0 is correct is called error I; the error of accepting H_0 when, in fact, H_1 is correct is called error II. These may be written as follows:

$$\text{I: Acc } H_1 | H_0 \text{ correct} \quad (3.2)$$

$$\text{II: Acc } H_0 | H_1 \text{ correct} \quad (3.3)$$

The probabilities of making these two errors are symbolized by α (alpha) and β (beta). α is the probability associated with I: accepting H_1 when H_0 is correct. β is the probability associated with II: accepting H_0 when H_1 is correct. In a shorthand notation, α and β are written:

$$\alpha = \Pr(\text{Acc } H_1 | H_0 \text{ correct}) \quad (3.4)$$

$$\beta = \Pr(\text{Acc } H_0 | H_1 \text{ correct}) \quad (3.5)$$

Remark. Errors I and II, and α and β , must be kept straight in order that each of you will be talking about the same thing when discussing problems in the future.

The suggestion in Sec. 3.1 may now be restated, namely, that the statistical decision rule be based upon a selection of R and A in the sample space so as to ensure that:

1. The probability of accepting H_1 when, in fact, H_0 is correct is small.

2. The probability of accepting H_0 when, in fact, H_1 is correct is small.

The statistical-decision rule should be based upon a selection of R and A in the sample space so as to ensure that α and β are small. Since the probabilities of making the two errors work against each other (one becomes larger as the other is made smaller), it may not be possible in a given test situation to make both α and β small simultaneously. (See Exercise 1.10 for an earlier demonstration of this.) In selecting the proper statistical decision rule, therefore, the procedure is to *fix* α at some small preassigned value, and then to select a region R which *minimizes* β . Thus, the probability (α) of accepting H_1 when H_0 is correct is controlled at some small level; the probability (β) of accepting H_0 when H_1 is correct is minimized.

3.3 A Decision Rule for the Bead-box Problem

The third experiment with the bead box (Sec. 3.1) consists of testing the following hypotheses:

$$H_0: P_0 = .5 \quad n = 10$$

$$H_1: P_1 = .2 \quad x = \text{number of red beads in sample}$$

Sample points in the region R of the sample space will lead to the decision to accept H_1 and in the region A to the decision to accept H_0 . R is called the region of rejection (points in it lead to rejecting H_0); A is called the region of acceptance (points in it lead to accepting H_0). What values of x should comprise the region R ? That is to say, what sample values should lead to the acceptance of H_1 ? Does it make sense to accept H_1 if 10 or 9 red beads are observed in the sample? Samples with 10 or 9 red beads are more likely to occur if $P = .5$ than if $P = .2$.

What values of x should comprise the region A and thus lead to the acceptance of H_0 ? Does it make sense to accept H_0 if 0 or 1 red beads occur in the sample? Samples with 0 or 1 red bead are more likely to occur if $P = .2$ than if $P = .5$. It seems to make more sense if values of x like 0 and 1 are put in R (accept H_1) and values of x like 10 and 9 are put in A (accept H_0). The region of rejection R is placed in the left-hand side of the sample space when it is diagramed as a horizontal straight line; the region of acceptance A is placed in the right-hand side of the sample space. These are referred to as the *lower tail* and *upper tail* of the sample space, respectively. Placing the region R in the lower tail and A in the upper tail of the sample space in this problem minimizes β .

Granted that R should be in the lower end of the sample space, how many values of x should it include? The other part of the general rule—to fix α —is the determining factor. Successive values of x beginning at the lower end of the sample space are included until the *sum of their probabilities (if H_0 is true) is equal to α* . The process of “fixing alpha” means the arbitrary selection of some probability of committing this error I. The value chosen for this problem is $\alpha = .055$. The region R is built up as follows:

1. Include in R the point $x = 0$, the smallest value of x in the lower tail. $\Pr(x = 0|n = 10, P_0 = .5) = .001$ (from Table A.2) rounded to three digits. This probability is less than .055 (α). Therefore, add another value of x to R .

2. Include $x = 1$ in the region R . $\Pr(x = 1|n = 10, P_0 = .5) = .010$. The sum of the probabilities is $.001 + .010 = .011$ (addition theorem). This probability is less than .055 (α). Therefore, include another value of x in R .

3. Include $x = 2$ in R . $\Pr(x = 2|n = 10, P_0 = .5) = .044$. The sum of the probabilities is $.011 + .044 = .055$. This is exactly equal to α .

4. Include $x = 3$ in R . $\Pr(x = 3|n = 10, P_0 = .5) = .117$. The sum of the probabilities is $.055 + .117 = .172$. This probability is greater than α .

5. Stop at the end of step (3) and include only the values $x = 0, 1$, and 2 in the region of rejection R . A consists of the points 3 to 10.

Statistical-decision Rule. If 2 or fewer red beads occur in the sample, reject H_0 (50 per cent red) and accept H_1 (20 per cent red); otherwise, accept H_0 (50 per cent red).

This rule unambiguously determines the decision which will be reached after selecting a random sample. The probability of committing error I of accepting H_1 when H_0 is true is controlled at .055; the probability of making error II of accepting H_0 when H_1 is true is minimized.

EXERCISES

3.1. Write out the meaning of each of the following in terms of accepting and rejecting H_0 and H_1 : error I; α ; error II; β .

3.2. Translate into shorthand notation each of the following statements:

a. The probability that x will equal 8 in a sample of 10 from a population with a proportion of 50 per cent.

b. The probability that x will be less than 8 in samples of 10 from a population with a proportion of 20 per cent.

3.3. Write out in words the meaning of each of these statements:

a. $\Pr(x = 15 | n = 25, P = .7)$

b. $\Pr(x \leq 10 | n = 20, P = .1)$

c. $\Pr(x \geq 4 | n = 5, P = .5)$

3.4. Given these conditions for a statistical test:

$$\begin{array}{ll} H_0: P_0 = .5 & n = 100 \\ H_1: P_1 = .2 & R: x \leq 10 \end{array}$$

a. Write in words the decision rule.

b. Write in shorthand notation the meaning of α and β .

c. If a value of $x = 15$ is observed, what decision is reached?

3.5. Given these conditions for a statistical test:

$$\begin{array}{ll} H_0: P_0 = .5 & n = 10 \\ H_1: P_1 = .2 \end{array}$$

a. For six regions of rejection (0), (10), (0,1), (9,10), (0,1,2), (8,9,10), find values for α and β , and enter them in a table.

b. Indicate the best region of rejection if α is allowed to be no more than .001; .011; .055. Why?

c. Are the regions of rejection chosen in part (b) in the lower end of the sample space?

3.4 Solution to the Horse-racing Problem

Statistical hypotheses deal with probability distributions of random variables. Most hypotheses in the real world are nonstatistical. In the horse-race problem of Example 1.1, for instance, the hypothesis that the friend has a "good system" or a "winning system" is not really statistical. It can, however, be translated into a precise hypothesis that the system will pick 50 per cent winners. The statistical test of this hypothesis includes selecting 10 races at random from a large group of races, applying the system to each of these races, and counting the number of winners. The hypothesis that the system will pick 50 per cent winners can be tested statistically because (1) the above experiment determines a value of a random variable x , the number of winners; (2) the random

variable x has a probability distribution described mathematically by the binomial distribution; and (3) the hypothesis that the system selects 50 per cent winners is a hypothesis about the parameter P of a binomial distribution.

One statistical hypothesis to be tested is $H_0: P_0 = .50$. Against what alternative should this hypothesis be tested? The alternative hypothesis is that the friend's system is a "bad" or a "losing system." Does this preclude any values of P ? The system is bad only if the percentage of winners is less than 50. If the system will pick 60 or 70 per cent winners, it is not a losing system. Therefore, only values of P less than 50 per cent are interesting as worth detecting. This is written $H_1: P_1 < .50$. The two possible decisions are to accept H_0 or accept H_1 . The actions consequent upon these decisions are:

<i>Decision</i>	<i>Action</i>
Acc H_0	Invest in the scheme
Acc H_1	Do not invest in the scheme

Since either of these decisions may be wrong when based upon sample observations, it follows that the actions may also be wrong. The two errors in decision, in terms of the statistical hypotheses, are:

- I: Acc $P < .50$ when, in fact, $P = .50$
 II: Acc $P = .50$ when, in fact, $P < .50$

In terms of the nonstatistical hypotheses these errors are:

I: Accept as a losing system one which is, in fact, a winning system.

II: Accept as a winning system one which is, in fact, a losing system.

In terms of the actions to be taken, the consequences of these two errors are:

I: Refuse to invest money in a winning system.

II: Invest money in a losing system.

Remark. The cost of investing in a losing system is a loss of actual cash. Is a cost also involved in refusing to invest in a winning system? Is a cost involved in missing an opportunity? Earlier it was stated that the two errors which can be made are qualitatively different from each other. Do you consider the costs of the two errors in this problem the same or different?

The horse-race test may be formally set out as follows:

$$\begin{array}{ll} H_0: P_0 = .50 & n = 10 \\ H_1: P_1 < .50 & R: ? \\ & x = \text{number of winners} \end{array}$$

The problem is to select the appropriate statistical decision rule or those values of x which, if they occur, will lead to an acceptance of H_1 and the

decision not to invest, i.e., the region R . What values of x should be included in the region of rejection R ? Should H_1 be accepted and money not invested if 10 winners are observed? 9 winners? When this problem was first introduced in Sec. 1.1, most students decided to invest (accept H_0) if 5 or more races were won. In other words, the region of acceptance included values of $x = 5, 6, 7, 8, 9$, and 10; the region of rejection included values of $x = 0, 1, 2, 3$, and 4. Again, in Sec. 3.3, it seemed sensible to put the region of rejection in the lower tail of the sample space in order to minimize β . The region of rejection included small values of x : 0, 1, 2, etc. When the alternative hypothesis is stated in the form $P_1 < P_0$ (i.e., it does not specify exactly a value of P but only that P_1 is less than P_0), the region of rejection R is put in the *lower end* of the sample space to *minimize* β .

Granted that the region R should be in the lower end of the sample space, how many values of x should be included in R ? Values of x are included in R until the sum of their probabilities of occurrence, if H_0 is true, is equal to but no greater than α . The decision as to the size of α is more or less arbitrary. Its magnitude depends upon the seriousness and the costs involved in making error I. Here the value arbitrarily selected is $\alpha = .055$.

The horse-race test may be summarized:

$$\begin{array}{lll} H_0: P_0 = .50 & n = 10 & R: x \leq 2 \\ H_1: P_1 < .50 & \alpha = .055 & \end{array}$$

Decision Rule. Accept H_1 and do not invest in the scheme if two races or fewer are won; accept H_0 and invest in the scheme if three races or more are won.

The region of rejection is found by including $x = 0$, then $x = 1$, then $x = 2, \dots$, until α is satisfied. This may be shown in a table:

x	$\Pr(x n = 10, P_0 = .5)$ (from Table A 2)	Cumulative probability
0	.001	.001 (less than α)
1	.010	.011 (less than α)
2	.044	.055 = α (stop here)
3	.117	.172 (greater than α)

The region of rejection R includes values of x equal to 0, 1, and 2—values equal to or less than 2. It may be written: $R: x \leq 2$. In general, a region of rejection in the lower tail of a sample space is written $R: x \leq x_r$, the subscript r standing for “reject,” meaning reject H_0 and accept H_1 .

This decision rule ensures that:

1. The probability is .055 (α) of not investing money in the scheme when the claim is true ($P_0 = .5$).
2. The probability (β) of investing money when the claim is not true ($P_1 < .5$) is smaller than for any other rule with the same α and n .

Remark. Some may think that β is small because it has been minimized. This is not a correct assumption β may be large, although it is smaller than it would be for another region of rejection.

The problem with which the subject matter of statistics was first introduced has been partially solved. The number selected for the statistical decision rule (2) is less than the number most commonly chosen by students (4). If the number 4 were selected, the probability of missing the opportunity to invest in a good system would be .377. The author is unwilling to accept this large a probability of not investing in the scheme when, in fact, it selects 50 per cent winners. If this large a probability is acceptable, the decision rule with $R: x \leq 4$ is a good one.

3.5 Operating Characteristics of a Decision Rule

With the statistical decision rule to invest if three or more winners occur and not to invest if two or fewer winners occur, the probability of investing depends upon the true percentage of winners which the system will select. This percentage is unknown, but if it is less than .5, the correct decision is not to invest. The probability that money will be invested if P is less than .5 depends upon the exact value of P_1 . The values of β which are associated with different values of P_1 form the *operating-characteristic curve*. The height of this curve shows the probability of accepting H_0 for different admissible values of P_1 under H_1 . The value of β varies with the true value of P_1 . How it varies with P is shown on the OC curve.

The OC curve is most important in tests of statistical hypotheses. A friend comes along with a wild scheme to bet on race horses, making a specific claim about the percentage of winners ($P_0 = .5$). A scientific attitude toward such a proposition dictates a test of the claim in which the system is applied to a random sample of races and the decision to invest or not to invest made only after the results are observed. The actual percentage of winners under the scheme is unknown. What if it should select winners only 40 per cent of the time? With a rule to accept H_0 and invest if three or more races are won, what is the probability of deciding to invest with this percentage of winners? If the system selects only 10 per cent winners, what is the probability of investing in the scheme?

The computation of points on the OC curve is fairly simple when a

table of binomial probabilities is available (like Table A.2). In general, β is the probability of accepting H_0 when H_1 is true. H_0 is accepted when values of x from 3 to 10 are observed in the sample. Therefore,

$$\begin{aligned}\beta &= \Pr(\text{Acc } H_0 | H_1 \text{ true}) \\ \beta &= \Pr(x = 3 \text{ to } 10 | n = 10, P_1)\end{aligned}$$

By the *addition theorem*, β is the *sum* of the separate probabilities for $x = 3$ to $x = 10$, given a particular value of P_1 . The exact probabilities for the OC curve of the horse-racing problem are computed in Table 3.2.

The probability of accepting H_0 and investing in the system if the system selects winners 40 per cent of the time is .833; if the system selects winners only 10 per cent of the time, the probability of investing is only .07. The probabilities of investing for other values of P_1 are given in

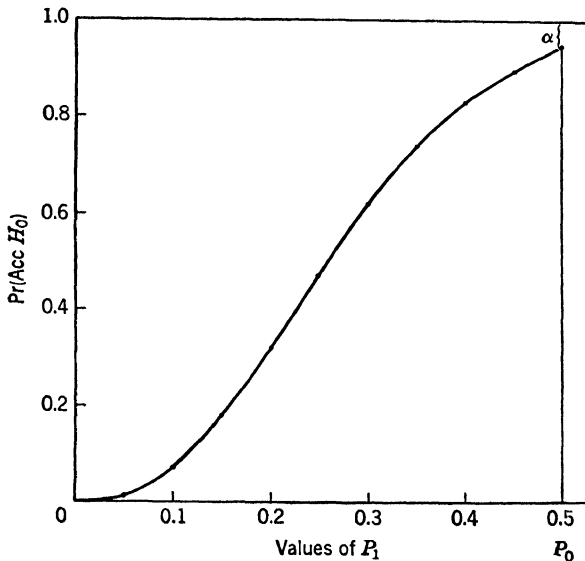


FIG. 3.1 Operating-characteristic curve of the horse-race test.

Table 3.2. A graph of the OC curve is shown in Fig. 3.1. The height of the curve is the probability of accepting H_0 for different admissible values of P_1 under H_1 . The graph dramatically shows that the chance of investing becomes smaller as the true, but unknown, percentage of winners becomes small, until finally, if the system will not pick any winners ($P = 0$), there is no chance of investment.

The OC curve is the plan of operation which was stressed in the early part of Chap. 1. It is not enough to make a decision rule based upon a claim H_0 and a more or less arbitrary value of α associated with error I,

It is true that β will be minimized in such a procedure, but since "minimum" is not synonymous with "small," β may be quite large. The probability of deciding to accept H_0 when the alternative hypothesis is correct is known only through the OC curve. *The plan of operation should cover a large number of alternative possibilities for the value of P_1 .* The OC curve does just this.

It should be emphasized that statistical tests are probability tests. The region of rejection which unambiguously determines the decision depends upon the arbitrarily selected value of α . To this extent, then, the decision rule is arbitrary. Two different individuals can very well arrive at two different decision rules because they evaluate differently

TABLE 3.2. COMPUTATION OF THE OPERATING-CHARACTERISTIC CURVE OF THE HORSE-RACE TEST

$$\begin{array}{lll} H_0: P_0 = .5 & n = 10 & \alpha = .055 \\ H_1: P_1 < .5 & R: x \leq 2 & A: x \geq 3 \end{array}$$

Region of accepting H_0 $x =$	Probability of x for $P_1 =$									
	.00	.05	.10	.15	.20	.25	.30	.35	.40	.45
3	.000	.010	.057	.130	.201	.250	.267	.252	.215	.166
4	.000	.001	.011	.040	.088	.146	.200	.238	.251	.238
5	.000	.000	.001	.008	.026	.058	.103	.154	.201	.234
6	.000	.000	.000	.001	.006	.016	.038	.069	.111	.160
7	.000	.000	.000	.000	.001	.003	.009	.021	.042	.075
8	.000	.000	.000	.000	.000	.000	.001	.004	.011	.023
9	.000	.000	.000	.000	.000	.000	.000	.001	.002	.004
10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
$\beta =$	.000	.011	.069	.179	.322	.473	.618	.739	.833	.900

the seriousness of error I. For example, a person who thinks that missing the opportunity of investing (when, in fact, the system selects 50 per cent winners) is most serious would set a very low value on α , say, .01 or less. He should recognize that at the same time he will run a larger risk of investing in a bad scheme. The person to whom a lost opportunity to invest is less important than the actual cash loss associated with investing in a bad scheme should select a larger value of α , say, .10 or more. The actual probabilities of committing error II can be obtained from the OC curve which is appropriate for the selected value of α . This is an individual choice which must be made according to a subjective evaluation of the cost of making errors. It is not a problem which the statistician can solve. He can only point out that statistical tests hinge upon these probabilities and ask for a selection of α according to individual judgments.

3.6 Selection of Decision Rule for Different Alternatives

It is now possible to generalize the analysis to cover all possibilities which may arise in practice. For testing H_0 against a *one-sided* alternative $H_1: P_1 < P_0$, the region R which leads to the acceptance of H_1 (and the rejection of H_0) should be put in the *lower tail* of the sample space. The region R is built up by including $0, 1, \dots, x_r$ until α is satisfied. Where should the region R be placed for an alternative $H_1: P_1 > P_0$, or $H_1: P_1 \neq P_0$? $P_1 \neq P_0$ means that either $P_1 > P_0$ or $P_1 < P_0$. The appropriate regions are summarized in Table 3.3.

TABLE 3.3. SUMMARY OF REGIONS OF REJECTION CORRESPONDING TO DIFFERENT ALTERNATIVES TO H_0

Alternative	Location of R	Procedure to build R
$P_1 < P_0$	Lower tail	Add $0, 1, \dots, x_r$ until α is satisfied
$P_1 > P_0$	Upper tail	Add $n, n-1, \dots, x_r$ until α is satisfied
$P_1 \neq P_0$	Both tails	Add $0, 1, \dots, x_r$ until $\alpha/2$ is satisfied; Add $n, n-1, \dots, x_r$ until $\alpha/2$ is satisfied

In the two-tailed test, α is divided between the two tails, usually with one-half in each tail. For example, if $\alpha = .05$, .025 is in the upper tail and .025 in the lower tail. Thus the region of rejection has two parts. The lower part is built up by adding $0, 1, \dots, x_r$ until the sum of the probabilities (if H_0 is true) is equal to $\alpha/2$, and the upper part by adding $n, n-1, \dots, x_r$ until the sum of the probabilities is equal to $\alpha/2$.

Remark. The one-tail tests just described satisfy the criterion: fix α and minimize β . Splitting the region of rejection in the two-tailed test, however, *does not minimize* β for testing a hypothesis against all possible alternative values of P . Nevertheless, the two-tailed test is a "best" test in another sense which will be explained later.

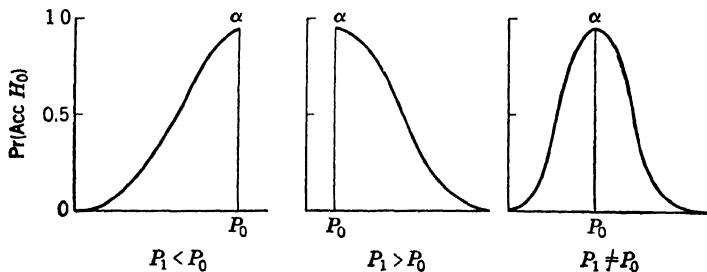


FIG. 3.2. OC curves for different alternative hypotheses.

The OC curves of statistical decision rules used for each of these alternative hypotheses have the general shapes shown in Fig. 3.2. In each, the probability of accepting H_0 becomes smaller the farther P_1 departs from P_0 .

3.7 Effect of Sample Size on Operating Characteristics

What effect will a larger sample have on the OC curve of the horse-racing test if P_0 and α remain the same? If the OC curve of Fig. 3.1 and Table 3.2 is unacceptable because it leads to the decision to invest with a larger probability than is desirable for values of P_1 like .45 and .40, how may the decision rule be changed so that the OC curve is lower, or the probability of investment is smaller, for these values of P_1 ? The horse-racing problem is solved for three different sample sizes, and the OC curve is computed for each test in Table 3.4.

The value of α is approximately .05 for each sample size. It cannot be exactly the same for each n because of the discrete nature of the

TABLE 3.4. OPERATING CHARACTERISTICS FOR THE HORSE-RACE TEST WITH THREE DIFFERENT SAMPLE SIZES

P	$H_0: P_0 = .5$		$H_1: P_1 < .5$	
	$n = 10$	$\alpha = .055$	$n = 50$	$\alpha = .059$
	$R: x \leq 2$		$R: x \leq 19$	
	Pr(Acc H_0)*		Pr(Acc H_0)†	
	Pr(Acc H_0)†		Pr(Acc H_0)†	
$P_0 = .5$	$1 - \alpha = .94$		$1 - \alpha = .94$	
$P_1 = .45$	.90		.80	
.40	.83		.55	
.35	.71		.27	
.30	.62		.08	
.25	.47		.01	
.20	.32		.001	
.15	.18		.000	
.10	.07		.000	
.05	.01		.000	
.00	.000		.000	

* From Table 3.2.

† From Harry R. Romig, "50-100 Binomial Tables," John Wiley & Sons, Inc., New York, 1953.

binomial distribution. The decision rule changes for each value of n . With $n = 10$, the rule is to accept H_1 and decide not to invest if 2 or fewer winners are observed, and to accept H_0 and invest if 3 or more winners are observed. With $n = 50$, the decision rule is to accept H_1 and not invest if 19 or fewer winners are observed, and to accept H_0 and invest otherwise. With $n = 100$, the decision rule is to accept H_1 and not invest if 41 or fewer winners are observed, and to accept H_0 and invest otherwise. The probability (α) of error I (missing an opportunity) is approximately the same for each rule. The probabilities of accepting H_0 and investing for values of $P_1 < P_0$ become smaller for the same value

of P_1 with an increase in the sample size. The OC curves are drawn in Fig. 3.3. If $P_1 = .35$, the probability of accepting H_0 and investing is .74 when $n = 10$, .27 when $n = 50$, and .09 when $n = 100$. The probabilities for other values of P may easily be read from Fig. 3.3.

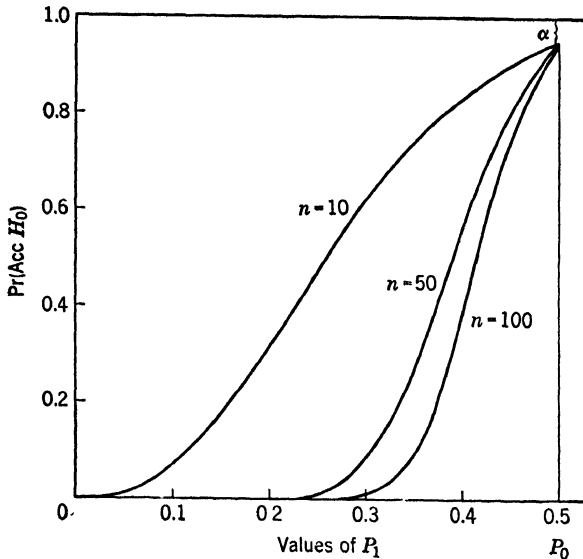


FIG. 3.3. OC curves for the horse-race test with $n = 10$, $n = 50$, and $n = 100$.

At the beginning of this course many students insisted that they could not reach a decision in the horse-racing problem with a sample of only 10 races. The author not only claimed that decisions could be made with only 10 observations, but insisted that they be made. It is not impossible to make decisions based upon small samples. It can be done, and it often must be done, although it is reasonable to hesitate before deciding with so few as 10 observations. The probability of error I (accepting H_1 when H_0 is true) is controlled and can be made very small with any sample size. But, with $n = 10$, the probability of error II (accepting H_0 when H_1 is true) is large even for values of P considerably less than P_0 . The hesitation to make decisions from a few observations is well founded, because the probability of making incorrect decisions of error II is large with small values of n . This probability of error is reduced by taking larger sample sizes, as shown by the OC curves in Fig. 3.3.

EXERCISES

Solve Exercises 3.6 to 3.10 by answering in order the following questions. Each problem involves a sample of 100 or less for which tables of binomial probabilities are available.

- a. What population is sampled?
- b. What two hypotheses are suggested?
- c. What actions are contemplated as a result of accepting each of these hypotheses?
- d. What are the errors I and II, what do they mean, and what are their costs or consequences?
- e. What is the random variable? What is the sample space? What is the location of the regions R and A ?
- f. What values of α , associated with error I and region R , are proposed?
- g. What region R satisfies the value of α selected in (f)?
- h. What is the statistical decision rule stated in terms of the actions to be taken?
- i. What are the operating characteristics of the test? For several values of P_1 under H_1 , find the value of β . Present these in a table and draw a graph of the OC curve.

j. Interpret the meaning of α and at least one value of β on the OC curve.

3.6. Solve Example 1.2.

3.7. The City Planning Commission of Teeswater, Ontario, draws a sample of 100 buildings from a map of the city showing all buildings in the community as of a given date. The sample shows that 21 per cent of these buildings are one-story houses used as residences. The property record of the city construction department shows that 35 per cent of the buildings are one-story residences. Is the sample in disagreement with the records? If so, how could you account for this difference? Of what value is a sample check of this kind?

3.8. Physicians say that one-half of the offspring of families in which one parent is diabetic will be diabetic. If, in a sample of 80 children, 35 are diabetic, do you accept the physicians' claim?

3.9. The transportation analyst of a railroad-car manufacturing company has noted a coupling condition in the design of railroad cars which, if present on more than 10 per cent of all cars in service, is sufficiently serious to warrant modification of the design. A competent inspector is located in the railroad yard of a large city to check this condition. Out of 90 cars examined, 14 had the condition present which led the coupling to break when the railroad cars were being assembled into trains. Should the car design be modified? Can you think of any more analyses that should be made before a final answer is given to the company?

3.10. The 1950 data for the whole United States show that 7040 out of 8554 wildcat oil wells drilled were dry. You are thinking of investing in a company that drills wildcat wells. In 1950, 44 out of its 70 drillings were dry. Is this experience consistent with that of the whole United States, or does it indicate that this company drills after more careful geophysical judgment by its geologists?

3.11. The X company, a leading soft-drink manufacturer, is thinking of introducing to the market a new carbonated fruit drink to supplement its already popular noncarbonated X -brand beverage. An advertising firm is asked to conduct the research necessary to decide whether commercial production of the new drink should be undertaken.

The advertising agency conducts a small study preliminary to a large-scale market survey to find out something about consumer preferences for the new

beverage. If the results of this study are encouraging, a costly market-wide analysis will follow.

The preliminary study consists of delivering six bottles of X-brand noncarbonated drink and six bottles of X-brand carbonated drink free to each of 100 homes selected randomly within a small section of the city. Each household is advised that one carton of bottles is marked with a blue band and one with a white band. The significance of the markings is not divulged (white band, noncarbonated; blue band, carbonated), but the members of the household are asked to try the drinks alternately.

After 1 week, return calls are made to each of the 100 homes originally selected, and each household is offered a free carton of drinks, either white or blue band. Call backs were made to households at which no one was home until finally a total of 80 households had been contacted. Of these, 47 preferred the white band, 22 preferred the blue band, and 11 were unable to detect a difference or had no preference.

- a. The company has decided that 30 per cent or more of the households must prefer the blue band to warrant a large market analysis. Should the analysis be undertaken?
- b. What do you think the advertising firm should advise the company to do?
- c. What would you do about the 20 households for which no return call information was available? Does this affect the randomness of the sample?
- d. Can you suggest a method of statistically handling the 20 nonresponses?

3.8 Tests of Statistical Hypotheses and Everyday Decision Problems

Have you ever made decisions using methods similar to those described in the last few pages? Many of life's decisions are based upon limited observations or, in other words, incomplete information. You do not have all the facts and it is too costly to get all of them. You must therefore rely on a sample from all the facts. The complex adjustments involved in decisions and consequent actions may be called *inductive behavior*. As an example of inductive behavior, consider the following everyday decision problem.

The weather this afternoon may be either of two possible alternatives: rainy or clear. Actually there are different degrees of rain: cloudburst, downpour, drizzle, etc. There are different degrees of clear weather: bright, hazy, partly cloudy, smoggy, etc. The problem is simplified to the dichotomy in order to emphasize the important components of the analysis without confusing the issue with details. *In the morning* you are allowed to gather information about the weather from three different sources: the sky, the newspaper, and a barometer in your living room. On the basis of these observations, you must decide whether to wear your raincoat when you leave for the office (or school). Here you have a very simple and a very common decision problem.

Assume that there are only three different observations possible from

each source of information. The observations from each source available may be:

<i>Observations</i>	<i>Sky</i>	<i>Newspaper</i>	<i>Barometer</i>
O_1	Raining	Rain	Low
O_2	Cloudy	Rain(?)	Steady
O_3	Clear	Clear	High

Let your decision be symbolized by D_i , and further let

D_r = decision that it will be raining this afternoon

D_c = decision that it will be clear this afternoon

Let your possible actions be symbolized by A_i , and further let

A_r = act of wearing your raincoat

A_c = act of leaving your raincoat at home (or not wearing it)

A rule of inductive behavior relates each A_i to a D_i , and each D_i to an O_i . That is to say, each action is consequent upon a decision, which is consequent upon an observation. For example, the action A_r (wear raincoat) follows the decision D_r (it will be raining this afternoon). Action A_c (leave raincoat at home) follows decision D_c (it will be clear this afternoon). The real problem is to relate each D_i to the observations O_i , and thus to connect the action to be taken with the observations O_i .

The rule of inductive behavior to be adopted may be analyzed in terms of probabilities, with probability defined here in the ordinary sense

Pr(O_i in the morning|state of the weather in the afternoon)

Observations which are made in the morning	State of the weather in the afternoon	
	Rain	Clear
O_1 : Sky: raining Newspaper: rain Barometer: low		
O_2 : Sky: cloudy Newspaper: rain(?) Barometer: steady		
O_3 : Sky: clear Newspaper: clear Barometer: high		

as a relative frequency of occurrence. *Each probability is the probability that the appropriate observation will occur in the morning if the weather is truly in a certain state in the afternoon.* As an example, the probability of O_1 in the morning if the weather is rainy in the afternoon means this: of all the days on which there actually was rain in the afternoon, what proportion had rain in the morning, a newspaper prediction of rain, and a low barometer?

These probabilities may be entered in a table, as shown on page 68. Although you can hardly fill in these probabilities in objective mathematical terms, you do make decisions as if these cells were filled in with intuitive or subjective or personal probabilities. You will derive value from the experience of assigning numbers, however subjective they may be, to these cells. Precisely because this is so largely an intuitive process, you may find yourself in disagreement with the author's choice of probabilities in the six cells, but it will serve as a basis for discussion.

PERSONAL PROBABILITIES OF R. C. SPROWLS
(O_i in the morning|state of the weather in the afternoon)

Observations which are made in the morning	State of the weather in the afternoon	
	Rain	Clear
O_1 : Sky: raining Newspaper: rain Barometer: low	.8	.1
O_2 : Sky: cloudy Newspaper: rain(?) Barometer: steady	.1	.1
O_3 : Sky: clear Newspaper: clear Barometer: high	.1	.8

Remark. The meaning of these probabilities, often confusing to students, is interpreted as follows: $\Pr(O_1|\text{rain}) = .8$ means that on 80 per cent of all the days on which it rained in the afternoon, it was raining in the morning, the newspaper predicted rain, and the barometer was low. $\Pr(O_1|\text{clear}) = .1$ means that on 10 per cent of all the days on which it was clear in the afternoon, it was raining in the morning, the newspaper predicted rain, and the barometer was low.

The hypotheses being tested are:

- H_0 : rain
- H_1 : clear

The sample space consisting of the three points O_1 , O_2 , and O_3 must be divided into a region R which will lead to the acceptance of H_1 and a region A which will lead to the acceptance of H_0 . Common sense would lead the author to put O_3 into R because it is more probable under H_1 : clear, than under H_0 : rain. Therefore, if O_3 is observed, the decision is to accept H_1 , reach decision D_c : clear, and take action A_c : leave raincoat at home. The point O_1 will go into A because the probability that O_1 will occur is greater under H_0 than under H_1 . If O_1 is observed, the decision is to accept H_0 , reach decision D_r : rain, and take action A_c : wear raincoat. If O_2 is observed, what is the decision to be reached and the action to be taken? Probabilities are not enough to determine whether O_2 should be in the region of rejection or acceptance because the probabilities are equal under either hypothesis.

The decision problem thus far may be formulated in the following schematic diagram:

Decision	True state of the weather	
	Rainy	Clear
D_r : rain	Acceptable	Error
D_c : clear	Error	Acceptable

D_r is obviously the correct decision on a rainy day and D_c the correct decision on a clear day. D_r is an incorrect decision on a clear day; D_c is an incorrect decision on a rainy day. Following the convention of statistical tests of hypotheses, these errors may be labeled:

I: Acc clear|rain

II: Acc rain|clear

Errors in decisions involve costs. What are the costs of making these two errors?

I: Reaching decision D_c on a rainy day means that you will not be wearing your raincoat on a rainy day. You will get wet; you may catch cold or get pneumonia; death may follow, etc.

II: Reaching decision D_r if it is clear means that you will be wearing your raincoat on a clear day. You will look silly; the neighbors will shake their heads; you will get hot and be uncomfortable, etc.

The rule of inductive behavior depends not only upon the probabilities involved but also upon these costs. If you do not mind getting wet but abhor the thought of being seen wearing a raincoat on a clear day, O_2 should be placed in the region R . If you dislike chancing pneumonia, O_2 should be placed in the region A . Suppose the region R in this problem were $R: O_3$. In other words, you will accept clear and not wear your

raincoat only when you observe the evidence of O_3 . What are the magnitudes of α and β ?

$$\alpha = \Pr(O_3|\text{rain}) = .1 \text{ (from the table)}$$

$$\beta = \Pr(O_2 \text{ or } O_1|\text{clear}) = .1 + .1 = .2$$

These probabilities have the following meaning in this problem:

$\alpha = .1$: On one-tenth of all the rainy days, you will not be wearing your raincoat.

$\beta = .2$: On two-tenths of all the clear days, you will be wearing your raincoat.

Now suppose that you do not mind getting wet but hate to be out in the sunshine wearing a raincoat. You decide to put O_2 in the region R along with O_3 . Now what are α and β ?

$$\alpha = \Pr(O_2 \text{ or } O_3|\text{rain}) = .2$$

$$\beta = \Pr(O_1|\text{clear}) = .1$$

Note what has happened. There is now only one chance in ten that you will be wearing your raincoat on a day with a clear afternoon (β). The probability of making this error has been reduced, a desirable solution for you because the error involved the greatest cost for you. There are now two chances in ten that you will not be wearing your raincoat on a day with a rainy afternoon (α). You must pay for a lower β with a larger α . The costs of making errors affect your test by influencing your choice of these probabilities, especially α . You are free to choose any α that you want. Its choice is not, strictly speaking, a statistical problem. Its size depends upon the seriousness of the errors of an incorrect decision. It is most important for you to recognize that statistical tests are methods of making decisions which have certain probabilities of being correct or incorrect, or methods which give certain degrees of assurance of reaching correct or incorrect decisions.

Many real-life decisions can be simplified to a choice between two alternatives. Although in most instances objective probability calculations cannot be made, the basic principles of statistics are useful in analyzing these everyday decision problems because the statistical approach brings out into the open the use of subjective probabilities. Statistics is applied in those fields where one can measure or one thinks one can measure objective probabilities. It is not employed in fields where one cannot measure, but instead relies upon intuitive feeling for these probabilities. It is a consequence that among men of good judgment, differences usually come down to different estimates of the probabilities of the decision being right or wrong, or different evaluations of the associated costs of being right or wrong. This course is intended to give you an

insight into the decision-making process which you can use all the rest of your life whether or not you can compute the probabilities.

EXERCISES

3.12. A certain couple lived in a furnished apartment on the third floor of an apartment house on the South Side of Chicago. Both the man and his wife were working at the time, and every day they closed all the windows when they left for work.

a. Analyze their behavior in terms of the weather (rainy or clear), stating exactly what are the errors I and II.

b. How large are α and β ?

c. From your knowledge of α and β , can you say anything at all about their personal probabilities of rainy and clear days, given information about the weather?

d. Do you think that they believe these probabilities are 0 or 1?

e. If your answer to (d) is no, what must they believe about the costs of making these errors?

f. What do you think of the procedure of neglecting any information you may have in making decisions?

3.13. Suppose you are a student taking a course where the instructor gives short unannounced examinations. There may be times when the pressure of other work causes you to miss a class. Set up the decision problem in terms of the fourfold classification, stating the decisions, actions, errors, and consequences of the errors. Do probabilities enter into your decision rule? Can you get any objective measure of these probabilities?

3.14. Review the decisions you have made in the past few weeks or expect to make in the next few weeks. Take any one of them and set it up in the fourfold classification. State specifically the actions you contemplate, the errors you can make, the costs and consequences of those errors, and your method of arriving at your decision. Does probability play any part in your decision? Suggested decisions which you may face are dating a girl, coming to statistics class, cheating on an examination, etc.

SUPPLEMENTARY READING

Coates, R. M.: The Law, *The New Yorker*, vol. 23, no. 41, pp. 41-43, Nov. 29, 1947.

This humorous story has implications for statistics.

Fisher, R. A.: Biometry, *Biometrics*, vol. 4, no. 3, pp. 217-219, September, 1948.

This is a short and highly philosophical discussion of the mode of logic known as induction written by one of the leaders in the modern development of statistics.

Kendall, M. G.: The Statistical Approach, *Economica*, vol. 27, no. 66, pp. 127-145, May, 1950. Kendall surveys many ideas, applications, and problems in the field of statistics.

Snedecor, George W.: On a Unique Feature of Statistics, *Journal of the American Statistical Association*, vol. 44, no. 245, pp. 1-8, March, 1949. A prominent American statistician discusses what he believes to be the essential nature of statistics.

CHAPTER 4

STATISTICAL INFERENCE IN THE DISCRETE CASE: BINOMIAL DISTRIBUTION (II)

For sample sizes of 100 or fewer observations, the statistical decision rules and the operating characteristics of these rules may be computed by using tables of the binomial probability distribution. For n greater than 100, other methods of computation must be used. Of the several methods available, one will be discussed in detail in this chapter. The basic ideas of making decision rules and evaluating them in terms of the OC curve will not change in the problems discussed in this chapter. *Only the method of computing the probabilities α and β will change.*

4.1 An Approximation to the Binomial Distribution: the Normal Distribution

The most important probability distribution in statistics is the *normal distribution*. The term "normal" as used here does not mean "average" or "not deviating from an established standard" or "regular." The normal distribution is a mathematical probability distribution of a continuous variable which theoretically may take on any value from minus infinity to plus infinity. A graph of the normal distribution is a hump-shaped or bell-shaped distribution which is symmetrical about its center at the mean μ . The constant which characterizes the population for which the binomial distribution is the appropriate probability distribution is P , the proportion of elementary units in a particular category in the population. The normal distribution¹ is characterized by two con-

¹ The equation of the normal probability distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where x = a continuous variable which may take on values from $-\infty$ to $+\infty$

π = a constant = 3.14159 . . .

e = 2.71828 . . .

μ = arithmetic mean

σ^2 = variance

σ = standard deviation

stants: the mean μ and the standard deviation σ . μ is the value at the center of the distribution; σ is a measure of the variation or the dispersion of the observations about the center of μ .

Figure 4.1 shows several normal distributions with different combinations of μ and σ . Larger values of σ mean a wider spread or dispersion

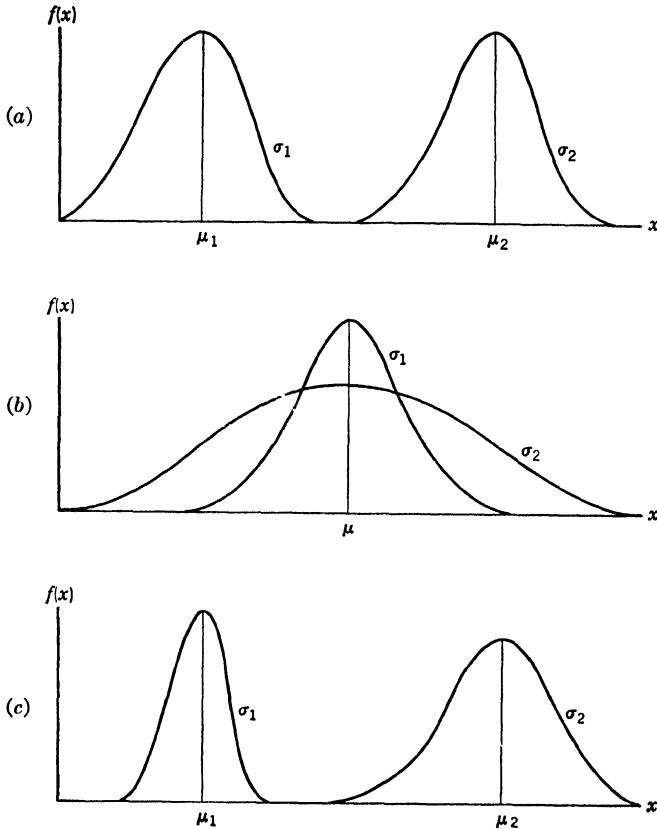


FIG. 4.1. Graphs of normal distributions with different means and standard deviations: (a) Unequal means, $\mu_2 > \mu_1$, equal standard deviations, $\sigma_1 = \sigma_2$. (b) Equal means; unequal standard deviations, $\sigma_2 > \sigma_1$. (c) Unequal means, $\mu_2 > \mu_1$; unequal standard deviations, $\sigma_2 > \sigma_1$.

about μ , and these are clearly indicated by the individual graphs. The shape of the normal distribution depends upon both μ and σ and is not always the same.

The normal distribution is a probability distribution, and therefore the total area under its graph is equal to 1. Probabilities associated with various values of x may be represented as areas under the normal curve corresponding with these values. For example, let x take on two values,

x_1 and x_2 , as shown in Fig. 4.2*b*. The probability that x will lie between x_1 and x_2 is the area between ordinates erected at these two values. Expressed in symbols, the statement is $\Pr(x_1 \leq x \leq x_2) = \text{area between } x_1 \text{ and } x_2$. The areas corresponding to the three probability statements

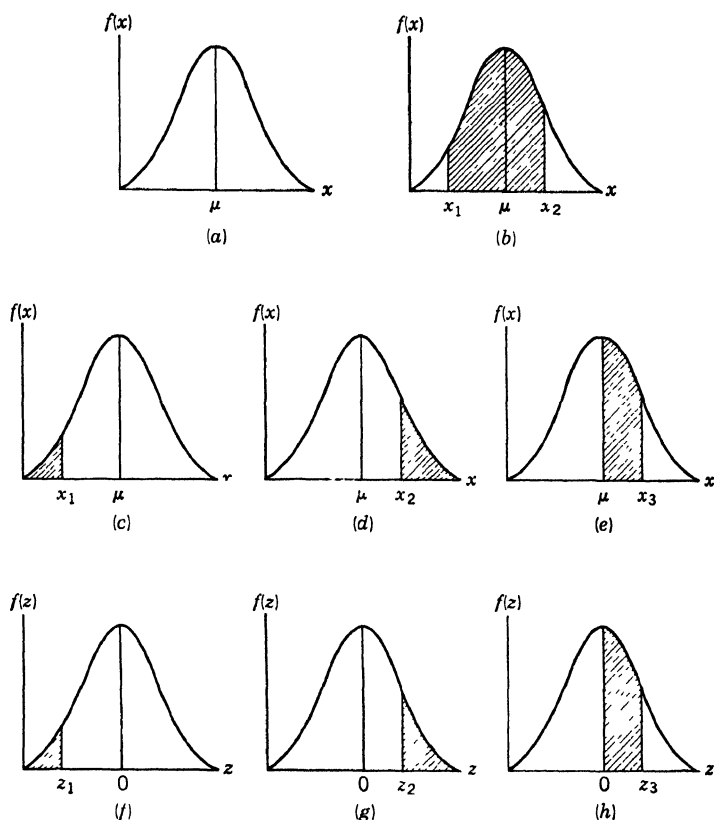


FIG. 4.2 Areas and probabilities of the normal distribution: (a) Normal distribution. (b) $\Pr(x_1 \leq x \leq x_2)$. (c) $\Pr(x \leq x_1)$. (d) $\Pr(x \geq x_2)$. (e) $\Pr(\mu \leq x \leq x_3)$. (f) $\Pr(z \leq z_1)$. (g) $\Pr(z \geq z_2)$. (h) $\Pr(0 \leq z \leq z_3)$.

that follow are shown in Fig. 4.2*c* to *e*: (c) $\Pr(x \leq x_1)$; (d) $\Pr(x \geq x_2)$; (e) $\Pr(\mu \leq x \leq x_3)$.

Fortunately, to arrive at probabilities of normally distributed variables, these areas have been tabled for the *standard normal distribution*. Any normal distribution function $f(x)$ can be transformed into the standardized form, written $f(Z)$, using the relation

$$Z = \frac{x - \mu}{\sigma} \quad (4.1)$$

In other words, Z is the difference between the variable x and its mean μ ,

divided by the standard deviation σ . The standard normal distribution $f(Z)$ has a mean equal to 0 and a standard deviation equal to 1. For different values of Z , probabilities may be represented by areas, just as with different values of x , as shown in Fig. 4.2*f* to *h*.

Table A.3 of Appendix A is a table of areas of the standard normal distribution. The area tabled is that area shown in Fig. 4.2*h*, i.e., the area between $Z = 0$ and a positive value of Z . This same diagram is drawn at the top of Table A.3 to indicate what area is tabled. Z is a pure number (having no units of measure) which may range from $-\infty$ to $+\infty$.

4.2 Table of Areas of the Normal Distribution

In order to demonstrate how Table A.3 is read, a part of it has been reproduced here. Z may be read to two decimal places, the first decimal place in the row stub and the second decimal place in the column caption.

Example 4.1. What is the probability that Z will be between 0 and 1.00? $\Pr(0 \leq Z \leq 1.00)$. Read down the first column labeled Z to 1.0. Follow this row across to the column labeled .00. The entry in this cell is .3413. $\Pr(0 \leq Z \leq 1.00) = .3413$.

SELECTED VALUES FROM TABLE A 3

Z	.00	.04	.05	.06	.08
0 20	.0793				
1 0	.3413				
1 2					.3997
1 6	.4452	.4495	.4505		
1 9	.4713			.4750	

Example 4.2. Find $\Pr(0 \leq Z \leq +1.28)$. Read down to $Z = 1.2$ and then across to .08 in this row. The entry is .3997. $\Pr(0 \leq Z \leq 1.28) = .3997$.

Not all problems are so simple as these. Several more difficult ones will be explained in the text and followed by homework exercises. A graph of the exact area to be found is most helpful in solving these problems.

Example 4.3. Find $\Pr(.20 \leq Z \leq 1.90)$.

- The area between $Z = 0$ and $Z = +.20$ is found in the table to be .0793.
- The area between $Z = 0$ and $Z = +1.90$ is found in the table to be .4713.
- Therefore, the area between $Z = .20$ and $Z = +1.90$ is the difference $.4713 - .0793 = .3920$. $\Pr(.20 \leq Z \leq 1.90) = .3920$.

Example 4.4. Find $\Pr(-1.60 \leq Z \leq 0)$. The area between 0 and -1.60 is not in the table, but because the distribution is symmetrical about $Z = 0$, the

area between $Z = -1.60$ and $Z = 0$ is the same as the area between $Z = 0$ and $Z = +1.60$, which is .4452. $\Pr(-1.60 \leq Z \leq 0) = .4452$.

Example 4.5. Find $\Pr(-1.60 \leq Z \leq -1.28)$. The area between 0 and -1.60 is .4452. The area between 0 and -1.28 is .3997. Therefore, the area between -1.60 and -1.28 is $.4452 - .3997 = .0455$.

Example 4.6. Find $\Pr(Z > 1.65)$. The area between 0 and 1.65 is .4505. The area from 0 to $+\infty$ is .5000 because the distribution is symmetrical and half of its area is above 0. Therefore, the area above 1.65 is the area in the upper tail of the distribution; it is $.5000 - .4505 = .0495$.

EXERCISES

4.1. Find each of the following probabilities from areas of the standard normal distribution, $f(Z)$.

(a) $Z \geq +1.96$

(c) $Z \geq +1.645$ (use linear interpolation)

(b) $Z \leq -1.96$

(d) $Z \leq -1.645$ (use linear interpolation)

4.2. Find the value of Z in the upper tail of the distribution which is exceeded by an area of .05; by an area of .025. Find the value of Z in the lower tail which cuts off an area of .05; of .025.

4.3. What values of Z cut off .05 of the total area of the curve with .025 in each tail?

4.4. What values of Z cut off .10 of the total area of the curve with .05 in each tail?

4.3 The Normal Distribution as an Approximation to the Binomial Distribution

Figure 4.3a is a histogram of the binomial distribution for $n = 10$ and $P = .5$. The total area under the curve is 1.00, and the area of each bar of the histogram, representing the probability of the corresponding value of x , is shown. These values are taken from Table A.2. A smooth curve is drawn to fit the general shape of the histogram. The areas of the bars may now be approximated by areas under this smooth curve; the latter may be computed if the mathematical equation of the curve is known. The standard normal distribution is the equation of the curve which, under some conditions, gives a good approximation to the binomial histogram. A review of the notation is:

n = sample size

x = random variable; number of occurrences in the sample

P = population proportion

$Q = 1 - P$

$\mu = nP$ = mean number of occurrences

$\sigma = \sqrt{nPQ}$ = standard deviation of the number of occurrences

The standard normal variable Z is the deviation of a normally distributed variable x from its mean value, divided by the standard deviation: $Z = (x - \mu)/\sigma$. Since the mean and standard deviation of the binomial distribution are known,

$$Z = \frac{x - nP}{\sqrt{nPQ}} \quad (4.2)$$

The normal distribution is *only an approximation* to the binomial distribution because the number of occurrences is a discrete variable which

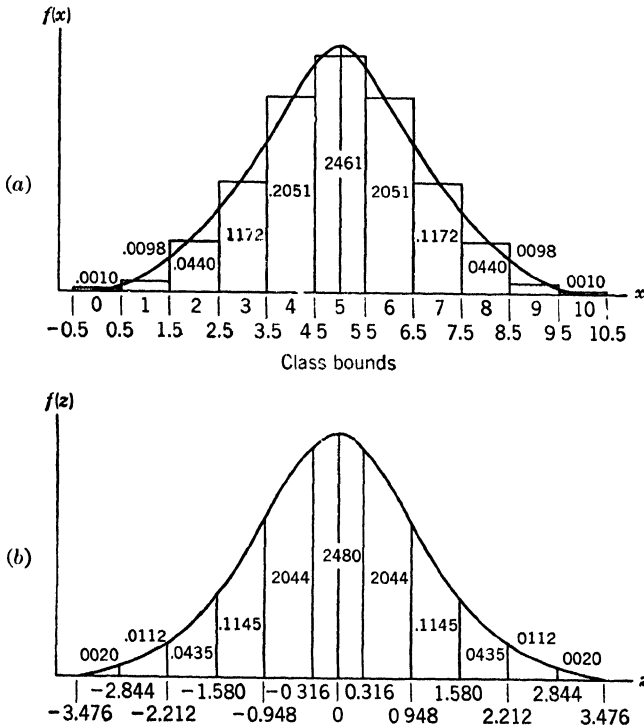


FIG. 4.3. Normal approximation to the binomial distribution with $n = 10$ and $P = 5$. (a) Histogram of binomial probabilities with smooth curve approximation. (b) Normal distribution with $\mu = 5$ and $\sigma = 1.58$ as an approximation to binomial probabilities; $z = (x - nP)/\sqrt{nPQ}$.

can assume only the values of $0, 1, 2, \dots, n$, while Z is a continuous variable which may assume any value from $-\infty$ to $+\infty$. The probability approximation will be better if a correction factor of one-half is either added or subtracted, depending upon the required probability, thus:

$$Z = \frac{x \pm \frac{1}{2} - nP}{\sqrt{nPQ}} \quad (4.3)$$

If a small diagram is drawn for each problem, the appropriate correction will be obvious.

The normal distribution which approximates the binomial distribution for $n = 10$ and $P = .5$ is drawn in Fig. 4.3b; the probabilities are labeled in the areas under the curve. The method of computing these probabilities is shown in the following examples.

Example 4.7. Find $\Pr(x = 5)$. The bar of the histogram which represents $x = 5$ has bounds of 4.5 and 5.5. The approximate probability is the area under the curve from $x = 4.5$ to $x = 5.5$.

If $x = 4.5$,

$$Z = \frac{4.5 - 5}{1.58} = -.316 \quad [\text{from Eq. (4.3)}]$$

If $x = 5.5$,

$$Z = \frac{5.5 - 5}{1.58} = .316 \quad [\text{from Eq. (4.3)}]$$

From Table A.3, the area between $Z = -.316$ and $Z = .316$ is .2480.

Example 4.8. Find $\Pr(x = 2)$. The bounds of the bar of the histogram are 1.5 and 2.5. The corresponding values of Z are -2.212 and -1.580 . From Table A.3, the area between these two values of Z is .0435.

The values of Z corresponding to the bounds of the histogram are labeled on Fig. 4.3b. The areas between different values of Z may be checked in Table A.3.

Example 4.9. Find $\Pr(x \geq 6)$. This probability is the sum of the probabilities for $x = 6, 7, 8, 9$, and 10 , which is represented by the area of these bars of the histogram. Note carefully that the bar for $x = 6$ starts at a lower bound of 5.5. Thus, in finding the value of Z , the appropriate value of x is $6 - .5$, or 5.5. $Z = .316$. The *continuity correction* in this instance is of the form $x - 1/2$. $\Pr(Z \geq .316) = .376$.

Example 4.10. Find $\Pr(x \leq 3)$. This probability is the sum of the probabilities for $x = 0, 1, 2$, and 3 , which is represented by the area of these bars of the histogram. Note carefully that the bar for $x = 3$ ends at an upper limit of 3.5. Thus, in finding the value of Z , the appropriate value of x is $3 + .5$, or 3.5. $Z = -.948$. The *continuity correction* in this instance is of the form $x + 1/2$. $\Pr(Z \leq -.948) = .172$.

In general, when the probability is of the form equal to or greater than a value of x ($\geq x$), the correction is to subtract $1/2$; when the probability is of the form equal to or less than x ($\leq x$), the correction is to add $1/2$. A rough diagram of the histogram and the normal curve approximation will serve as a reminder of the continuity correction in any specific problem.

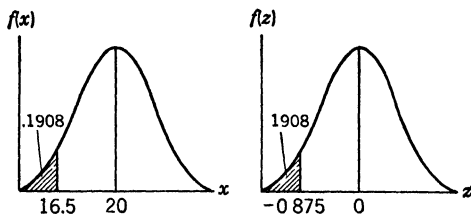
Even for so small a value as $n = 10$, the normal approximation is good if $P = .5$. If $P \neq .5$, the approximation becomes better as n becomes larger. Since tables are available for values of n up to 100 (see Sec. 2.4 for the references), the normal approximation will be used only for samples larger than 100. For very small or very large values of P , the approximation may not be good even for larger values of n . A general rule of thumb is that the approximation is fair if $nPQ \geq 4$.

Example 4.11. Find $\Pr(x \leq 16 | n = 100, P = .2)$.

$\mu = nP = 20$. $\sigma = \sqrt{nPQ} = 4$. $Z = (16 + .5 - 20)/4 = -.875$.

$$\Pr(Z \leq -.875) = .1908.$$

Therefore, $\Pr(x \leq 16 | n = 100, P = .2) \doteq .2$. Diagrams are useful aids in learning to use the normal distribution.

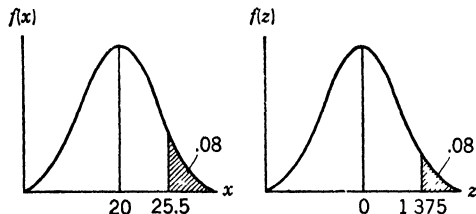


Example 4.12. Find $\Pr(x \geq 26 | n = 100, P = .2)$.

$\mu = 20$ and $\sigma = 4$. $Z = (26 - .5 - 20)/4 = 1.375$.

$$\Pr(Z \geq 1.375) = .0845.$$

Therefore, $\Pr(x \geq 26 | n = 100, P = .2) \doteq .08$.



Remark. The exact tabled probabilities are .192 (versus .1908) and .087 (versus .0845).

4.4 More on the Meaning of the Standard Deviation

Let x be a normally distributed variable with mean μ and standard deviation σ . Let x take on two values: $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$. What is the probability that x will lie within these two limits, that is, $\Pr[(\mu - 3\sigma) \leq x \leq (\mu + 3\sigma)]$?

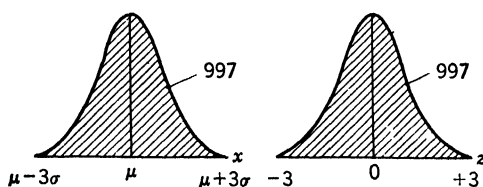
If $x = \mu + 3\sigma$,

$$Z = \frac{x - \mu}{\sigma} = \frac{\mu + 3\sigma - \mu}{\sigma} = +3$$

If $x = \mu - 3\sigma$,

$$Z = \frac{x - \mu}{\sigma} = \frac{\mu - 3\sigma - \mu}{\sigma} = -3$$

$\Pr(-3 \leq Z \leq +3) = .997$. The graphs of x and Z are shown below:



In other words, for all practical purposes, the range of Z from -3 to $+3$ includes the whole distribution of values of Z (.997) and the range $\mu \pm 3\sigma$ includes the whole range of values of x (.997).

Example: $n = 400$; $P = .5$; $\mu = nP = 200$; $\sigma = \sqrt{nPQ} = 10$.
 $\mu \pm 3\sigma = 200 \pm 30 = 170$ to 230

In samples of 400 from a population with $P = .5$, the values of x may range from 0 to 400 (the sample space), but the probability is .997 that x will lie between 170 and 230.

The standard deviation measures the dispersion or variability of the values of a variable about the mean value. If the variable is normally or approximately normally distributed, the standard deviation may be used to find the probability of values occurring inside or outside certain deviations from the mean and, thus, be given real meaning in a problem.

EXERCISES

Find each of the probabilities indicated. Draw graphs for the distribution of x and Z , and darken the required area under each curve. Use the proper continuity correction, which is indicated for some problems.

- 4.5. a. $\Pr(x \geq 56 | n = 100, P = .5)$ $(-1/2)$
 b. $\Pr(45 \leq x \leq 55 | n = 100, P = .5)$ $(-1/2; +1/2)$
 c. $\Pr(15 \leq x \leq 25 | n = 100, P = .2)$ $(-1/2; +1/2)$
 d. Compare each of these with tabled probabilities.
- 4.6. a. $\Pr(x \leq 117 | n = 256, P = .5)$
 b. $\Pr(x \leq 180 | n = 400, P = .5)$

Find the value of x_r such that the probability relationship indicated is true. [*Hint.* Find the value of Z for each probability from Table A.3, and then solve for x_r in Eq. (4.3).]

- 4.7. a. $\Pr(x \leq x_r | n = 100, P = .5) = .05$ $(+1/2)$
 b. $\Pr(x \leq x_r | n = 100, P = .2) = .025$ $(+1/2)$
 c. $\Pr(x \geq x_r | n = 100, P = .5) = .05$ $(-1/2)$
 d. $\Pr(x \geq x_r | n = 100, P = .2) = .025$ $(-1/2)$
- 4.8. a. $\Pr(x \leq x_r | n = 256, P = .5) = .025$
 b. $\Pr(x \geq x_r | n = 256, P = .5) = .025$
 c. $\Pr(x \leq x_r | n = 400, P = .5) = .025$

4.9. Suppose 100 "fair" coins ($P = .5$) are tossed randomly. Let x = number of heads. What values of x may occur? Within what limits will values of x occur with probability .997? With probability .95? With probability .50?

4.5 Use of the Normal Approximation in Testing Hypotheses

Example 1.3 may be solved by use of the normal approximation to the binomial. The statement of the problem is repeated here. Suppose that you, as the manager of the china department of a large department store, are responsible for the ordering of inventory stock. Two different patterns of china, pattern A and pattern B , are ordered from the same manufacturer. The two patterns are ordered in equal quantities on the assumption that they sell equally well. From time to time a random sample of 256 sales of the two patterns is examined, usually just before more sets are ordered for stock. What number of pattern A sales will lead you to accept the hypothesis that sales of the two patterns are equal, and what number to accept the hypothesis that the sales of the two patterns are not equal?

The statistical test is set out in the following form.

$H_0: P_0 = .5$ (50 per cent of the sales are pattern A)

$H_1: P_1 \neq .5$ (if not 50 per cent sales of A , the percentage can be either more or less than 50)

$\alpha = .05$ (set arbitrarily; the probability of accepting that other than 50 per cent of the sales are A when, in fact, this is true)

$n = 256$ (sample size)

$R: x \leq x_r \text{ and } x \geq x_r$ (a two-tailed test because of the alternative hypothesis $P_1 \neq P_0$)

The region R which leads to the acceptance of H_1 is divided between the two tails of the distribution because of H_1 . The probability α is divided equally, with .025 in the upper tail and .025 in the lower tail. The normal approximation to the binomial is used because n is 256.

The values of the standard normal variable Z which cut off .025 in each tail of the normal distribution are -1.96 and $+1.96$. The region R in the Z distribution is graphed in Fig. 4.4a. To find the region R in

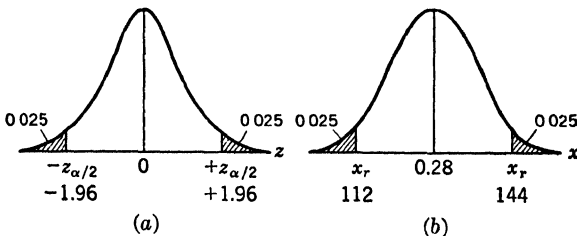


FIG. 4.4. Graphic representation of the china sales problem. (a) Distribution of z . (b) Distribution of x .

the x distribution, the number of pattern A sales, it is necessary to solve the equation which relates Z to x , using both values of Z and the proper continuity correction; in general

$$Z = \frac{x \pm \frac{1}{2} - nP}{\sqrt{nPQ}} \quad (4.3)$$

Lower tail:

$$-1.96 = \frac{x_r + \frac{1}{2} - 256(.5)}{\sqrt{256(.5)(.5)}} \quad x_r = 128 - .5 - 8(1.96) = 112$$

Upper tail:

$$+1.96 = \frac{x_r - \frac{1}{2} - 256(.5)}{\sqrt{256(.5)(.5)}} \quad x_r = 128 + .5 + 8(1.96) = 144$$

The region R is diagramed in Fig. 4.4b. It consists of two parts: a lower part containing values of x from 0 to 112 and an upper part containing values of x from 144 to 256. This is written $R: x \leq 112$ and $x \geq 144$.

Statistical-decision Rule. Accept H_1 , that the sales of patterns A and B differ, if the observed number of sales of pattern A is 112 or less, or 144 or more; otherwise accept H_0 .

The statistical test of the china sales is not complete. The probability of accepting H_1 when, in fact, H_0 is correct is controlled at $\alpha = .05$. What is the probability of accepting H_0 when, in fact, H_1 is correct? The operating characteristic curve answers this question and is shown in Table 4.1 and Fig. 4.5. Values of β are the probabilities of accepting

TABLE 4.1. OPERATING-CHARACTERISTIC CURVE OF THE CHINA SALES TEST

$P_1 =$	$Pr(Acc H_0)$
.30	.00
.40	.11
.42	.28
.45	.65
.47	.85
$.50 = P_0$	$.95 = 1 - \alpha$
.53	.85
.55	.65
.58	.28
.60	.11
.70	.00

that 50 per cent of the sales are pattern A associated with various alternative values of the percentage of pattern A sales. If only 40 per cent of the sales are truly of pattern A , the probability of accepting the hypothesis that 50 per cent are sales of A is .11; if 45 per cent of the sales are truly of A , the probability is .65. The OC curve describes the sampling plan in detail. Without this curve the probabilities of accept-

ing H_0 when, in fact, H_1 is correct are unknown. In repeated samples the hypothesis that 50 per cent of the sales are pattern A may be accepted and equal quantities of the two patterns ordered only because there is a very high probability of accepting this hypothesis even when it is false.

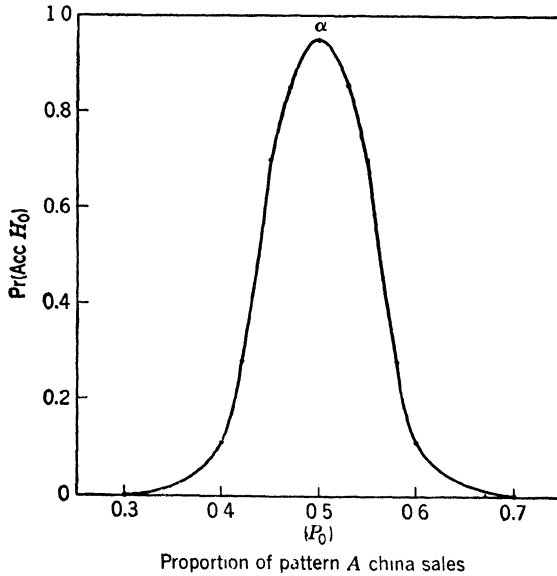


FIG. 4.5. OC curve of the china sales test.

The OC curve is low at both ends and high in the middle. The alternative hypothesis is that $P_1 \neq P_0$, an alternative which calls for a two-tailed test to guard against both high and low values of P_1 . The OC curve shows that the probability of accepting H_0 becomes less with both smaller and larger values of P_1 , that is, the test affords protection against undesirable values of P_1 alternative to P_0 .

Remark. The values of β on the OC curve in a two-sided test are not minimum values against all alternative values of P_1 . The two-tailed test is used because the probability of accepting H_0 is larger for $P = P_0$ than for any other value of P . In this sense, the two-tailed test is the *best unbiased test*.

Another example of the use of the normal approximation in testing hypotheses about percentages is a straw vote in a campus election. The vote indicates that 190 out of 400 students polled will vote for candidate A. Test the hypothesis that candidate A will get 50 per cent or more of the votes in the election.

$H_0: P_0 = .5$ (50 per cent of the votes)

$H_1: P < .5$ (candidate A will get less than half the votes)

$\alpha = .025$ (arbitrarily selected probability of making error I)

$n = 400$

$R: x \leq x_r$ (a one-tailed test with region R in lower tail)

The value of the standardized normal variable Z which cuts off .025 in the lower tail of a normal distribution is -1.96 . The region of rejection may be found by solving this equation for x_r :

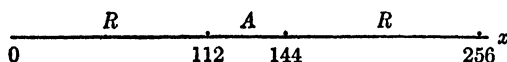
$$-1.96 = \frac{x_r + \frac{1}{2} - 400(.5)}{\sqrt{400(.5)(.5)}} \quad x_r = 179.9 \doteq 180$$

Statistical-decision Rule. Accept H_1 (that candidate A will get less than half of the votes) if 180 or fewer students in the straw vote indicate that they favor candidate A ; otherwise, accept H_0 .

Decision. Accept H_0 because 190 students favor candidate A . The operating characteristics of this test are not given here.

4.5a Computation of the Operating-characteristic Curve of the China Sales Test

The OC curve may be readily computed using the normal approximation to the binomial distribution. The sample space is divided into the regions R and A in the china sales test as follows:



The probability of accepting H_0 is the probability that x will be between 112 and 144. $\beta = \Pr(\text{Acc } H_0 | H_1 \text{ true})$. Therefore,

$$\beta = \Pr(112 < x < 144 | n = 256, P_1)$$

For different values of P_1 , the values of $x = 112$ and 144 can be converted into Z values and the area between these Z values found in the table of areas of the normal distribution, Table A.3.

Example 4.13. Let $P_1 = .47$. The value of β will be the same for

$$P_1 = 1 - .47 = .53$$

because of the complementary relationship of P and Q . Two points on the OC curve of a two-sided test are obtained with only one computation.

If $P_1 = .47$,

$$\begin{aligned} \mu &= nP_1 = 256(.47) = 120.32 \\ \sigma &= \sqrt{nPQ} = \sqrt{256(.47)(.53)} \end{aligned}$$

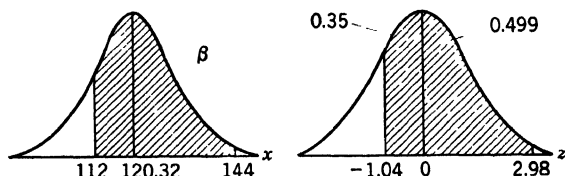
With $x_r = 112$,

$$\begin{aligned} Z &= \frac{112 - 120.32}{\sqrt{256(.47)(.53)}} \\ &= -1.04 \end{aligned}$$

With $x_r = 144$,

$$Z = \frac{144 - 120.32}{\sqrt{256(.47)(.53)}} \\ = 2.98$$

The probabilities may be shown as areas under the curves for x and Z and computed from the table of normal areas (Table A.3).



$$\beta = .35 + .499 \doteq .85$$

If $P_1 = .53$,

$$nP_1 = 256(.53) = 135.68$$

With $x_r = 112$,

$$Z = \frac{112 - 135.68}{\sqrt{256(.53)(.47)}} \\ = -2.98$$

With $x_r = 144$,

$$Z = \frac{144 - 135.68}{\sqrt{256(.53)(.47)}} \\ = 1.04$$

Only the signs of the two values of Z are changed. Since the normal distribution is symmetrical, the probabilities and, therefore, β remain the same. If

$$P_1 = .53, \beta = .85.$$

Example 4.14. Let $P_1 = .45$ or $.55$. β will be the same for both values of P_1 . $P_1 = .45$, $nP_1 = 115.2$.

With $x_r = 112$,

$$Z = \frac{112 - 115.2}{\sqrt{256(.45)(.55)}} \\ = -.4$$

$$\Pr(-.4 < Z < 0) \doteq .15$$

With $x_r = 144$,

$$Z = \frac{144 - 115.2}{\sqrt{256(.45)(.55)}} \\ = 3.62$$

$$\Pr(0 < Z < 3.62) \doteq .4998$$

$$\beta = .15 + .4998 \doteq .65$$

Example 4.15. Let $P_1 = .42$ or $.58$.

If $P_1 = .42$,

$$nP_1 = 256(.42) = 107.52$$

With $x_r = 112$,

$$Z = \frac{112 - 107.52}{\sqrt{256(.42)(.58)}} \\ = .57$$

$$\Pr(0 < Z < .57) \doteq .2157$$

With $x_r = 144$,

$$Z = \frac{144 - 107.52}{\sqrt{256(.42)(.58)}} \\ = 4.62$$

$$\Pr(0 < Z < 4.62) \doteq .5000$$

$$\beta = .5000 - .2157 \doteq .28$$

EXERCISES

4.10. *a.* In the china sales problem, let $P_1 = .4$ or $.6$. If $P_1 = .4$,

$$nP_1 = 256(.4) = 102.4.$$

With $x_r = 112$,

$$Z = \frac{112 - 102.4}{\sqrt{256(.4)(.6)}} \\ = 1.32$$

With $x_r = 144$,

$$Z = \frac{144 - 102.4}{\sqrt{256(.4)(.6)}} \\ = 5.3$$

Check the statement that $\beta \doteq .11$.

b. Let $P_1 = .3$ or $.7$. Check these values: With $x_r = 112$, $Z = 4.8$; with $x_r = 144$, $Z = 9.2$. β is approximately equal to 0.

c. Interpret the values computed in (*a*) and (*b*).

4.11. *a.* The solution of a cryptogrammic code depends upon knowledge of relationships which exist in a language. In the English language about half of the words contain the letter *e*. Set up a statistical test of this hypothesis, using a sample size of 400 words. State both hypotheses, and find the statistical decision rule, using your own value of α .

b. Take a random sample of 400 words from some source, and apply the decision rule according to the plan outlined in (*a*). What decision do you reach? What is your chance of accepting H_1 if 50 per cent of the English words do contain the letter *e*?

c. Select various values of P_1 under the alternative hypothesis, and find values of β . Collect these values in a table, and plot as the operating-characteristic curve. Write out the meaning of the value on the curve for $P_1 = .4$ and $P_1 = .55$.

4.12. The bakeshop superintendent of a baking company is considering the purchase of a new machine which will ice cakes faster with no increase in the current 5 per cent which are improperly iced. The superintendent makes a test run with the new machine on 500 cakes and finds that 28 cakes are improperly iced.

- a. Would you advise the superintendent to buy the new machine?
- b. What do the errors I and II mean in this problem?
- c. What are the costs of making these errors?
- d. How do these costs affect your statistical decision rule?
- e. If the new machine actually ices cakes with 10 per cent bad icings, what is the chance that the superintendent will buy the machine if he uses your statistical decision rule? (*Hint: This is one point on the OC curve.*)

4.6 Sample Size in Testing Hypotheses

The horse-racing problem used to introduce this course in statistics was finally solved as follows in Sec. 3.4.

$$\begin{array}{lll} H_0: P_0 = .5 & n = 10 & R: x \leq 2 \\ H_1: P_1 < .5 & \alpha = .055 \end{array}$$

Decision Rule. Accept H_1 and do not invest in the scheme if 2 or fewer winners are observed in a sample of 10 races; otherwise invest.

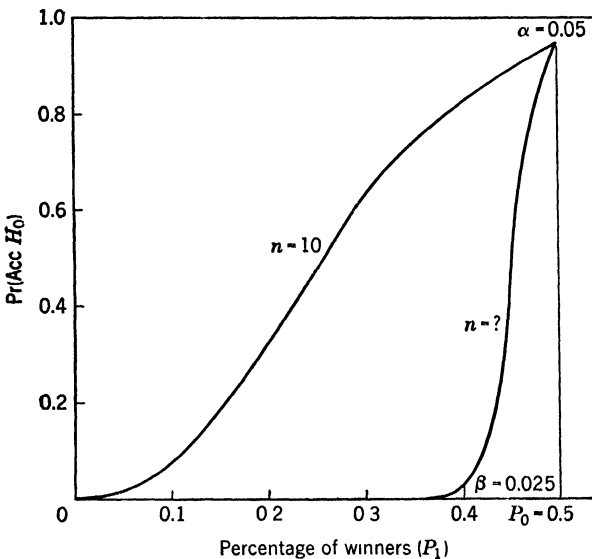


FIG. 4.6. Sample size of a specified OC curve.

The OC curve of Table 3.2 and Fig. 3.1 is reproduced in Table 4.2 and Fig. 4.6. If the system wins in 50 per cent of the races in which it is applied, the probability of accepting H_0 and investing money is .945 ($1 - \alpha$). If the system wins in only 40 per cent of the races, the probability of investing money is .833. If the system wins in only 25 per cent of the races, the probability of investing money is .474. These probabilities, though a minimum with respect to the location of the region of rejection, are still large, but may be reduced by taking a larger

sample, as shown in Sec. 3.7 where the OC curves for $n = 10, 50$, and 100 were compared in Fig. 3.3. How large a sample should be selected?

The number of observations depends upon the specification of four quantities: P_0 and α , and P_1 and β . For example, let the specifications of a proposed horse-race test be the following:

$$\begin{aligned} P_0 &= .50 & \alpha &= .05 \text{ (approximately the same as with } n = 10) \\ P_1 &= .40 & \beta &= .025 \end{aligned}$$

These quantities have the following meaning: If 50 per cent winners are picked by using the system, the probability of accepting H_1 and not investing in the system is controlled at .05. If only 40 per cent winners are selected by using the system, the probability of accepting H_0 and investing in the system is only .025. Both α and β are fixed at small values. The test specifications are two points on an OC curve, as shown in Fig. 4.6 and in Table 4.2. What sample size is necessary to make the OC curve pass through these two points? The sample size n may be determined from the basic equation for α and β using the normal approximation to the binomial.

TABLE 4.2. OPERATING CHARACTERISTICS OF HORSE-RACE TEST

$H_0: P_0 = 5$	$n = 10$	$H_0: P_0 = .5$	$\alpha = .05$
$H_1: P_1 < 5$	$\alpha = .055$	$H_1: P_1 = 4$	$\beta = .025$
$R: x \leq 2$		n and R unknown	

P	$Pr(\text{Acc } H_0)$	$Pr(\text{Acc } H_0)$
$P_0 = 5$	$1 - \alpha = .945$	$1 - \alpha = .95$
$P_1 = 45$	$\beta = .90$	
40	.83	$\beta = .025$
35	.74	
30	.62	
25	.47	
20	.32	
15	.18	
.10	.07	
.05	.01	
.00	.00	

$$\alpha = \Pr(x \leq x_r | n, P_0) \quad \text{and} \quad \beta = \Pr(x > x_r | n, P_1)$$

Using the normal approximation to the binomial, the values of Z_α and Z_β are (for R in the lower tail),

$$-Z_\alpha = \frac{x_r - nP_0}{\sqrt{nP_0Q_0}} \quad \text{and} \quad Z_\beta = \frac{x_r - nP_1}{\sqrt{nP_1Q_1}}$$

These are two equations in two unknowns n and x_r . The solution to these two equations follows:

$$x_r - nP_1 = Z_\beta \sqrt{nP_1Q_1} \quad (4.4)$$

$$x_r - nP_0 = -Z_\alpha \sqrt{nP_0Q_0} \quad (4.5)$$

Subtracting Eq. (4.4) from Eq. (4.5),

$$n(P_0 - P_1) = Z_\alpha \sqrt{nP_0Q_0} + Z_\beta \sqrt{nP_1Q_1} \quad (4.6)$$

$$n(P_0 - P_1) = \sqrt{n} (Z_\alpha \sqrt{P_0Q_0} + Z_\beta \sqrt{P_1Q_1})$$

$$\sqrt{n} = \frac{Z_\alpha \sqrt{P_0Q_0} + Z_\beta \sqrt{P_1Q_1}}{P_0 - P_1}$$

$$n = \left(\frac{Z_\alpha \sqrt{P_0Q_0} + Z_\beta \sqrt{P_1Q_1}}{P_0 - P_1} \right)^2 \quad (4.7)$$

The value of x_r may be found by substituting in either Eq. (4.4) or Eq. (4.5).

Remark 1. Z_α and Z_β are always positive as derived.

Remark 2. If the alternative hypothesis is $H_1: P_1 > P_0$, the denominator of the formula for n becomes $P_1 - P_0$, and the signs of Z_α and Z_β are changed in Eq. (4.4) and Eq. (4.5).

For the proposed test,

$$P_0 = .5 \quad \alpha = .05 \quad Z_\alpha = 1.645$$

$$P_1 = .4 \quad \beta = .025 \quad Z_\beta = 1.96$$

$$n = \left(\frac{1.645 \sqrt{(.5)(.5)} + 1.96 \sqrt{(.4)(.6)}}{.5 - .4} \right)^2 = (17.83)^2 \doteq 318$$

With $n = 318$ and the region of rejection in the lower tail ($P_1 < P_0$), x_r is determined from Eq. (4.5).

$$x_r - 318(.5) = -1.645 \sqrt{318(.5)(.5)}$$

$$x_r = 159 - 1.645(8.91) \doteq 144$$

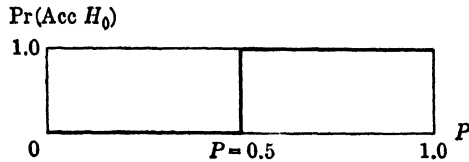
Decision Rule. Accept H_1 and do not invest if the observed number of winners is equal to or less than 144; otherwise, accept H_0 and invest.

A part of the planning of a statistical experiment which will yield a decision rule for deciding between two alternative hypotheses is a determination of the number of observations in the sample. The number of observations needed in the experiment depends upon the specification of the OC curve, i.e., two points on this curve. An $n = 318$ in the example is not an exact sample size. In a practical experiment, 300 observations would be used, or possibly 325 or 350. The magnitude of n is a *rough approximation* to the number of observations needed. If the experimenter is thinking of using 100 observations, he finds out that this is

not a large enough sample. If he is thinking of using 1000 observations, he finds that this is too many and a smaller sample will suffice. An important contribution of statistics to scientific decision making is the method of determining the sample size from a specification of the operating characteristics of a desirable statistical test.

4.7 Statistical Significance and Practical Significance

The ideal OC curve for the horse-race test is shaped like this:



For all values of P less than .5, the probability of accepting H_0 and investing money is zero; for all values of P equal to or greater than .5, the probability of accepting H_0 and investing money is one. Theoretically, this OC curve exists only when the whole population is examined. The larger the sample size, the sharper is the discrimination between acceptable and unacceptable values of P .

Remark. This function is only an ideal even when the whole population is sampled because mistakes in classification can be and often are made

In choosing the proper sample size for testing hypotheses, one of the determinants is a value of P_1 under the alternative hypothesis. The claim for the horse-racing system is 50 per cent winners; therefore anything less than 50 per cent winners means it is a losing system. As a *practical matter*, what percentage of winners is a borderline case between a good and a bad system? For what values of P_1 should the probability of accepting H_0 and investing money be small? Should 49 per cent winners be avoided? 48 per cent? The practical unacceptable value may be 40 per cent, as in the proposed sampling plan of Sec. 4.6. If the system wins this percentage of races, the probability of investing money in the scheme is fixed at some small value.

The probabilities of making errors in sample testing of hypotheses cannot be made equal to zero, but they can be controlled at small values through the proper choice of the sample size. The determination of the borderline value of P which specifies unacceptable quality or a value to be avoided is not simply a problem in statistics. Such a decision is often difficult to make, but once the value has been selected, the sample size is readily computed.

A discrimination between two values of P which are very close together (say, $P_0 = .5$ and $P_1 = .49$) can always be made in a statistical test, pro-

vided a large enough sample is taken. Very large samples are usually very costly. By determining the sample size in the planning stage prior to actually conducting the experiment, attention is directly focused on the sample size and the cost of attaining such a sharp discrimination. If the cost is too high, either the experiment must be abandoned or the sharpness of the discrimination lessened by increasing the difference between P_0 and P_1 . Too many experiments are carried on with arbitrary sample sizes. In some instances, the samples are not large enough to discriminate satisfactorily probabilitywise between acceptable and unacceptable values of P . Sometimes, the samples are too large and the experiment could have been conducted more cheaply and still have provided the necessary operating characteristics had they been specified in advance.

4.8 Acceptance Sampling in Industrial Quality Control

An industrial application of the methods of testing hypotheses is in acceptance sampling in quality-control procedures. Each of the concepts of the more general tests of hypotheses is given a special and standardized name.

Population: A *lot* consisting of a large number of product units produced under the same conditions.

P : The *quality level* of the lot.

Accept H_0 : Accept the lot, which means to *buy* the lot.

Reject H_0 : Reject the lot, which means *not to buy* the lot.

P_0 : *Acceptable quality level* (ΔQI), the worst quality at which the consumer is willing to accept (buy) almost all lots.

P_1 : *Lot tolerance percent defective* (LTPD), the best quality at which the consumer is willing to accept (buy) almost none of the lots.

α : *Producer's risk*, a measure of the risk that the producer incurs of rejecting good lots.

β : *Consumer's risk*, a measure of the risk that the consumer incurs of accepting bad lots.

4.9 How to Judge Sample Results

The very first section of this book demonstrated two different ways of making statistical decisions: planning in advance what values of the random variable will lead to a decision to take one or the other course of action, and judging the sample results after they have occurred in order to reach the proper decision. Up to this point, this book has emphasized the planning of experiments. These statistical plans take the form of deciding how many observations are necessary and making a decision rule which takes into account both errors I and II and their associated

risks, α and β . This section will explore the problem of judging sample results after they have occurred.

In large samples, the normal distribution is used as an approximation to the binomial distribution to test hypotheses about proportions. The first step in selecting the region R for a random variable x is to find the appropriate region R in the standard normal distribution $f(Z)$. This region is placed in one or both tails, depending upon the alternative hypothesis, and the value or values of Z are selected according to the arbitrary probability α . Thus, if the test is a one-tail test with all the region R in the upper tail ($H_1: P_1 > P_0$), and $\alpha = .025$, the appropriate value of $Z_\alpha = 1.96$. The region $R: x \geq x_r$ is determined from the relationship between x , and Z_α , namely,

$$Z_\alpha = \frac{x_r - nP_0}{\sqrt{nP_0Q_0}}$$

The relationship is shown in Fig. 4.7a.

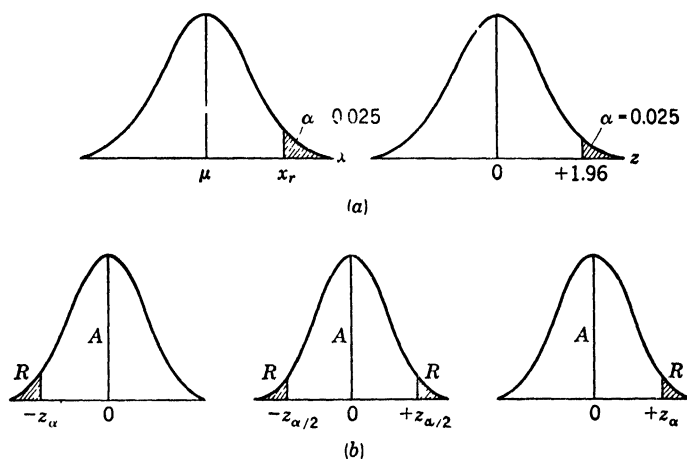


FIG. 4.7. Decision rules stated in terms of z . (a) Relationship between z and x_r . (b) Regions R and A in the z distribution

The characteristic of the sample can also be defined in terms of the variable Z and the decision rule stated in terms of Z . The range of values of Z is divided into two parts, R and A , such that if the sample Z falls into R , the decision is to accept H_1 and if it falls into A , the decision is to accept H_0 . In the example given above, $Z = 1.96$ is the dividing line between R and A . The decision rule is stated: if $Z \geq 1.96$, accept H_1 ; if $Z < 1.96$, accept H_0 . The sample value of Z is found by inserting the sample value of x in the equation

$$Z = \frac{x - nP_0}{\sqrt{nP_0Q_0}}$$

The location of the regions R and A in the Z distribution is the same as in the distribution of x for the same alternative hypothesis H_1 . These regions are shown in Fig. 4.7b. The value of Z which divides R and A is labeled Z_α to correspond to the probability level of α . That is, Z_α is the value of Z which cuts off the probability α (or $\alpha/2$) in the appropriate tail of the normal distribution. The three statistical decision rules which apply to the three alternative hypotheses are:

1. $P_1 < P_0$: Accept H_1 if $Z \leq -Z_\alpha$; otherwise accept H_0 .
2. $P_1 > P_0$: Accept H_1 if $Z \geq Z_\alpha$; otherwise accept H_0 .
3. $P_1 \neq P_0$: Accept H_1 if $Z \geq Z_{\alpha/2}$ or $Z \leq -Z_{\alpha/2}$; otherwise accept H_0 .

Example 4.16. The voter problem of Sec. 4.5 is solved by using the decision rule stated in terms of the variable Z .

$$\begin{array}{lll} H_0: P_0 = .5 & \alpha = .05 & Z_\alpha = -1.645 \\ H_1: P_1 < .5 & n = 400 & x = 190 \text{ (observed)} \end{array}$$

Decision Rule. Accept H_1 if the sample Z is equal to or less than -1.645 ; otherwise accept H_0 .

$$\begin{aligned} \text{From the sample, } Z &= \frac{190 - 400(.5)}{\sqrt{400(.5)(.5)}} \\ &= -1.0 \end{aligned}$$

Decision. Accept H_0 because $Z = -1.0$ falls in the region A which accepts H_0 . This is the same decision reached in the previous analysis.

Example 4.17. The china sales problem of Sec. 4.5 may be set up as follows:

$$\begin{array}{llll} H_0: P_0 = .5 \text{ (sales } A) & \alpha = .05 & \frac{\alpha}{2} = .025 & Z_{\alpha/2} = \pm 1.96 \\ H_1: P_1 \neq .5 \text{ (sales } A) & n = 256 & x = 150 \text{ (observed)} & \end{array}$$

Decision Rule. Accept H_1 if $Z \leq -1.96$ or $Z \geq +1.96$; otherwise accept H_0 .

If the observed number of sales of pattern A is 150,

$$Z = \frac{150 - 256(.5)}{\sqrt{256(.5)(.5)}} = 2.75$$

Decision. Accept H_1 because Z falls in the region R which accepts H_1 .

The mechanics of this method of formulating statistical decision rules are quite simple, but the ideas developed in the more detailed analysis of planning experiments are necessary in order to give the proper consideration to the two errors which can be made and to α and β . The real danger in using this method of judging the sample results *after the*

sample has been taken is that one forgets to plan for the proper sample size and to ask anything about the operating characteristics of the test. The entire emphasis is upon error I and setting α , but α is only one part of the whole problem of decision making. Also, with this method, the tendency is to think in terms of accepting and rejecting hypotheses and not in terms of taking specific actions as a result of arriving at a decision to accept one hypothesis or the other.

The two methods are equivalent for the same experiment, i.e., the same n and α . Planning and judging according to the statistical principles set forth here amount to the same thing. The planning of experiments is a preferable procedure because attention is directly focused upon the alternative hypothesis, upon all possible values of the true population proportion, and upon the risk involved in taking action for a wide range of population values. In judging the sample results, one may often forget to inquire what the consequences will be of a particular decision rule for many possible alternative population values.

EXERCISES

4.13. A wholesaler offers a shipment of 12,000 assorted items, which normally sell for from \$2.50 to \$4.95, to the head of a chain of jewelry stores at a discount from the regular wholesale price. The jewelry store could feature these items at a special pre-Christmas sale at prices from \$1 to \$1.79 and still make a substantial profit because of the large number of items. The jewelryman estimates that he must sell 95 per cent of all items bought in order to avoid a large leftover stock of merchandise. He also estimates that if only 88 per cent of the items are sold he does not want to undertake the venture. Before buying the large assortment of items, the jewelryman wants to experiment with their salability in his stores. He therefore persuades the wholesaler to sell him a small number of items so that he can test consumer response before buying the whole shipment.

a. What are the consequences of the two kinds of error which may be committed?

b. Assuming an alpha of .02 and a beta of .05, what sample size is necessary for testing these hypotheses?

c. If the wholesaler will sell only in multiples of 100 items, what number of items should the jewelryman order for his pretest?

d. If 186 of these items are sold during the pretest of the sale, would you advise the jewelryman to buy the consignment?

e. Comment on the randomness of the sample from the population being sampled.

4.14. In the experience of a cafeteria, 15 to 20 per cent of the customers eat the "special luncheon." Is the experience of one day when 93 out of 498 customers ate the special luncheon consistent with the previous experience?

4.15. It is assumed in an aircraft-parts manufacturing plant that the weekly turnover of employees should not be more than 50 employees out of the total

employed of about 2500. This is roughly a 2 per cent turnover. If terminations exceed this rate, the situation is serious enough to warrant investigation. One week 60 employees terminated employment. Is this a large enough number to warrant investigation?

4.16. Suppose that you are purchasing several thousand items from a manufacturer. Since you have just finished a course in statistics, you ask him if his company does any sampling inspection to control the fraction of defective items which it sells. A representative of the company replies that indeed it does have sampling inspection of its final product and that its sampling plan for the items you are to purchase has the following characteristics:

AQL: 2 per cent	Producer's risk: .019	$n = 150$
LTPD: 7 per cent	Consumer's risk: .021	$R: x \geq 8$

What do each of these terms mean? Would you suggest changing the test specifications in any way?

4.17. Walnuts are graded according to standards of size, color, shape, and other characteristics which are set by state law. In order to qualify for a grade A label, at least 16 per cent of the grower's crop must meet the specifications of size, color, etc. Obviously, not all walnuts in a crop can be inspected. The sampling procedure in operation consists of selecting a random sample of 100 walnuts from the entire crop with a mechanical selector, grading each one according to the standard specifications, and assigning a grade A label to the crop if 13 or more walnuts meet the state specifications for this label. What is the value of α in this test? As a grower of a crop which is inspected according to this sampling plan, how do you feel about this value? Why? As a consumer of walnuts, how do you feel about this value?

In this sampling plan, if $P = .12$, $\beta = .42$; if $P = .10$, $\beta = .20$; if $P = .05$, $\beta = .002$. What do each of these values mean? As a grower of walnuts, are you content to have your crop inspected with a plan for which these values obtain? As a consumer of walnuts? If not, specify your own sampling plan and determine the number of observations which will meet these specifications. What is the statistical decision rule you arrive at?

4.18. A city is under pressure from consumer organizations to adopt rent controls on houses and apartments which are for rent in that city. The city manager has decided to advocate the adoption of rent controls if there are no more than 5 per cent vacant units (both apartments and houses) in the city. He reasons that if more than 5 per cent of these units are vacant, housing is not critical in the city and rent controls should not be introduced. The city has only enough money in the treasury to finance a sample of 2500 observations, but it would like to spend less money and sample fewer observations, if possible.

Suppose that you were hired as a statistician on this problem, and that this city has a well-defined list of all rental units from a recent census. (In practice, such a list might be very hard to obtain and substitute methods of sampling employed.) Devise a sampling plan which requires no more than 2500 observations. Determine the number of observations you would need and the statistical decision rule to be followed in recommending an action on rent controls. Take

into account the two errors which can be made and the consequences of making them.

SUPPLEMENTARY READING

Moroney, M. J.: "Facts from Figures," Chaps. 7 and 12, Penguin Books, Inc., Baltimore, 1953.

Neyman, Jerzy: "First Course in Probability and Statistics," Chaps. 1 and 5, Henry Holt and Company, Inc., New York, 1950.

CHAPTER 5

STATISTICAL INFERENCE: ESTIMATION OF A PROPORTION

Many practical problems in statistics are not concerned with testing hypotheses or claims about a population proportion because there is no preconceived notion about the value of P . In this case, the investigator simply wants to *estimate* the value of P in the population. For example:

1. A housing authority may wish to estimate the percentage of vacancies in apartment houses.

2. The Gallup Poll estimates the percentage of voters for party candidates.

3. The purchaser of a product may want to estimate the percentage of defects in the items he has bought.

4. A cigarette company may want to estimate what percentage of all smokers smoke its brand of cigarette.

In none of these problems is there any claim or hypothesis about the unknown percentage. Instead, the unknown percentage is to be estimated. Sometimes a complete count of the population is made, in the decennial Census of Population, for example, an attempt is made to enumerate certain characteristics of every unit in the population. More frequently, samples are used because they will allow sufficiently precise estimates at a much lower cost.

5.1 Point Estimates

Suppose that 4 winners are observed in a sample of 10 races in Example 1.1. What is the estimate of the unknown percentage of winners based upon this sample? In everyday usage, the percentage of winners in the sample ($p = x/n = 4/10$, or 40 per cent) is commonly used as an estimate of the percentage of winners in the population. Is it justifiable to use the sample percentage as an estimate of the unknown population percentage? Whether it is a justifiable practice depends upon whether the estimate so obtained has the properties desired for an estimate. What prop-

erties are desirable for estimates of this kind? Two are suggested here.

In the first place, it is desirable to have an estimate such that if it were computed for a large number of random samples of the same size from a given population, the average of all the sample estimates would equal the true population value. Such an estimate is said to be *unbiased*. In the second place, it is desirable to have an estimate which will fluctuate from sample to sample as little as possible. If one method leads to an estimate with a standard deviation of, say, 2, and another to a standard deviation of, say, 4, the first method is generally to be preferred to the second. The method of estimation which leads to estimates with the *smallest* or minimum sampling variation is said to be *efficient*.

A class of estimates which are, under very general conditions, both unbiased and efficient, is the *maximum-likelihood estimate*. This estimate is reported as one number or as a *single-valued estimate*. The meaning of maximum likelihood will be explained in terms of a problem. The probability that the sample value ($x = 4$) will occur in the sample ($n = 10$) depends upon the unknown parameter P . In general this probability is written $\Pr(x = 4|n = 10, P)$ and is a term of the binomial distribution:

$$f(4) = \frac{10!}{4!6!} P^4 Q^6 \quad Q = 1 - P$$

The maximum-likelihood estimate of P , call it $\hat{p}$, is that value which, if substituted for P in the probability equation, makes that probability a maximum: $\Pr(x = 4|n = 10, \hat{p}) = \max$. The determination of the maximum-likelihood estimate is a problem for the differential calculus, but a numerical demonstration will illustrate the principle. Binomial probabilities for $x = 4$ and $n = 10$ are given in Table 5.1 for various values of P . These terms are taken from Table A.2.

TABLE 5.1

P	$\Pr(x = 4 n = 10, P)$
	$\Pr(x = 4)$
.1	011
.2	088
.3	200
.39	2503
.40	2508 (maximum)
.41	2503
.5	205
.6	.111
.7	.037
.8	.0055
.9	.000

The maximum probability occurs when $P = .4$; this is also the sample proportion. The maximum-likelihood estimate of P , $\hat{p}$, is the *sample proportion* x/n . This is the best single-valued estimate of P according to the criterion of maximum likelihood. The common practice of using the sample percentage to estimate the unknown population percentage thus has this statistical justification.

The criterion of maximum-likelihood estimation may be confusing at first because it is a conditional statement. A formal definition is, therefore, repeated. A maximum-likelihood estimate of an unknown population parameter is that value computed from the sample which, if it were the true parameter value, would maximize the probability of occurrence of the observed sample. In the simple case of a percentage, the sample percentage is the maximum likelihood estimate of the unknown population proportion P .

5.2 Confidence-interval Estimates of a Proportion

Although it is a common practice, and a statistically justifiable practice, to use the sample percentage as a single-valued estimate of the unknown population percentage, it is unwise to think that the population percentage P is equal to the sample percentage p . Randomness in the selection of the sample will cause p to differ from P . The difference between p and P is often referred to as the *random error*. This difference is not an error in the sense of a mistake; it is merely a difference which occurs because only a sample of observations is taken from which to estimate the population value of P . It seems reasonable, then, to add something to the sample p to make allowance for this random error. A recommended procedure is to make an interval of the form $p \pm$ "something" and to state that this interval includes the true value of P . It will never be known with certainty whether the interval includes P ; the statement that it does include P could be wrong. Therefore, a probability is also attached to the interval—the probability that the interval does include the true value of P .

The longer the interval, the greater is the probability that it contains P ; the shorter the interval, the smaller is the probability that it contains P . But, the longer the interval, the less precise is the estimate of P ; the shorter the interval, the more precise the estimate of P , that is, the narrower the range which attempts to locate P . It is desirable to have a long interval for a high probability that it contains P and a short interval for greater precision of the estimate. The inherent conflict is resolved in Sec. 5.3.

The problem of establishing a confidence-interval estimate of the unknown population proportion P may be stated this way. First, a

random sample is selected. Second, on the basis of this sample, lower and upper limits designated by $\underline{P}$ and $\bar{P}$ are set which form an interval $\underline{P}$ to $\bar{P}$. $\underline{P}$ is called the *lower confidence limit*; $\bar{P}$ is called the *upper confidence limit*. Because of the random sampling variation, the interval may not contain P . The probability that the interval will contain P must, therefore, be set in advance. This probability is called the confidence coefficient and is written C.C. It is an arbitrarily selected probability that the interval contains P . It may be a probability of .90, .95, .50, or any other value.

As an example, suppose that 4 winners are observed in 10 races. The maximum-likelihood estimate of P is 40 per cent. What is the interval estimate which will contain P with 95 per cent confidence? This interval may be read directly from Table 5.2. For $n = 10$ and $x = 4$, the limits are 12 to 74 per cent. The statement that this interval includes the true value of P is made with 95 per cent confidence.

If $n = 100$ and $x = 40$ winners, the 95 per cent confidence interval estimate read from Table 5.2 is from 30 to 50 per cent. If $n = 1000$ and $x = 400$, the observed fraction is .40 and the 95 per cent confidence limits are from 37 to 43 per cent.

Remark. The only table of confidence limits reproduced in this book is Table 5.2 for 95 per cent confidence limits. The following references are to published tables and charts for setting confidence intervals with other values of the confidence coefficient:

C. J. Clopper and E. S. Pearson, The Use of Confidence or Fiducial limits, *Biometrika*, vol. 26, p. 410, 1934.

Wilfrid J. Dixon and Frank J. Massey, Jr., "Introduction to Statistical Analysis," pp 320-323, McGraw-Hill Book Company, Inc., New York, 1951.

Statistical Research Group, Columbia University, Churchill Eisenhart, Millard W. Hastay, and W. Allen Wallis (eds.), "Techniques of Statistical Analysis," Chap. 10, McGraw-Hill Book Company, Inc., New York, 1947.

Several comments are in order at this point. First, the confidence coefficient is selected exactly as α is selected in testing some hypothesis about the value of P . A value of x is observed in a random sample of n observations. What admissible values of P would be accepted by this sample value and what values of P would be rejected by this value of x ? For example, if $n = 10$ and $x = 4$, both $P = 1.0$ and $P = 0.0$ for the population values are rejected, because if $P = 1.0$, x would equal 10, and if $P = 0.0$, x would equal 0 in the sample. The actual occurrence of $x = 4$ rules out both $P = 1.0$ and $P = 0.0$. All those values of P accepted constitute the interval $\underline{P}$ to $\bar{P}$. All those values of P rejected are outside the interval. The confidence coefficient C.C. is equal to $1 - \alpha$. The size of α depends upon the cost of error I: accepting H_1 when H_0 is true. The size of C.C. depends upon how serious it is to form

TABLE 5.2. NINETY-FIVE PER CENT CONFIDENCE INTERVAL FOR BINOMIAL DISTRIBUTION

Number observed x	Size of sample						Fraction observed x/n	Size of sample	
	10	15	20	30	50	100*		250†	1000†
0	0 31	0 22	0 17	0 12	0 7	0 4	00	0 1	0 0
1	0 45	0 32	0 25	0 17	0 11	0 5	01	0 4	0 2
2	3 56	2 40	1 31	1 22	0 14	0 7	02	1 5	1 3
3	7 65	4 48	3 38	2 27	1 17	1 8	03	1 6	2 4
4	12 74	8 55	6 44	4 31	2 19	1 10	04	2 7	3 5
5	19 81	12 62	9 49	6 35	3 22	2 11	05	3 9	4 7
6	26 88	16 68	12 54	8 39	5 24	2 12	06	3 10	5 8
7	35 93	21 73	15 59	10 43	6 27	3 14	07	4 11	6 9
8	44 97	27 79	19 64	12 46	7 29	4 15	08	5 12	6 10
9	55 100	32 84	23 68	15 50	9 31	4 16	09	6 13	7 11
10	69 100	38 88	27 73	17 53	10 34	5 18	10	7 14	8 12
11		45 92	32 77	20 56	12 36	5 19	11	7 16	9 13
12		52 96	36 81	23 60	13 38	6 20	12	8 17	10 14
13		60 98	41 85	25 63	15 41	7 21	13	9 18	11 15
14		68 100	46 88	28 66	16 43	8 22	14	10 19	12 16
15	..	78 100	51 91	31 69	18 44	9 24	15	10 20	13 17
16			56 94	34 72	20 46	9 25	16	11 21	14 18
17			62 97	37 75	21 48	10 26	17	12 22	15 19
18			69 99	40 77	23 50	11 27	18	13 23	16 21
19			75 100	44 80	25 53	12 28	19	14 24	17 22
20	...	...	83 100	47 83	27 55	13 29	20	15 26	18 23
21				50 85	28 57	14 30	21	16 27	19 24
22				54 88	30 59	14 31	22	17 28	19 25
23				57 90	32 61	15 32	23	18 29	20 26
24	...	...		61 92	34 63	16 33	24	19 30	21 27
25				65 94	36 64	17 35	25	20 31	22 28
26	...	...		69 96	37 66	18 36	26	20 32	23 29
27	...	...		73 98	39 68	19 37	27	21 33	24 30
28				78 99	41 70	19 38	28	22 34	25 31
29	...			83 100	43 72	20 39	29	23 35	26 32
30	...			88 100	45 73	21 40	30	24 36	27 33
31					47 75	22 41	31	25 37	28 34
32					50 77	23 42	32	26 38	29 35
33					52 79	24 43	33	27 39	30 36
34					54 80	25 44	34	28 40	31 37
35					56 82	26 45	35	29 41	32 38
36					57 84	27 46	36	30 42	33 39
37					59 85	28 47	37	31 43	34 40
38	..				62 87	28 48	38	32 44	35 41
39					64 88	29 49	39	33 45	36 42
40	...				66 90	30 50	40	34 46	37 43
41					69 91	31 51	41	35 47	38 44
42					71 93	32 53	42	36 48	39 45
43	...				73 94	33 53	43	37 49	40 46
44					76 95	34 54	44	38 50	41 47
45					78 97	35 55	45	39 51	42 48
46					81 98	36 56	46	40 52	43 49
47					83 99	37 57	47	41 53	44 50
48					86 100	38 58	48	42 54	45 51
49					89 100	39 59	49	43 55	46 52
50					93 100	40 60	50	44 56	47 53

SOURCE. This table is reprinted with the permission of the author and the publishers from Table 1.1 of George W. Snedecor, "Statistical Methods," 4th edition, Iowa State College Press, Ames, Iowa, 1946

* If x exceeds 50, read $100 - x$ = number observed, and subtract each confidence limit from 100

† If x/n exceeds 0.50, read $1.00 - x/n$ = fraction observed, and subtract each confidence limit from 100.

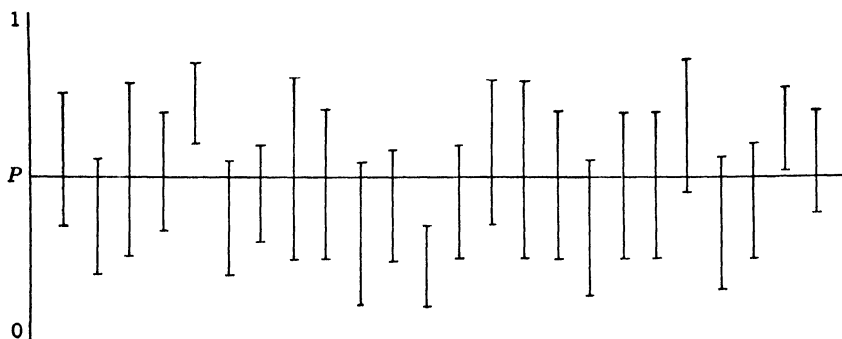
an interval which does not include P or, what is the same thing, how important it is to form an interval which does include P .

Second, what is the meaning of ".95 probability that the interval includes P "? The confidence coefficient, a probability which may be used to predict the outcomes of a physical experiment, is interpreted in terms of the relative frequency of occurrence in the experiment. The physical experiment of an estimation problem consists of selecting a random sample, forming an interval, and making the statement that the interval contains the parameter. The confidence coefficient has meaning only in relation to this procedure. If the sampling process is repeated in exactly the same manner for a large number of samples, if an interval is computed for each sample in exactly the same way, and if the statement is made that each interval contains the true value, the confidence coefficient is a prediction about the fraction of these statements that are true. The truth of any one statement is uncertain; the uncertainty is measured by attaching to it a probability interpreted in terms of a relative frequency of occurrence in a long series of random trials. This interpretation is a subtle and important point.

With a confidence coefficient of .95, 95 per cent of the intervals from a large number of random samples formed in the same manner as in the example will include the true value of P . Or, if the statement is made that each interval contains P , 95 per cent of such statements may be expected to be correct, on the average. Also, the 95 per cent holds true if a person samples from different populations in the course of his lifetime. If the application of this technique is different in each case, 95 per cent is still the fraction of statements which are predicted to be true.

An erroneous, but unfortunately popular, probability statement about the confidence coefficient is frequently made. The statement that the probability is .95 that the true value of P lies in a certain interval (40.1 to 49.9 per cent in the voter problem) is incorrect. P either does or does not lie within the interval. It is a constant and therefore has no probability distribution; no statement can be made about its probability of lying within an interval. Whether it does or does not lie within such an interval can never be known unless the whole population is examined, and there is then no estimation problem. Students should be very careful not to make this incorrect statement about the meaning of the confidence coefficient.

Probabilities must be attached to variables. The parameter P does not vary; the interval varies because it is computed from sample values which vary from sample to sample. A diagram may help to clarify the meaning of confidence. Suppose that the true value of P is located on a scale running from 0 to 1 and a straight line is drawn horizontally at this point as shown below:



In a series of repeated samples, values of p and, therefore, the confidence intervals computed from these values will vary. If these intervals are plotted on the diagram, some will overlap P and some will not. The confidence coefficient is the proportion of these intervals that will overlap the parameter P on the average or in the long run.

Remark. The significance of the correct statement of the meaning of confidence is a subtle point. A very homely example may help. The game of horseshoe consists of throwing a horseshoe at a peg. A ringer occurs when the shoe rings or encloses the peg. What is the interpretation of the statement that the probability is .95 that a ringer will be thrown? Does it mean that 95 per cent of the shoes thrown will ring the peg? Or does it mean that 95 per cent of the time the peg will come up through the horseshoe? The correct statement is fairly obvious. It is a particularly good example, because the wider the spread of the shoe, the more chance there is of making a ringer, other things being equal.

Third, why is an interval that misses the true parameter ever constructed? Why not make an interval with 100 per cent confidence that it contains the true value of P ? Let the random variable be the proportion of races that are won, as in the example above. What is the 100 per cent confidence interval? The interval from 0 to 100 per cent is absolutely certain to contain P , but this is a useless answer. Therefore, the interval is shortened from 100 units in length to approximately 10 units in length, and this interval is located along the scale of P (it makes a difference if the interval is from 20 to 30 per cent, 40 to 50 per cent, or 80 to 90 per cent) with a loss of only 5 per cent confidence.

EXERCISES

5.1. How could you make a shorter interval and have more confidence that it contains the true parameter?

5.2. For the same sample size, what happens to the confidence coefficient as the interval is made shorter?

5.3. Suppose that the interval were to be made shorter and shorter until it were to be just the point p , that is, the sample proportion. How much confidence would you have in this point estimate? Do you find it confusing to say that the sample proportion is the best estimate in which you have no confidence? You should not if you know what confidence means here.

5.4. At a university a full load of courses for students is 15 semester hours. The registrar wants to estimate what percentage of students take a full load of work. There are about 10,000 undergraduates in the university and each student has a card in a file. This file constitutes a frame from which to draw a sample. Suppose that you wished to help the registrar on this problem and decided to draw a random sample of 250 students.

a. How would you go about drawing the sample?

b. If you observed that 115 out of the 250 took 15 or more hours' work, what is the 95 per cent confidence interval for the percentage of students taking a full load?

c. Write out the exact meaning of this interval and the confidence coefficient.

5.3 Sample Size in Estimation

If a guess is made that the true proportion of winners that the horse-racing system of Sec. 1.1 will pick is 40 per cent, how large a sample is needed to estimate the true proportion within plus or minus 6 per cent at the 95 per cent confidence level? An answer that is sufficiently good for most purposes can be obtained from Table 5.2 by trial and error.

Suppose $n = 100$. For $P = 40$ per cent, $x = 40$. The 95 per cent confidence interval is from 30 to 50 per cent, which is an estimate of P within plus or minus 10 per cent. A larger sample than this is called for in order to shorten the interval. If $n = 250$ and $x/n = .4$, the 95 per cent confidence interval is from 34 to 46 per cent, which is an estimate within plus or minus 6 per cent. An interval computed from a sample of 250 will not exactly estimate P within plus or minus 6 per cent because of the unpredictable sampling variability. It will be close to plus or minus 6 per cent if the guessed value (.4) is not too far wrong. Table 5.2 can be used to give a rough approximation to the sample size; in practice, perhaps a little larger sample than indicated should be taken if it can be afforded.

Example 5.1. If the percentage of persons owning a particular make of car is estimated to be .20, how large a sample is necessary to estimate the true proportion within plus or minus 4 per cent at the 95 per cent confidence level? With $n = 250$ and $x/n = .2$, the interval is from 15 to 26, or an estimate approximately within plus or minus 5 per cent. If $n = 1000$ and $x/n = .2$, the interval estimate is from 18 to 23 per cent, or within plus or minus 2 per cent. The sample should be a little larger than 250 but it need not be as large as 1000. If they can be afforded, 400 or 500 observations would be satisfactory.

The determination of the sample size depends upon the specification of three things:

1. The length of the desired interval
2. The confidence coefficient
3. An estimate of P

The length of the interval depends upon how closely P is to be located. The confidence coefficient depends upon how important it is to make an interval which includes the true value of P . An estimate of P can always be obtained somewhere. Previous experience with the problem, the opinion of someone with experience, or a small pilot study will all give estimates which are good enough in practice for determining an approximate sample size. Just as in problems of testing hypotheses, if the sample sizes are too large, either the experiment must be abandoned or the specifications of the estimate changed. Without planning beforehand, an arbitrary sample size may lead to an estimate which is too vague (i.e., the interval is too long) to be of any practical value, or it may lead to an interval which is much too short (i.e., it is more precise than is needed). In the first instance, the question is not answered; in the second, the question could have been answered more cheaply.

EXERCISES

5.5. The number of red beads in samples of 10 beads from a bead box filled with 20 per cent red beads was recorded in the order of appearance in Exercise 1.4, part (b). For each of the 100 samples selected, record the 95 per cent confidence interval from Table 5.2. Out of the 100 intervals, how many included or trapped the true value of P ? How many missed the true value of P ? Compare the percentage of intervals trapping P with the confidence coefficient.

5.6. Previous retail experience of the owner-manager of a small shoe store in a city indicates that about 78 per cent of the potential customers who enter his store buy at least one pair of shoes. This owner, now engaged in the retail shoe business in another city, desires to check his current experience with that in the other store. If he wishes to estimate the percentage of customers who enter his store and buy at least one pair of shoes within ± 3 per cent with a 95 per cent confidence interval, how many observations should he take? How would you suggest that he select these observations to get a random sample? Using your sample size, what is 95 per cent confidence-interval estimate if $x = 750$ customers who enter the store and buy at least one pair of shoes? Does the length of this interval meet the specifications?

5.7. A new shampoo was distributed to 257 housewives for their use. Interviewers returned to question them regarding their opinions on the new product. One question was: Would you be likely to buy this shampoo on the basis of what you know about it now? This question was answered in the affirmative by 195 of the 257 housewives. From these data, what per cent of the housewives do you estimate will buy the shampoo?

Remark. This is an actual problem. Later call backs showed that only 10.5 per cent of these same housewives actually did buy the shampoo after finishing the first bottle.

5.8. The furniture department of a large store wished to estimate what percentage of its customers buying leather-top desks enter claims due to natural scars and markings of the leather. What estimate would you make if 404 out of 500 customers are satisfied? (*Note:* Table 5.2 does not carry a column for $n = 500$. Can you approximate the interval estimate without this entry?) As a result of these data, what policy would you recommend to the furniture department?

5.9. Postcard questionnaires were sent to a random sample of 1000 subscribers to a particular magazine asking them if they had received the latest issue of the magazine. This sample showed 54 wrong addresses. Estimate the percentage of addresses needing correction in the magazine's address file and the amount of error in this estimate. After the individuals with incorrect addresses were contacted, the sample showed that 66 individuals did not receive the latest issue of the magazine. Estimate the percentage of the magazine's subscribers who do not receive the magazine, and the amount of error in this estimate. What conclusions would you draw about the failure of subscribers to receive the magazine?

5.10. In a simple random sample of 250 from a population of residents in a small suburban area, 150 had attended high school or college, 71 had not gone as far as high school, and 29 did not answer. Estimate 95 per cent confidence limits for the number of residents in the community who had attended high school or college. Do the results of the sample furnish conclusive evidence that a majority of the residents have attended high school or college?

5.11. If you accepted H_1 in Exercise 3.7, estimate the percentage of buildings that are one-story residences.

CHAPTER 6

STATISTICAL-RESEARCH METHODOLOGY

Enough theory has been developed in the preceding chapters to allow each of you to undertake a research problem of your own in which you either test a hypothesis or make an estimate of an unknown population proportion. This chapter will introduce you to a research methodology by describing a simple problem for the classroom and then summarizing research principles useful in guiding your own work. This same classroom problem will also lead to the further development of statistical theory and applications in succeeding chapters.

6.1 A Classroom Investigation of Students' Weights

Statement of the Problem. A study of the weights of students is a statistical investigation which can be carried on in the classroom and which will illustrate research technique.

Problem. To investigate the weights of students at this university.

Procedure for Measuring Weight. There are two different methods of measuring weight. The first is to weigh students on a scale; the second is to question students about their weight. Using the scale is the preferable procedure because it will give results closer to the "true" weight of the individual. This method is more costly and time-consuming because there is no scale in the classroom, and the less accurate but more convenient procedure of questioning people will be used.

Measurement Procedure. Ask each student to record his weight to the nearest pound.

Population and Sample. According to the statement of the problem, the population studied includes all students at this university. Since it is not possible to ask all students their weight, only a sample will be questioned; those students who were present in statistics class on Sept. 24, 1951, constitute the sample. Each of them was asked to record on a card his weight to the nearest pound. (He was also asked to record his height to the nearest half inch and his age at his last birthday.) The data for 40 students are shown in Table 6.1.

TABLE 6.1. WEIGHTS OF STUDENTS PRESENT IN STATISTICS CLASS ON SEPT. 24, 1951

(Weight recorded to nearest pound)

145	195	150	140
129	160	150	162
132	155	158	155
185	153	135	152
185	155	140	191
142	139	157	155
124	120	188	165
156	147	166	170
148	185	142	165
182	165	180	165

6.2 Critical Examination of the Numbers Generated

Numbers in themselves have no meaning. The number 165 means nothing until it is related to a measuring procedure. It could mean 165 years, 165 dollars, or 165 hours. In the present problem it is supposed to mean 165 pounds—the weight of a particular student measured to the nearest pound. At the outset of any problem, whether conducting your own research or evaluating the research of others, you should critically examine the numbers recorded. There may be reasons inherent in the measuring procedure why the number 165 (or any other number) does not accurately represent the specified weight.¹ Among possible reasons are the following:

1. The student may not have weighed himself recently.
2. His weight may vary with the time of day.
3. Sickness may lead to a currently abnormal weight.
4. The weight may be recorded with or without clothing.
5. The student may not tell the truth about his weight.

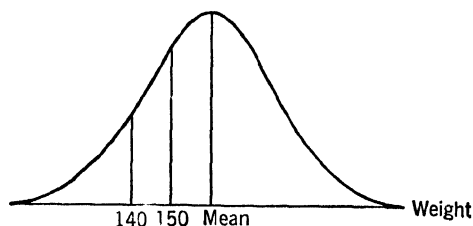
Thus the number (165), though associated with a student's weight, may reflect other things as well.

6.3 Test to See if Weights Were Measured to the Nearest Pound

There is reason to doubt that the questioning procedure results in weights measured to the nearest pound. Since the direct solution of weighing each class member on a scale and comparing his weight with that recorded on the card is inconvenient in the absence of a classroom scale, another test of the hypothesis that weights are measured to the nearest pound must be devised.

¹ The use of weights reported by students in the classroom to dramatize limitations on gathering data was first brought to the author's attention by Prof. Allen Wallis in a course which he offered in the Executive Program of the University of Chicago School of Business.

Imagine a histogram of the "true" weights of a large number of people (men or women). The weights are called "true" because, in general, measurements of this kind cannot be taken perfectly. A histogram is so constructed that the areas of the bars are proportional to the number of items falling in the corresponding intervals on the abscissa. The larger the total number of observations in the whole distribution, the smaller are the intervals used. With very many observations, the intervals become smaller and smaller until the histogram approaches a smooth curve. Statisticians know from prior experience that the distribution almost always has a bell shape like that shown below:



Now consider the interval 140 to 150. If weights are measured to the nearest pound, there is a tendency for slightly more 9s than 1s in this interval; i.e., there are more 149 weights than 141 weights. This is true for all intervals up to the mean. Above the mean, the tendency is just the reverse; there are slightly more 1s than 9s. In the whole distribution, the final digits 0, 1, . . . , 9 will be very nearly equally frequent.

Deduction. In a large number of weighings to the nearest pound, there will be equality in the frequency of final digits.

On the basis of this deduction, four observations are expected for each final digit. The actual distribution of weights by final digit is shown in Table 6.2.

TABLE 6.2 DISTRIBUTION OF WEIGHTS BY FINAL DIGIT

	Final digit										Total
	0	1	2	3	4	5	6	7	8	9	
Observed frequency.	8	1	6	1	1	11	2	2	3	2	40
Expected frequency	4	4	4	4	4	4	4	4	4	4	40

The outstanding characteristic of Table 6.2 is the large number of weights ending in either 0 or 5. Whereas 8 observations (or 20 per cent) are expected in these two cells, 22 observations from the sample (or 55 per cent) actually occur in them. This suggests a tendency to round to the

nearest 5 pounds. How should the difference between the observed and expected frequencies be judged? Is the difference consistent with the hypothesis that weights are measured to the nearest pound?

The real hypothesis to be tested is this: do the numbers represent weights to the nearest pound? As in most practical problems in statistics, this hypothesis is not statistical because it does not deal with a probability distribution of a random variable. Since the hypothesis is both nonstatistical and inconvenient to test, it has been changed to one more easily tested: if weights are recorded to the nearest pound the final digits 0, 1, . . . , 9 will be equally frequent, with some sampling variation. Since there are 10 possible final digits, 20 per cent of those in a large number of weights recorded to the nearest pound should be either 0 or 5, and 80 per cent some other number. A null hypothesis based upon this deduction from the original hypothesis may be stated as follows:

H_0 : In a large number of weights measured to the nearest pound, the proportion ending in 0 or 5 should be 20 per cent (or .2).

This is a statistical hypothesis because (1) the outcome of an experiment in which people record their weights determines a value of a random variable x , the number of weights which end in either 0 or 5; (2) this random variable is known to be binomially distributed and the hypothesis is an assumption about that binomial distribution. The statistical hypothesis to be tested is $H_0: P_0 = .2$, where P_0 is the true proportion of weights which end in 0 or 5. Against what alternative will H_0 be tested? Prior experience has shown that people tend to round their weights to the nearest 5 pounds. A reasonable alternative hypothesis is $H_1: P_1 > .2$.

The two types of possible error are:

- I: Accept H_1 when, in fact, H_0 is true.
- II: Accept H_0 when, in fact, H_1 is true.

What do these errors mean in terms of the experiment, and what are their consequences? To commit an error I means, for the statistical hypothesis, accepting $P > .2$ when, in fact, $P_0 = .2$. Its meaning for the nonstatistical hypothesis is to accept that the weights were not measured to the nearest pound when, in fact, they were measured to the nearest pound. An error II means, for the statistical hypothesis, accepting $P_0 = .2$ when, in fact, $P > .2$. For the nonstatistical hypothesis, it means to accept that weights were measured to the nearest pound when, in fact, they were measured not to the nearest pound but, perhaps, to the nearest 5 pounds.

The consequences of making either of these errors cannot be determined from the mere acceptance or rejection of the hypotheses tested. They must relate to the *action* which will result from the statistical decision. What will be done if H_0 is accepted, and what will be done if H_1 is accepted? To answer these questions, further explanation of practical statistical work is required.

As previously discussed, numbers generated in statistical work must bear some relation to the real world. Statistical variables must be defined, and a definition of a variable or characteristic is acceptable and complete only if it is given in terms of a procedure for measuring that characteristic. A "best way" of measuring a characteristic usually exists, i.e., a method which gives results closer to those that are needed than other methods. Such procedures may well vary from time to time, and may be more time-consuming and expensive than others. The value or values of numbers generated may be considered as "true" values, the ideal toward which experimenters strive. Such a procedure for the weight variable is to weigh each individual on a scale to the nearest pound—a fairly costly, time-consuming, and inconvenient procedure. In practice, then, the procedure of asking each individual to record his weight to the nearest pound is substituted. The shortcomings, possible errors, and biases inherent in this procedure have already been discussed.

The actions to be taken as a result of accepting or rejecting the null hypothesis relate to these procedures of measurement and can be so stated:

Accept H_1 : Conclude that the procedure of asking individuals to record their weights is not an acceptable substitute for the procedure of actually weighing the individuals on a scale. Study a new sample in which students are weighed.

Accept H_0 : Conclude that the substitute procedure is acceptable; continue to use this less costly procedure for the analysis of weights.

The consequences of the two kinds of errors in testing hypotheses may be stated thus:

- I: Accept H_1 when H_0 is true: say that the substitute procedure is not acceptable when it really is acceptable. It is costly to repeat the present experiment and to weigh students on a scale unnecessarily in future experiments.
- II: Accept H_0 when H_1 is true: adopt a procedure which is not a good substitute. An elaborate statistical investigation is carried on with numbers which do not measure the desired characteristic.

In the class problem, it is possible to state the test of significance as follows:

$H_0: P_0 = .2$ (no tendency to round to 0 and 5)

$H_1: P_1 > .2$ (tendency to round to 0 and 5)

$\alpha = .05$ (an arbitrary significance level)

$n = 40$ $x = 22$ (0s and 5s)

Since this is a one-sided test (see H_1), the region R which leads to the acceptance of H_1 is placed in the upper tail of the sample space. From a table of binomial probabilities, for $n = 40$, $P_0 = .2$, and $\alpha = .05$, the value of x_r is 13.

Decision Rule. Accept H_1 and conclude that the substitute measuring procedure is not acceptable if 13 or more of the sample weights end in 0 or 5.

Decision. Accept H_1 because the sample value of $x = 22$ falls in the region R which accepts H_1 . In operational terms, the statistician concludes that weights are not reported to the nearest pound and rejects the procedure of asking people their weights as an acceptable substitute for the procedure of weighing them on a scale. The action to be taken is to repeat the present experiment using a scale and to adopt a scale in future experiments involving weights.

6.4 Reexamination of the Classroom Research Problem

As a result of the analysis indicating that the recorded weights are probably not measured to the nearest pound, the whole research problem should be reformulated. It is necessary to define more exactly the concept of weight and how it is to be measured; to specify precisely both the population to be studied and the sample to be drawn from that population. What weight can be defined in terms of the operation for measuring it. The definition of weight is as follows: (1) It is obtained by weighing each student on a scale to the nearest pound. (2) The scale to be used is the one located in the student health office. (3) The weight is to be recorded in the morning after breakfast on a specified day. (4) The individual is to be weighed without his clothing as in a physical examination. (5) The same scale is to be used for men and women.

The population is defined as all students regularly enrolled on a full-time basis in the university by the registrar as of a given date during the semester. Each student who meets this specification is one sampling unit in the population.

Which students are to be included in the sample or how should the sample be taken? Statistical decisions assume *randomness* in the selection of the observations comprising the sample. The making of decisions based upon probabilities is strictly valid only if the assumption of randomness is met. Where, in the weight data, did the element of ran-

domness enter? The answer is, of course, that it did not enter at all into the process of selecting the sample. In fact, it is doubtful whether there was any planning in the process of selection. Those students whose weights were recorded happened to be in class that day and were selected because they were conveniently at hand. It is important to recognize that this sample is not a random sample. When the experiment is repeated, randomness should be ensured in the sampling process. A random sample can easily be obtained from the specified population because the registrar's list of all students is an excellent frame from which to draw a sample using a table of random numbers.

Finally, how large a sample should be taken? The basic principles of determining sample size in testing hypotheses have been discussed. The same hypothesis used when students were questioned about their weight can be tested here, even though the new numbers generated will represent weight to the nearest pound. Let the test specifications be:

$$\begin{array}{lll} H_0: P_0 = .2 \text{ (0s and 5s)} & \alpha = .05 & Z_\alpha = 1.645 \\ H_1: P_1 = .3 \text{ (0s and 5s)} & \beta = .025 & Z_\beta = 1.96 \end{array}$$

These test specifications mean that if no tendency exists to report weight to the nearest 5 pounds ($P_0 = .2$), the probability of accepting H_1 is .05 (α). If the tendency to round to the nearest 5 pounds is such that 30 per cent of the weights reported are rounded to the nearest 5 pounds ($P_1 = .3$), the probability of accepting H_0 is .025 (β). $P_1 = .3$ specifies a percentage of weights rounded to 5 pounds which is to be avoided.

The value of n is found from Eq. (4.7).

$$\begin{aligned} n &= \left(\frac{Z_\alpha \sqrt{P_0(1-P_0)} + Z_\beta \sqrt{P_1(1-P_1)}}{P_1 - P_0} \right)^2 \\ &= \left(\frac{1.645 \sqrt{(.2)(.8)} + 1.96 \sqrt{(.3)(.7)}}{.3 - .2} \right)^2 = 242 \end{aligned}$$

To find the region R , solve this equation for x_r .

$$\frac{x_r - nP_0}{\sqrt{nP_0(1-P_0)}} = Z_\alpha \quad \frac{x_r - 242(.2)}{\sqrt{242(.2)(.8)}} = 1.645 \quad x_r = 58.6 \doteq 59$$

Decision Rule. Accept H_1 if $x \geq 59$; otherwise, accept H_0 .

The 242 students in the sample may be selected by numbering each student in the registrar's list and selecting 242 numbers from a table of random numbers. No data have been gathered for the reformulated weight problem, but the preliminary investigation has served a purpose by indicating a reasonably good procedure to follow in solving the problem. Although simple and not very practical, the procedure has implications for more general statistics problems because:

1. It emphasizes the action taken as a result of statistical decisions and not merely the acceptance of hypotheses.
2. It shows the importance of establishing a relationship between statistical hypotheses and corresponding nonstatistical hypotheses.
3. Many external statistical data used in business research—advertising and marketing—and in government surveys which sample human populations use substitute methods like asking people on questionnaires or in interviews instead of a more costly and often impossible measurement procedure.

EXERCISES

- 6.1. Test the hypothesis that people do not tend to round to the nearest 5 pounds when asked to record their weights to the nearest pound against the alternative hypothesis that they do tend to round to the nearest 5 pounds. Use the data recorded in Exercise 2.24.
- 6.2. Give an example from your everyday experience in which a number reported by an individual as a result of having been asked about it may differ from the true value.
- 6.3. With the increased use of petroleum products, especially gasoline, in recent years, concern has developed that consumption may deplete the available supplies, and data on the extent of oil reserves are periodically published. What does the term “oil reserve” mean to you? When you see a reference to oil reserves, what is called to mind? See if you can find a definition of oil reserves. Does it agree with your own idea about the meaning of the term? Does it indicate that the total amount of petroleum under the earth’s surface is larger or smaller than the “reserves?”

6.5 Some Elementary Research Principles

Although this section will take up very general principles of research, it will also from time to time refer specifically to your individual research problem. Although it may sound like a manual on “How to Do It,” this popular but offensive form of writing can hardly be avoided in this instance.

Statement of the Problem. State and restate, define and redefine the problem. Establish limits within which you will confine the problem. Define carefully the population you want to study. (Your problem: Use of the binomial distribution involves a dichotomy. What dichotomy? What elementary unit? What collection of units form the population?)

Formulate a Hypothesis. A hypothesis is a trial approach to a problem. Many hypotheses are developed by analogy with other problems of similar nature. An imagination is most important. The hypothesis must pose a precise and testable question since it will be subject to empirical test. Imagine the kind of data which can answer the ques-

tion. Can you get these data? If the hypothesis itself is not directly testable, certain logical deductions from it may be empirically testable. (Your problem: Make a statement about the exact proportion of the dichotomy in the population. $H_0: P_0 = ?$)

Apparatus for Measuring the Variable. Specify exactly the conditions under which you will measure the variable. What kind of an apparatus will you use to generate a number? Specify a procedure for measurement which will give you true values. If this procedure is too costly to use, choose another from the available alternatives. State exactly what you think it measures. Few apparatuses will measure exactly what one wishes; they will give only approximations.

Conduct a sort of dress rehearsal in which you generate some data, i.e., pretest your measuring apparatus. Examine your numbers in an effort to find out what the apparatus measures. Even in so simple a problem as a binomial classification, a practical definition of categories is often complicated and it may be necessary to amend definitions as a result of the preliminary run. (Your problem: What is the dichotomy? How is an object to be identified with a particular category?)

Formulation of Alternative Hypotheses. In testing statistical hypotheses, the alternatives to the main hypothesis usually take the form of alternative values of the parameter. How large a difference in the parameter is worth looking for? This depends upon the use to which the experimental results are put. Ideally, the alternative should be specified in advance and not after examining the sample results. It can usually be based upon prior knowledge or other experiments. (Your problem: Set the values of P_1 under H_1 .)

Risks and Costs of Error. State the errors which can be made in the decisions based on the test, and the costs involved in making these errors. Some prior knowledge can certainly be brought to bear on the seriousness of the errors. Statistics does not replace judgment, and judgment must be used in selecting the more or less arbitrary values of alpha and beta. Set alpha in advance and use the proper test procedure in controlling beta, or else, set both alpha and beta in advance and determine the appropriate sample size. Set one or both in advance because it is too easy to account for an unpopular experimental result by shifting alpha and beta after the experiment is over. (Your problem: Specify alpha with P_0 and beta with P_1 ; determine n and the region of rejection in testing a hypothesis. If estimating an unknown percentage, set the aimed-at precision of the estimate, the confidence coefficient, and determine n .)

Sampling Procedure. Valid statistical inferences depend upon randomness in selection of the sample. Sampling variation affects conclusions and these sampling variations can be controlled and measured only

with probability samples. A researcher with a vested interest in a problem cannot be trusted to select by judgment the sample to be used in the test.

A simple random sample may be selected from a frame or a list of all the units in the population by numbering the list and selecting a sample from a table of random numbers. For example, if there are 10,000 students on the registrar's list and $n = 400$, each student is given a number, a sample of 400 numbers is selected from the table of random numbers, and the 400 students identified are placed in the sample. This method may be costly because of the need to search the frame for special numbers.

In practice a substitute method of selecting the sample is often used. Since there are 10,000 students and $n = 400$, a sample of 10,000/400, or 1 in every 25 students, is needed. Select a *random number* between 1 and 25, e.g., 14. Student 14 is in the sample and every twenty-fifth student following him: 39, 64, 89, etc. This method is called *systematic random sampling*. It is often a good substitute for random sampling if there is no relationship between the order of the individuals in the frame and the variable studied. In this problem, is the number of hours a student carries related to his alphabetical listing? Additional comments on systematic random sampling are in Sec. 6.8.

In much practical work in problems related to business, adaptation of the experimental technique to obtain random samples is not possible. The random variables are often beyond the control of the researcher because observations are generated by a process which he has not set up himself; he must take the observations as they come and adapt the statistical theory to the circumstances. If it is not possible to select a random sample either by using a table of random numbers or by a systematic selection, what case can be made for the randomness of the sample? Is a group of students walking across the campus or a group of customers entering a store a random sample about which probability statements may be made? Section 6.8 of this chapter discusses this problem more fully.

The Research Record. Keep a notebook in which you record everything about your research project. Write down all ideas which occur to you so that you will not forget them. Record your data systematically, and label all data clearly lest you forget what the numbers mean. Be aware of the problems which arise and the decisions which you make at every stage of your research problem. Most of all, be critical of your own work and try to find flaws in it.

The Research Report. The final step in a research project is writing the research report which describes your problem and its solution for others. The type of report varies with the nature of the research and the audience to whom the report is addressed. The report should include

a carefully written explanation of each of the steps taken in the solution: statement of the problem, method of analysis, data, and conclusions. Tables and graphs are useful in presenting the data. Reference should be made to specific works consulted in connection with the problem. Sufficient detail should be given so that the reader knows exactly what was done to lead to the conclusions stated in the report.

6.6 Sources of Information

Finding the necessary data is very often a formidable problem in business research. No one is acquainted with all sources of information. The areas of interest in the fields of business and economics (as in other fields) are so numerous and so broad as to make a listing of all possible data sources impracticable. Therefore, only very general methods and publications will be mentioned.

Data may be generally classified as primary or secondary. Primary data are measured or observed by the original investigator; secondary data are those which have been recorded by someone else. As an example, the weight data gathered in class are primary data; if they were copied from the records of the student health service, they would be secondary data. Business research uses both kinds.

Primary Data. One method of collecting primary data is direct measurement or enumeration. Whatever the conditions for generating a number, the investigator determines the number within the limits of those conditions. For example, he may count the number of females in a group, or the number of vacant houses in a selected area. He may measure the diameter or the length of a steel bar produced by a machine.

Another method is the questionnaire; the data are generated by asking people questions. This may involve (1) mailing a questionnaire for a respondent to fill out at his leisure; (2) personally presenting a respondent with a questionnaire (e.g., an employment interview); (3) interviewing a respondent on a schedule of questions which he is to answer (Gallup Poll or Census). Interviewing usually involves less formal questions than those on a written questionnaire and is sometimes considered a separate method of obtaining data. Although interviewing does present special problems, it is here discussed as a variation on the questionnaire technique.

In soliciting responses to a questionnaire or interview, the employment of certain guiding principles does not automatically ensure success. Firsthand training and experience are necessary. One experience of conducting an actual interview or sending out a questionnaire is of more value than any textbook discussion. Again the text becomes a brief manual, this time of "what to do" rather than "how to do it." The principles are given for use both in personal research problems and in

reviewing the research of others. An intelligent comprehension of statistics is cultivated by recognizing the problems which arise, seeing the pitfalls ahead, and learning to ask the right questions.

Principles for Questionnaire Writing

1. Undertake a preliminary investigation to determine what answers are needed and what questions are likely to elicit those answers.
2. Not all questions can be included because a questionnaire, to be manageable, must be limited in size. Choose the most important and practicable questions.
3. Seek for clarity and simplicity in the questions. The language must be simple and familiar to the respondent. Questions must be readily and accurately answerable.
4. Be sure the concepts involved in the questioning are clearly defined. What is meant by income, unemployed, or child, for example?
5. Avoid leading or offensive questions. Do not be unduly inquisitive (unless, of course, the questioning is for a Kinsey report). Avoid questions which are emotionally charged or involve issues of pride or prestige.
6. Answers to the questions should be objective and capable of tabulation. Leave a proper space for answers.
7. Carefully plan and arrange the questionnaire. There should be a logical flow of thought, usually from a very general beginning to a more specific ending. Often questions are asked to give the interview continuity, even though the answers are not relevant to the study.
8. Pretest the questionnaire on a small sample, not to get information, but to test the questions.
9. Arouse interest in the study.

Mail Questionnaires

1. The questions should be self-explanatory.
2. There should be directions for filling in answers.
3. Questions should be simple. Questions in which the respondent checks one of a multiple choice are best.

Are you a veteran? Yes _____ No _____

Do you own a dog? _____ cat? _____ parrot? _____, etc.

4. A brief covering letter which explains the purpose of the survey is often helpful in eliciting response. Sometimes inducements to return the questionnaire, like money, books, merchandise, etc., are used. Include a self-addressed envelope.

Interviews

1. The selection, training, and supervision of interviewers is an important problem, for interviewers contact the respondents. They must be

able to elicit the required information and tabulate it properly for the use of others.

2. Instructions to interviewers, especially definitions, must be complete, concise, and clear.

3. Interviewing is a social process in which skill must be developed by practice. Training alone is not enough.

4. Since the interviewer can record answers in the proper category, the questions need not be asked in the multiple-choice form which is preferable in the mail questionnaire.

6.7 Sources of Secondary Information

Sources of business information are too numerous to be listed in detail. Knowledge of how to use a library is essential. The reference librarian and the card catalogue should be consulted for information about specific sources.

Professional journals are published in almost every field of interest. A researcher should be familiar with the more important publications in his major field of interest. A useful reference is Eillen C. Graves (ed.), "Periodicals Directory, A Classified Guide to a Selected List of Current Periodicals, Foreign and Domestic," 6th ed., R. W. Bowker Company, New York, 1951.

Journals primarily concerned with the theory and application of statistics are:

1. *Journal of the American Statistical Association*
2. *American Statistician*
3. *Annals of Mathematical Statistics*
4. *Biometrics*
5. *Biometrika*
6. *Applied Statistics*
7. *Journal of the Royal Statistical Society*
8. *Review of Economics and Statistics*

EXERCISES

6.4. Read and write a short report on one article of interest to you in either the *Journal of the American Statistical Association* or the *American Statistician*.

6.5. Examine one professional journal in your major field of interest. Write a summary of its contents. Does it contain any articles on statistics?

6.6. Locate both a government and a private source of statistical data in your major field of interest. Write down the complete reference and a brief summary of the data it contains.

6.7. During January to May, 1951, the Bureau of Labor Statistics conducted a survey of consumer expenditures as one of the major phases of its revision of the

Consumer Price Index. Questions in this survey on personal-care items are reproduced below.

3. PERSONAL-CARE ITEMS

When did you last buy items listed below? (Give date and amount spent)

Item	Date	Total expense
1. Tooth paste, mouthwashes, etc	_____	_____
2. Shaving soap, shampoos, etc	_____	_____
3. Cosmetics, perfume, hair preparations, lotions, etc	_____	_____
4. Combs, brushes, etc	_____	_____
5. Razors and blades, files, etc	_____	_____
6. Other personal-care items (what kind?)	_____	_____

If you were asked these questions, would you be able to answer them accurately? Are any of the item classifications ambiguous? Would you know exactly what items to include in each group? Would you know which of these items you had purchased within the last week and how much you spent for them?

6.8. A city newspaper sends a roving reporter into the business district to obtain public opinion on a matter of city-wide interest. This reporter is going to walk around the streets asking persons at random a specific question about this matter. The question requires a yes or a no answer. Comment critically on this method. List any factors which indicate that it is a good or a bad method.

6.8 Evaluating Sources of Information

The previous sections of this chapter have indicated that great care must be taken in planning and executing research projects, however simple they may seem. Serious pitfalls are open to the unwary in measurement processes, whether by questionnaire or direct observation, and in sampling procedures. When using another's data, question them. "How could he know?" is a good question to ask. To enable the reader to appraise the usefulness of the data, the definitions used, the measurement problems encountered, and the sampling error in the results should be carefully indicated. Think critically about the numbers reported. They do not by their mere presence make a work "scientific."

Age Definition and Measurement Problem in the United States Census of Population. The 1950 United States Census of Population is an example of a large-scale survey in which individuals are questioned about certain characteristics. To illustrate both the need for careful definitions for the enumerators and possible errors in the data, the following passage is quoted from "United States Census of Population," U.S. Department of Commerce, Bureau of the Census, vol. II, "Characteristics of the Population," part 5, California, page xvi. (Note: This passage is on the same page of each Census part which covers a different state.)

AGE

Definitions

The age classification is based on the age of the person at his last birthday as of the date of his enumeration, that is, the age of the person in completed years. The enumerator was instructed to obtain the age of each person as of the date of his visit rather than as of April 1, 1950.

Assignment of Unknown Ages

When the age of a person was not reported, it was estimated on the basis of other available information, such as marital status, school attendance, employment status, age of other members of the family, and type of household. Age was estimated by this procedure in the 1950 Census for 0.19 per cent of the population of the United States. This method of assigning unknown ages on the basis of related information was used for the first time in the 1940 Census when estimates of age were made for 0.16 per cent of the population of the United States. In previous censuses, with the exception of 1880, persons of unknown age were shown in a separate category. The summary totals for "14 years and over" and "21 years and over" for earlier censuses presented in this volume include all persons of "unknown age" since there is evidence that most of the persons for whom age was not reported were in the age classes above these limits.

Errors in Age Statistics

A considerable body of evidence exists which indicates that age is misreported in several characteristic ways and that certain age groups are less completely enumerated than others in censuses. A comparison of age distributions from the 1950 Census with age distributions based on figures from the 1940 Census and brought up to date from official records of births, deaths, and migration, suggests that this generalization is also true for the 1950 Census. This comparison shows that, for the United States as a whole, there appears to be an underenumeration of children under 5 of approximately 4.8 per cent as compared with about 7.6 per cent in 1940. Males between the ages of 18 and 24 also appear to have been relatively underenumerated. Likewise, there appears to be a deficit of persons in the age range 55 to 64 years which, however, is more than offset by an excess over the number expected in the age group 65 years old and over.

In addition to errors in the statistics for broad age groups arising from underenumeration and the misstatement of age, there is a tendency to report age in multiples of 5. This tendency is apparent in statistics for single years of age in which the frequencies for single years ending in 0 and 5 are frequently greater than those for the two adjoining years. This type of misreporting presumably occurs in situations in which the respondent, in the absence of specific knowledge as to his exact age or the age of the person for whom he is reporting, gives an approximate figure. The returns also exaggerate the number of centenarians, particularly among nonwhite persons. In general, the degree of inaccuracy in reported ages is greater for adults than for children.

Uses of Sampling in Gathering U.S. Government Statistics. The following passage is reprinted from "Statistical Services of the U.S. Government," revised edition, pages 16 and 17, June, 1952.

The use of modern methods of sampling has greatly improved the accuracy and speed of many Federal surveys, and reduced their cost and the burden of response. Sampling is used, for example, to determine month-to-month changes in prices and employment, to forecast conditions of crops and yield and to estimate household budgets and savings. It is also used to determine the number of inhabitants in an area and their characteristics, or the number of farms in an area, or the number that have particular characteristics. Sampling is especially economical for measuring changes that have taken place in various areas and classes since the last complete census. The economies are greatest when the last complete census is not far out of date.

The accuracy of a scientifically planned sample can be calculated closely in advance, and geared to the requirements of precision or to the limitations of cost. This is one of the most important features of modern sampling, because a survey can now be planned in accordance with the precision needed in the uses that are to be made of the data, or within the range of the permissible costs, whichever is the determining factor.

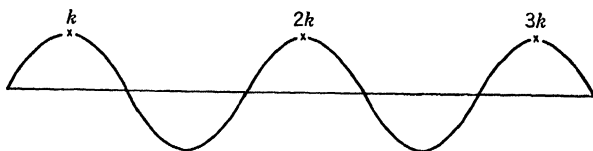
Almost any required degree of precision can be obtained through the use of sampling, provided the information sought is obtainable. A sample cannot produce information on subjects for which the techniques of the questionnaire and of the interviewing are yet so undeveloped that accurate information cannot be obtained from respondents. Even in these cases, however, sampling may make a contribution, for if a smaller number of respondents is covered more attention may be given to each at the same total cost.

As in the use of any precision tool, sampling requires expert knowledge. Sampling is not merely the substitution of a part for the whole, with the subsequent attempts at adjustments and corrections. Sampling is instead a methodical procedure, planned with the aid of the theory of probability, and carried out strictly in accordance with the demands of theory so that the errors arising from sampling can be calculated mathematically. In a sample that is properly designed, the precision of the results is properly assessed from the results themselves.

Additional Comments on Sampling Procedures in Surveys. The sampling theory developed in this course is predicated upon simple random sampling in which the observations comprising the sample are selected from a list by means of a table of random numbers. In practice, simple random samples are seldom used in surveys because lists are not available, because substitute methods are easier, or because other sample designs will give more reliable results for the same cost.

One substitute method has already been discussed—*systematic random sampling* (see Sec. 6.5, Sampling Procedure). Systematic random samples are selected by picking a *random start* and then taking every *k*th

observation, where k is the ratio of n/N . Systematic random sampling is quickly explained to and easily carried out by those relatively untrained in statistics. In some circumstances hidden difficulties may not readily come to light (e.g., in industrial sampling). If there is a systematic variation in the order of appearance or the order of numbering a series, the danger of selecting an unusual and nonrepresentative sample occurs when the sampling interval k happens to coincide with the period of the variation. An illustration is given below:



As an example, a bias may occur in a systematic sample of houses on a street. If every tenth house is to be selected and if there are 10 houses to the block, starting at a corner will give a "corner bias" to the sample. If the order of appearance or the order of numbering is unrelated to the characteristic under observation, the systematic sample is both simple and foolproof. It is usually a good substitute for a simple random sample in sampling alphabetical listings of human populations. Except for a recommendation that its use be confined to populations whose units may be numbered randomly, systematic sampling is beyond the scope of this book.

Other forms of sampling which make use of supplementary information are used in practice, especially in surveys. One of these methods is *stratification*. The population is divided into groups or strata which are more homogeneous than the whole population in respect to the characteristic being studied. For example, in a study of tract home prices, the strata might be defined by the builder; in a study of consumption expenditures, the strata might be defined by income levels. Strata may be defined by geography, measures of size (income, rent, sales, assets), location (rural or urban), etc. The stratifying variable is selected on the basis of available information and of high correlation with the variable under study. Once the strata have been determined, samples are selected *randomly* from within each stratum. For this purpose, a simple random sample or a systematic sample is used (with the cautions given earlier).

The number of observations taken in each stratum may be determined under several sets of conditions. Two of the most common are proportional and nonproportional allocations. In proportional allocation, the percentage of the sample taken from a particular stratum is the same as the percentage of the whole population in that stratum. If N is the size of the population, N_h the size of the h th stratum, n the sample size, and

n_h the number of observations in the sample to be selected in the h th stratum,

$$n_h = \frac{N_h}{N} \cdot n$$

The sample gives more weight to the larger strata relative to the whole population.

In nonproportional allocation, both the variability in a stratum and its relative importance in the population are taken into account. If σ_h is the standard deviation in the h th stratum,

$$n_h = \frac{N_h \sigma_h}{\sum_h N_h \sigma_h}$$

where the summation is over all h strata. The number of observations in a stratum depends upon both N_h and σ_h . The more homogeneous a stratum, the fewer observations are taken from it. The limit is a stratum for which $\sigma_h = 0$, where all observations are the same. Only one observation need be taken from this stratum to learn all that can be known. Selection of more than one observation, even though N_h/N may be large, is both unnecessary and costly. Under either method of allocation, a small stratum which is few in numbers but important in its contribution to the total magnitude of the variable (e.g., the 10 largest banks, or the 20 largest cities) is usually sampled 100 per cent. In other words, $n_h = N_h$.

Compared to simple random sampling, stratified sampling may lead to a smaller sampling variability with the same total sample size, or to the same sampling variability with a smaller sample. The effectiveness of the stratified sample in reducing the sampling variability depends upon the homogeneity of individual strata and the differences among the strata. The whole purpose in designing and executing a stratified random sample is to reduce the sampling variability over what it would be with the same number of observations in a simple random sample.

Another form of sampling is the *cluster sample*, in which the population is broken into clusters of elementary units. A sample of clusters is selected randomly from all the clusters. Within each cluster either all the sampling units are examined (100 per cent sampling) or a random sample of units is selected. Randomness is ensured at each stage of the sampling process. The main advantage of a cluster sample is its lower cost. However, since the items in a cluster are likely to be quite homogeneous with respect to the variable studied, a larger sampling error results. The conditions which make for effective stratified sampling lead to ineffective cluster sampling. The costs of obtaining the data in

cluster sampling may be so much reduced that the sample size can be increased and the sampling variability lowered for the same total cost as a simple random sample.

For example, in a study of a characteristic of the dwelling units in a large city, such as the number of inhabitants, a simple random sample would be almost impossible in the absence of a list of all dwelling units.

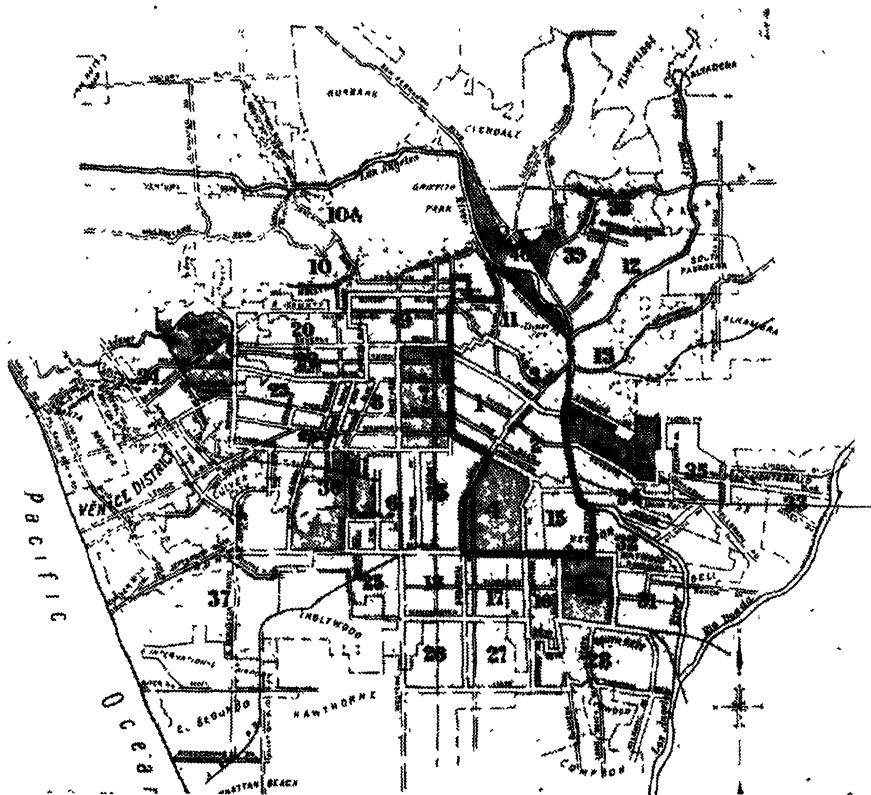


FIG. 6.1. General index to the volumes of Sanborn Maps for Los Angeles, California (4 refers to volume number). (This map and those following are reproduced with the permission of the Sanborn Map Co., New York.)

A cluster sample where the cluster is based upon a geographical area is most feasible. The city can be divided into city blocks (the cluster) and a sample of blocks selected from a map of the city. From the blocks selected at the first stage of sampling, either every dwelling unit or a sample of dwelling units may be examined. The most costly item is often transportation to the city block. Once the investigator has reached his destination, it takes little more time to enumerate the whole block than to take only a sample. If it is decided, however, to select a

sample of dwelling units, the map used must show all the buildings in the block identified by type of building. If the map does not do this, the investigator may sketch a small map of the block showing the dwelling units from which to select his sample. A list of dwelling units is necessary for random selection of the final sample. Such a listing for an entire city is almost impossible; a simple random sample consisting of a unit here and a unit there would involve the costly procedure of sending the investigator all over the city. The area sample reduces the required travel and, by providing a compact block as the primary sampling unit, makes possible a list of dwelling units for the final stage of sampling. Wherever maps are available, area samples are feasible. A whole country may be studied with a cluster of counties or other areas delineated by grid coordinates on the map.

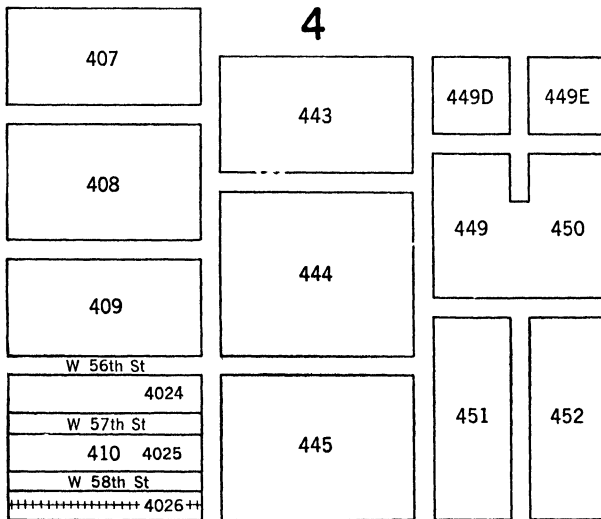


FIG. 6.2. A key to the pages of a part of Vol. 4 of Sanborn Maps for Los Angeles (For example, page 410 covers city blocks 4024, 4025, and 4026.)

Example 6.1. Figures 6.1 to 6.3 show a sequence of Sanborn Maps for the city of Los Angeles, California, which provide an enumeration of city blocks and structures on those blocks useful in selecting a random sample of blocks as clusters of dwelling units.¹ Figure 6.1 is a map of the whole city of Los Angeles with an index to the volumes which cover this territory. For example, more detailed maps covering the area designated by the numeral 4 are in Vol. 4 of the Sanborn Maps. Figure 6.2 is an index map to the pages of a part of Vol. 4, with each page containing the maps of several city blocks. Thus, page 410 of Vol. 4 maps the blocks numbered 4024, 4025, and 4026. Finally, Fig. 6.3 is a map of block

¹ These figures are reproduced with the permission of the Sanborn Map Company, New York.

4025, part of page 410 of Vol. 4, showing single-family dwelling units, apartments, and an elementary school.

Within the city, every block has a number and a random sample of city blocks may be selected by using a table of random numbers. Once a block has been selected for the sample, either all the dwelling units on the block can be covered or a subsample of dwellings units may be selected by numbering the dwelling units on the block map (Fig. 6.3) in a serpentine fashion and again using the table of random numbers. This sampling procedure contains two stages of sampling: blocks at the first stage and dwelling units on blocks at the second stage. It is a special case of cluster sampling called area sampling, because the basic clusters are defined by geographical areas.

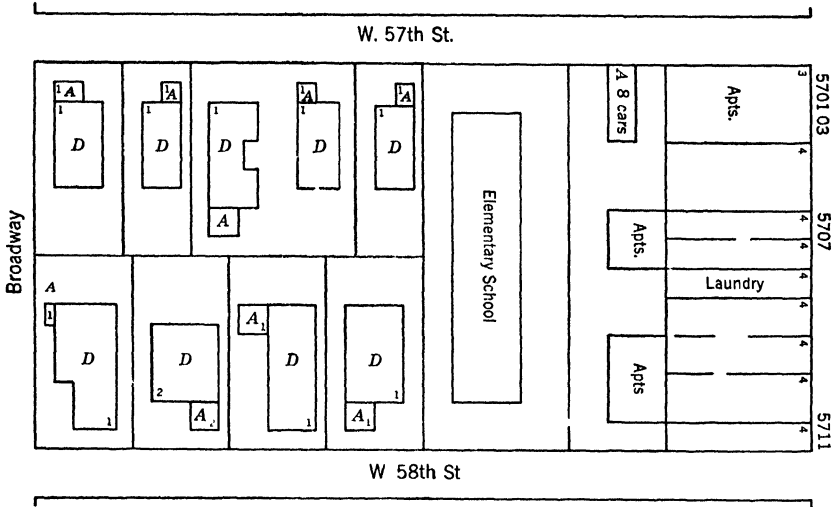


Fig. 6.3. Example of Sanborn Block Map showing apartments, elementary school, and single-family residences in block 4025, page 410, Vol. 4, of Los Angeles, California.

Each of the methods discussed—simple random, systematic, stratified, or cluster sampling—is a method of probability or random sampling. At each stage of the more complicated designs, a random sample is selected. In each instance the sampling variability may be estimated from the sample itself and probability statements made about the sample estimates of totals, averages, or percentages.

6.9 Nonrandom Samples

Information may also be obtained by nonrandom sample selections. The probability of selecting the units to comprise the sample is not known, and no measure of the reliability of the sample results in terms of the sampling error is available for these methods.

Expert opinion may be sought on a particular subject by questioning one or several individuals who claim to be experts in a given field. The

reliability of the results depends solely upon a criterion of faith in the validity of the opinions. An inherent danger of expert opinion is pointed out in the case of an individual who believes that he does not need an operation for a certain ailment. He goes to many doctors for a diagnosis, finally finds one who agrees with him, and adopts this doctor as the "expert."

A *purposive* sample is a sample selected as representative of the population, such as a presumably typical city, town, school, or family. The basic difficulty is that the sample selected may be representative of some characteristics but not necessarily of others. A typical family may contain a father, a mother, and two children, but may not be representative of incomes, of schooling, or of automobile ownership. The sample is known to match or to be representative on some characteristics; the *hope* is that it matches on others equally well.

A *quota* sample is much like a stratified random sample in that the population is broken down into more or less homogeneous strata and the sample size determined for each stratum. This is a quota, for a certain number of people must be found who meet the specifications of a given stratum. (If these individuals are selected randomly this sample is a perfectly good stratified random sample.) In practice, the quotas are usually filled by nonrandom methods. Convenience frequently dictates the selection, as the investigator chooses localities where he thinks he can fill his quota easily and quickly. Controls are built into the sampling process just as in stratified sampling, but probability sampling theory is not applicable because random methods of selection are not used.

Even surveys starting out with well-planned random samples may result in unknown probabilities of selection. Survey by mail questionnaire with a large core of nonrespondents is an example. Although the individuals to be questioned may originally have been selected by random methods, the voluntary returns do not constitute a random sample because of the nonrespondents. If those failing to answer are but a small part of the sample, allowance may be made for possible biases and nonmeasurable sampling errors. If, on the other hand, they are a substantial part of the sample, no methods are available to determine the reliability of the sample results.

Surveys very often use some form of nonrandom sampling coupled with well-sounding phrases of objective random sampling. Formulas for sample size and estimation procedures, and sometimes even formulas for measuring the sampling error, are given. Such surveys are misleading because in reality they are based on nonrandom samples; their reliability cannot be assessed. Nonrandom samples may suggest problems or hypotheses; they will never answer any with a known reliability of the answers.

EXERCISES

6.9. As a term project, you may undertake any research project which either estimates or tests a hypothesis about an unknown population proportion. You should set up the problem by defining the population you wish to study, what characteristic you will measure, how you will measure it, how large a sample you need, how you will select the sample, the statistical decision rule, and the final decision based upon the actual sample results.

Your project report should include a well-written summary of all phases of the research from the initial planning to the final conclusions. You should certainly discuss the method of measurement, possible sources of error, any irregularities which occur, and ways in which you would revise another study of the same sort. The method of selecting the sample is most important. Time spent in thinking about your problem before actually beginning physical labor is usually time well spent.

Almost all the practical problems in Chaps. 3 to 5 were taken from the research projects of students in elementary statistics. Reading over these problems may give you ideas about what you would like to do. The various sections of Chap. 6 and some of the references at the end of this chapter may help you in actually executing your research.

6.10. A newspaper recently carried the following item:

"When 18 boys were born consecutively in five days recently it was generally assumed that a run of girl babies would begin. But boys are still far in the lead. Since the run of 18 boys ended with the birth of a girl last Friday, this is how the new arrivals have been recorded: in the next six births, five were boys. Then, four boys in the next five births. From midnight Tuesday—five more boys and only one girl. Total score: 32 boys, 4 girls."

Assuming that $P = .5$ for a boy (actually the infant sex ratio favors boys slightly), the probability that 32 boys and 4 girls will be born is .0000009 or roughly 9 in 10 million. The probability that 32 or more boys will be born is .0000010. Do these data support or deny the hypothesis that $P = .5$ for boys? Give reasons for your answer.

6.11. A common practice in executing a survey is to ascertain the fraction of individuals that are in certain classes, e.g., 55 per cent male and 45 per cent female, and then to instruct that the sample be distributed in these same proportions. The interviewers are requested to find exactly the number of persons necessary to make the sample proportions equal the population proportions. What are some problems in this kind of survey? Is it a good or a bad method? Is this a method of probability sampling? How can the sampling error be estimated?

6.12. Railroads, restaurants, airlines, and theaters frequently leave blank questionnaires in convenient locations to solicit information from their patrons regarding their services. Criticize this method of obtaining information.

6.13. Senator *X* reports that people favor his recent proposal 50 to 1 on the basis of the first 2500 letters, telegrams, and telephone calls he has received. Comment on the conclusions that may be drawn from these data about public opinion in the whole country.

6.14. A list contains the names of all of the people in a given area who are registered to vote, with their party affiliation marked R for Republican and D for Democratic. The local party (Republican or Democrat) wishes to study certain characteristics of its registered voters from a tabulation of 10 per cent of the people in the list. The problem is to draw a 10 per cent systematic sample of people with the required registration. Which of the methods listed below will give the required sample?

1. From every successive 10 sheets of names, draw one at random and take all the voters on these sheets with the proper registration.
2. With a random start, take every tenth name.
3. With a random start take every tenth voter with the proper registration.
4. From every successive 10 names, draw one at random.
5. From every successive 10 names with the proper registration, draw one at random.
6. With a random start, take all the names with the proper registration on every tenth sheet.
 - a. Which of the procedures will be the most expensive and which the least expensive?
 - b. In practice, methods 3 and 5 usually give the smallest sampling variability. Are these also the most expensive? Can alert workers use these methods without making separate lists of voters with the required registration?

6.15. A manufacturer of baby foods wishes to study the incidence of pregnancy in a given state in which he does not have a retail outlet at the present time. From the latest Census report, he has a rough idea of the population of the whole state and of smaller geographical divisions within the state. Suggest a method of stratifying the total population for effective use of the funds available for the study.

SUPPLEMENTARY READING

General Methods

- Cochran, William G.: "Sampling Techniques," John Wiley & Sons, Inc., New York, 1953. Introduction
- Deming, William Edwards: "Some Theory of Sampling," John Wiley & Sons, Inc., New York, 1950. Chapters 1 and 2 discuss the planning of a survey and various errors which can occur in surveys.
- Goode, William J., and Paul K. Hatt: "Methods in Social Research," McGraw-Hill Book Company, Inc., New York, 1952.
- Hansen, Morris H., William N. Hurwitz, and William G. Madow. "Sample Survey Methods and Theory, vol. 1, Methods and Applications," John Wiley & Sons, Inc., New York, 1953. This is an advanced book in sample survey methods. The students in elementary statistics might read Chaps. 1 to 4. Chapter 12 presents some case studies of actual sample surveys of which parts B and E on the population survey and the 1950 Census are fairly elementary.
- Parten, Mildred: "Surveys, Polls, and Samples: Practical Procedures," Harper & Brothers, New York, 1950.
- Payne, S. L.: "The Art of Asking Questions," Princeton University Press, Princeton, N.J., 1946.
- Wilson, E. Bright, Jr.: "An Introduction to Scientific Research," McGraw-Hill Book

Company, Inc., New York, 1952. Chapters 1, 2, 3, 6, 7, and 13 are good on general methods of research.

Yates, Frank: "Sampling Methods for Censuses and Surveys," Hafner Publishing Company, New York, 1949. Chapters 2, 4, and 5 discuss the requirements of a good sample, and problems arising in the planning, execution, and analysis of a survey.

Applications

"Business Application of Statistical Sampling Methods: Proceedings," University of Illinois Press, Urbana, Ill., May, 1950.

Business Statistics Going Ph.D., *Business Week*, May 30, 1953, p. 96

Marks, Eli S., W. Parker Mauldin, and Harold Nisselson: A Case History in Survey Design: The Post-enumeration Survey of the 1950 Census, *Journal of the American Statistical Association*, vol. 48, no. 262, pp. 221-243, June, 1953.

"Modern Statistical Methods for Business and Industry: Proceedings," Carnegie Institute of Technology, May, 1953

Neter, John: Accounting Data for Control Purposes, *The Controller*, August, 1953, pp. 369-371.

Top Management by Mathematics, *Business Week*, May 30, 1953, p. 86.

Criticism

Cochran, William G., Frederick Mosteller, and John W. Tukey: Statistical Problems of the Kinsey Report, *Journal of the American Statistical Association*, vol. 48, no. 264, pp. 673-716, December, 1953.

——— -: Principles of Sampling, *Journal of the American Statistical Association*, vol. 49, no. 265, pp. 13-35, March, 1954.

Wallis, W. Allen: Statistics of the Kinsey Report, *Journal of the American Statistical Association*, vol. 44, no. 248, pp. 463-484, December, 1949.

CHAPTER 7

TOOLS—FREQUENCY DISTRIBUTIONS, AVERAGES, AND MEASURES OF DISPERSION

The first three chapters in this book dealt with populations characterized by the classification of elementary units into one of two categories. Statements about these populations referred to the proportion of elementary units in each category. In some populations, the numerical values assigned to the elementary units fall into more than just two categories; they vary quantitatively and form a distribution of values. The bowl of 998 numbers in Table 1.3, used in the random sampling experiment, is an example of such a population. A distribution of values has several important descriptive criteria: (1) the form or shape; (2) the central value or average; and (3) the dispersion about the central value.

Certain characteristics of the distribution are readily discernible when the whole population is available for study, as in Table 1.3. More numbers lie near the center of the distribution than at the extremes; i.e., the values tend to “pile up” around the middle. The “average value” is at the center of the distribution, and the individual values vary about the average. Moreover, the distribution is symmetrical about the average value. When the whole population is not at hand, information about these characteristics may be obtained from samples. The sampling procedure poses two general statistical problems: (1) describing the same characteristics in the sample; and (2) drawing inferences about the population characteristics from sample values. Inference problems will be taken up in the next chapters; the descriptive problems are illustrated in this chapter using the weight, height, and age data of Table 7.1.

7.1 The Array: Weight Data

What do the raw data of Table 7.1 show about students' weights? Is the distribution of weights readily apparent? It requires some study to discover even the simple fact that the smallest weight is 120 pounds and the largest 195 pounds; more observations would make this even more difficult. In order to make such facts easily discernible, the raw data

ELEMENTARY STATISTICS

TABLE 7.1. HEIGHTS, WEIGHTS, AND AGES OF STUDENTS PRESENT IN
STATISTICS CLASS ON SEPT. 24, 1951

Height (to nearest half inch)	Weight (to nearest pound)	Age (to completed year)
70 0	145	20
64 5	129	19
64 0	132	21
70 0	185	20
71.0	185	22
70 0	142	23
66.0	124	19
69.5	156	21
60 0	148	24
72 5	182	21
76 0	195	21
74 5	160	20
72 0	155	23
72 5	153	20
72 5	155	20
69 5	139	22
65 0	120	19
68 5	147	19
69 0	185	32
72 0	165	21
71 0	150	21
68 5	150	21
66 0	158	29
70 5	135	21
72 0	140	20
60 0	157	20
67 0	188	31
72 5	166	20
69 5	142	21
69 5	180	27
67.5	140	21
69.5	162	20
68 5	155	21
64 5	152	20
73.5	191	21
69.0	155	24
69 0	165	45
70.5	170	27
71 5	165	21
71 0	165	21

must be reorganized. An array provides a simple way to begin the analysis.

An array is an arrangement of the observations from smallest to largest or from largest to smallest. A quick method of making an array (of weights) is to set up a table with column headings corresponding to 10-pound intervals covering the whole range of weights, thus:

120	130	140	150	160	170	180	190
9	2	5	6	0	0	5	5
4	9	2	5	5		5	1
0	5	8	3	6		2	
		7	5	2		5	
		0	0	5		8	
		2	0	5		0	
		0	8	5			
			7				
			5				
			2				
			5				

The last digit in each weight in the list of 40 weights in Table 7.1 is recorded in the appropriate column. For example, the first weight is 145, and the numeral 5 is entered under the column headed 140. The next weight is 129 and the numeral 9 is entered under the column headed 120, etc. The results are shown above. The array is made by ordering the weights within each column.

120	139	145	152	155	162	166	185
124	140	147	153	156	165	170	185
129	140	148	155	157	165	180	188
132	142	150	155	158	165	182	191
135	142	150	155	160	165	185	195

The array, though simply a rearrangement of the raw data (the 40 numbers), reveals several things. The range of 120 to 195 pounds is immediately obvious. Moreover, the average weight appears to be about 155 pounds. The array clearly shows that more weights fall at the central values of 140 to 160 pounds than at the extreme values of 120 and 190 pounds. Little more than this can be learned from an array.

7.2 The Frequency Distribution: Weight Data

A more compact summary and a clearer picture of the underlying distribution result from grouping together those observations which for practical purposes need not be distinguished. For example, all weights from 160 to 170 may be grouped together. The actual individual

weights lose their identity in this interval, for the statistician knows only that seven of them lie between 160 and 170 pounds. Such a summary is called a *frequency distribution* and the groupings, such as 160-170, are known as *class intervals*.

The construction of a frequency distribution is simple in theory and also, with experience, in practice. The chief difficulty for beginners is determining where to start. Of various possible methods, the one here suggested will lead rather quickly to a frequency distribution.

The 10-pound interval used for the array is an obvious starting point. The intervals will then be: 120-130, 130-140, 140-150, Is this form of interval good? In what interval is an observation of 140 pounds placed? Obviously, it could be placed in either the interval 130-140 or the interval 140-150. Overlapping intervals like this are ambiguous and must be avoided. The first rule in making a frequency distribution is to keep the intervals *mutually exclusive* with respect to the measured values of the variable.

A possible variation of the 10-pound interval is the following: 119.5-129.5, 129.5-139.5, 139.5-149.5, These intervals are not mutually exclusive as written, for the upper limit of one is identical with the lower limit of the next. But *they are mutually exclusive in this example because no weights are reported to the half pound*. Each original weight is given in full pounds. In this respect, then, these intervals are satisfactory, and they should be examined more critically.

Each original observation loses its identity when placed in a class interval of a frequency distribution. Later it will be necessary to perform some computations with these numbers and, therefore, a numerical value must be assigned to the observations in any class interval. The value assigned is the *mid-point* of the interval. It may be computed by adding half the length of the interval to the lower class limit. The mid-points of the above intervals are:

<i>Interval</i>	<i>Mid-point</i>
119 5-129 5	124.5
129 5-139 5	134 5
139 5-149 5	144 5, . . .

It can easily be seen that these mid-points are hardly appropriate here, for they are reported in half pounds while the original data are reported in whole pounds. Moreover, since the original weights show a tendency to end in the whole numbers 0 or 5, intervals with mid-points ending in either 0 or 5 in the digits position are better for the weight data than intervals with other mid-points. A second rule in making a frequency distribution is to select intervals with mid-points at clustering values of the original data if such clusterings exist.

Both 5- and 15-pound intervals may have mid-points ending in 0 and 5 in the digits position. For example, 120 is the mid-point of the interval 117.5–122.5, and 125 of the interval 117.5–132.5. With data ranging from 120 to 195 pounds, selection of a 5-pound interval will result in a distribution with $7\frac{5}{5}$, or 15 intervals. Use of a 15-pound interval will require $7\frac{5}{15}$, or 5 intervals. Which is the better choice? The problem of choosing the number of class intervals is related to the problem of selecting the size of the interval. The smaller the class interval, the more intervals there will be, and vice versa. The final decision is essentially arbitrary. Realization of this element of subjectivity enables a student to analyze and interpret more clearly the frequency distributions made by others. He will understand that the data could have been arranged differently.

There should be neither too many nor too few intervals. A 1-pound interval for weights would result in 76 intervals for a range of 120 to 195. With a sample of only 40, many of these intervals would contain no observations. The distribution of weights of a large number of people is known from previous experience to be hump-shaped, and a 1-pound interval will not accurately reflect this distribution of the population. The purpose of a frequency distribution is, of course, to create a visual picture of the underlying distribution.

The other extreme is to have one large interval: 120–195 pounds. All 40 observations will be in this interval. This is ridiculous, for such a distribution tells even less than a 1-pound interval. The interval chosen must lie between these two extremes. Its size will depend upon (1) the number of observations in the sample, (2) the range of the data, and (3) the use for which the distribution is intended. A rough rule of thumb is to form a distribution with 6 to 16 intervals. It is clear that, with only 40 observations, 6 intervals are preferable to 16 intervals. Thus the 5-pound interval for weights is ruled out, and the 15-pound interval selected.

Since the smallest weight is 120 pounds, the first 15-pound interval may begin at 117.5. The resulting frequency distribution is shown in Table 7.2.

TABLE 7.2 FREQUENCY DISTRIBUTION OF WEIGHTS

Class bounds	Mid-point	Frequency
117.5–132.5	125	4
132.5–147.5	140	8
147.5–162.5	155	14
162.5–177.5	170	6
177.5–192.5	185	7
192.5–207.5	200	1

One of the mid-points is 155 pounds, which is a cluster value; the next one, 170, misses the cluster at 165. However, a 15-pound interval cannot catch both 155 and 165 as mid-points, and a 10-pound interval, which could catch both of them, is poor for other reasons. Of the four observations in the first class, two are below 125 and two above; the only observation in the last class is 195, which is below 200. Seven observations are in the interval 177.5–192.5 and only six in the preceding interval 162.5–177.5. These irregularities may be due to the small sample size.

The arbitrary element in the selection of the interval suggests the construction of another distribution with a 15-pound interval starting at 112.5, as in Table 7.3.

TABLE 7.3. FREQUENCY DISTRIBUTION OF WEIGHTS

Class bounds	Mid-point	Frequency
112 5–127 5	120	2
127 5–142 5	135	8
142 5–157 5	150	13
157 5–172 5	165	9
172 5–187 5	180	5
187 5–202 5	195	3

This distribution has mid-points which end in 0 or 5. It avoids the irregularity of a larger frequency in the higher of the two intervals above the mode. However, the two observations in the first interval are either equal to or greater than the mid-point (120), and the three observations in the last interval are either equal to or less than the mid-point (195). Both of these distributions are equally satisfactory.

Certain rules can be followed in making a frequency distribution, but even so a good deal of choice remains with respect to other elements of the distribution. Such choices involve decisions that are not particularly important; the decision is equally satisfactory in either way. The rules may be summarized as follows:

1. Make an array of the data.
2. Adopt class intervals which are mutually exclusive.
3. If the original data are measured to the *nearest unit*, make the class limits end in *one-half* the unit of measure to ensure unambiguous class intervals.
4. Adopt class intervals with mid-points at clustering values of the original data, if such clusterings occur.
5. Follow the rule of thumb that the right number of intervals is 6 to 16. The final decision depends upon the number of observations and the range of the data.

7.3 Additional Comments on Making Frequency Distributions

Despite all arguments to the contrary, statistical reports often make use of intervals like the following for variables measured to the nearest unit: 120 and under 130, 130 and under 140, 140 and under 150, These intervals are inherently ambiguous, for example, for weights reported to the nearest pound. A person weighing 139.7 pounds will report his weight as 140 pounds. This weight will be placed in the interval of 140 and under 150, when it should really be placed in the interval of 130 and under 140. Although errors of this kind will occur only at the class limits, they may be avoided by using intervals with class limits set at one-half the unit of measure in the original data.

Intervals of this form are useful where data are not measured to the nearest unit; for example, the ages of Table 7.1 are reported to the completed year. In this case intervals such as 18 and under 20, 20 and under 22, etc., are preferred. If the interval form 17.5–19.5, 19.5–21.5, etc., is adopted, the ambiguity will lead to errors. A person of 19 years and 8 months will report his age as 19. This age will be placed in the interval 17.5–19.5; it should be placed in the interval 19.5–21.5. These marginal cases can be handled correctly by selecting intervals of the form 18 and under 20, etc.

Example 7.1. Table 7.4 shows the age distribution for the Los Angeles Metropolitan Area as reported by the 1950 Census of Population.

More on Mid-points. Since individual observations lose their identity in the frequency distribution, it is necessary to assign a particular value to the observations in each class interval. This value is the mid-point of the interval, computed by adding one-half the difference between the class bounds to the lower class bound.

<i>Class bounds</i>	<i>Mid-point</i>
199.5 –299.5	249.5
199.95 –299.95	249.95
199.995–299.995	249.995
10 and under 20	15

When the limits are written as 200–299, it is usually understood that the class bounds are 199.5–299.5 and the mid-point is 249.5.

Mid-points may have to be rounded. The precision of the original data and the magnitude of the observations must be considered in such rounding, though the rules are not uniform. In general, do not round 249.5, but round 249.95 to 250 and 249.995 to 250. Such rounding will not introduce serious error. The observations in any class interval will

take on the value of the mid-point for further computation. Although inaccuracies will result, the errors will tend to cancel out if the distribution is symmetrical because positive differences between original observations and the mid-point will compensate for negative differences.

TABLE 7.4. AGE DISTRIBUTION FOR LOS ANGELES STANDARD METROPOLITAN AREA: 1950

<i>Age group</i>	<i>Number</i>
Under 5	425,470
5-9	332,689
10-14	250,301
15-19	232,551
20-24	297,077
25-29	381,306
30-34	376,731
35-39	368,397
40-44	326,704
45-49	291,457
50-54	264,749
55-59	226,658
60-64	195,382
65-69	160,489
70-74	111,506
75-84	106,807
85 and over	19,637
Total	4,367,911

SOURCE: U.S. Bureau of the Census, "Census of Population: 1950," vol. II, "Characteristics of the Population," part 5, California, Government Printing Office, Washington, 1952, Table 33.

Age is reported to the completed year (see Sec. 6.8 for a discussion of age measurement in the Census); yet the intervals are of the form 5 to 9, 10 to 14, etc. If the interval 5 to 9 means 5 and under 10, and if the interval 10 to 14 means 10 and under 15, this distribution is correct. If they mean 4.5 to 9.5 and 9.5 to 14.5, they are ambiguous.

Extreme Values. Extreme values often present a problem in the construction of frequency distributions. Suppose that one of the 40 weights recorded in the class is 350 pounds, and that upon investigation this is found to be a correct value. The class intervals including the first 39 observations of the array might run as follows:

<i>Class bounds</i>	<i>Frequency</i>
112 5-127 5	4
127.5-142 5	6
.	.
.	.
.	.
172 5-187 5	3
187.5-202.5	1

If class intervals are added so as to include the 350-pound value, all but one of them will be empty.

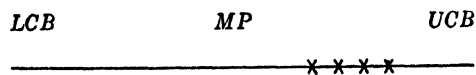
202	5-217	5	0
217	5-232	5	0
	.		.
	.		.
337	5-352	5	1

This is very untidy. A better solution is to make one *open-end interval*, as follows:

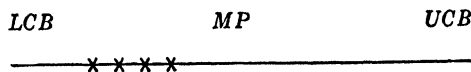
202.5 and over* 1

This will include the one extreme observation, whose actual value should be given in a footnote: *350 pounds. If there are several observations in an open-end interval, record the value of each of them in the footnote. The same method may be used for extremely low values.

Starting the Class Intervals. Class bounds need not start with the smallest sample observation or end with the largest (120 and 195 in the weight data). The lower class bound of the first interval may be less than 120 (it cannot be more), and the upper class bound of the last interval may be greater than the largest value (it cannot be less). Only in rare circumstances will the extreme values of a series of class intervals coincide with the range of the data. Even then the coincidence will not be exact because class bounds are given to one more decimal place than the original observations. The class bounds should neither be too far above nor too far below the extreme sample values. The figure below illustrates what may happen in the first interval if this rule is disregarded:



The four observations in this interval lie above the mid-point, which thus understates each observation. If the last class interval looks like this:



the error of understatement will be canceled by the error of overstatement. This may happen in an approximately symmetrical distribution, but it is preferable to have observations both above and below the mid-points of extreme classes:



Then errors of overstatement and understatement will tend to cancel within each interval as well as among intervals.

7.4 Frequency Distributions of Age and Height

The class interval for age must be of the form *x and under y* (for example, 20 and under 25), because age is measured to the *last* year. The range is from 19 to 45. If the whole range is divided into, say, 6 intervals, each interval will be $(45 - 19)/6 = 4$, or 5 years long. With a 5-year interval, the distribution is as follows:

<i>Class interval</i>	<i>Frequency</i>
19 and under 24	32
24 and under 29	4
29 and under 34	3
34 and under 39	0
39 and under 44	0
44 and under 49	1

This distribution shows that 32 of the 40 class members are in the youngest age group. In fact, only six of the original observations are above 24 (27, 27, 29, 31, 32, 45). A frequency distribution that will distinguish among the observations from 19 to 24 and treat the six extremes in another manner is more desirable.

With a two-year interval, the distribution is:

<i>Class interval</i>	<i>Frequency</i>
19 and under 21	14
21 and under 23	16
23 and under 25	4
25 and over*	6

* 27, 27, 29, 31, 32, 45 years.

This distribution is not completely satisfactory because the concentration has been divided between only two groups. To distinguish among the large number of observations in the years from 19 to 23, a 1-year interval must be used, as in Table 7.5.

TABLE 7.5. FREQUENCY DISTRIBUTION OF AGE

<i>Class interval</i>	<i>Frequency</i>
19 and under 20	4
20 and under 21	10
21 and under 22	14
22 and under 23	2
23 and under 24	2
24 and under 25	2
25 and over*	6

* 27, 27, 29, 31, 32, 45.

The 6 observations greater than 24 are placed in an open-end interval with their exact values footnoted.

In the first discussion of frequency distributions, a general rule was enunciated: class bounds should end in one-half the unit of measure when data are measured to the nearest unit. This rule is applied with little difficulty to weights reported to the nearest whole pound. Heights are recorded to the nearest $\frac{1}{2}$ inch. Therefore, the class bounds should end in either .25 or .75 inches. Since the original observations end in either .0 or .5, the mid-points of the class intervals should end in .0 or .5; 1-inch intervals (or multiples of 1 inch) will not give such mid-points.

Example:

<i>Class interval</i>	<i>Mid-point</i>
67 25-68 25	67 75
67.75-69.75	68.75

An odd interval must be used: .5 inch, 1.5 inches, or 2.5 inches, etc.

The range of heights is from 60.0 to 76.0, or 16 inches. A half-inch interval is obviously too small. A $1\frac{1}{2}$ -inch interval will necessitate about 11 or 12 intervals. However, there are four extreme observations: 60.0, 60.0, 74.5, and 76.0. The range of the remaining observations is from 64.0 to 73.5, or 9.5 inches. A $1\frac{1}{2}$ -inch interval for this range will make a distribution with 8 or 9 intervals, if the extreme values are put into open-end intervals. A frequency distribution made by following this procedure is shown in Table 7.6.

TABLE 7.6. FREQUENCY DISTRIBUTION OF HEIGHTS

<i>Class bounds</i>	<i>Frequency</i>
Under 63.75	2*
63 75-65 25	4
65 25-66 75	2
66 75-68 25	2
68 25-69 75	11
69 75-71 25	8
71 25-72 75	8
72 75-74 25	1
Over 74 25	2*

* 60.0, 60.0, 74 5, 76.0.

7.5 Summary of Frequency Distributions of Weight, Age, and Height

One of the purposes in making a frequency distribution of sample values is to obtain information about the distribution of the population from which the sample is selected. Although the actual distribution of the sample values varies somewhat because of the arbitrary choice of the class interval, general tendencies may be noted. Each of the three frequency distributions constructed in this chapter shows a tendency for

the values to cluster about a central value. One interval has more observations in it than any other interval, and there is an obvious dispersion of the observations about this central value. None of these distributions is symmetrical about the center; usually, values string out farther from the center in one direction than in the other. Such asymmetrical distributions are said to be *skewed*. When the long tail of the distribution is above the center, the distribution is skewed to the right or positively; when the long tail is below the center, the distribution is skewed to the left or negatively.

For example, the ages of students tend to cluster around values from 20 to 22 but with a positive skewness (i.e., several larger values), as shown in Table 7.5. Heights tend to cluster around the middle values of 68 to 70, although there is a fair amount of irregularity in the distribution. In fact, the skewness is hard to detect but is probably to the left because of the larger concentration of values just above the center. The concentration of weight is around the 150-pound level with a positive skewness.

These tendencies are valid descriptions of the sample frequency distributions. Is it valid to draw the inference that they are also characteristic of the underlying populations? The tendency to pile up around a central value and the dispersion around that central value probably are valid inferences for the population; skewness is questionable, because small samples from perfectly symmetrical distributions usually are skewed. In other words, these characteristics of samples are subject to sampling variations. They will differ from sample to sample and from the true population characteristics. Inferences about the population distributions are subject to the same uncertainties as were inferences about percentages. The difference between a description of the sample distribution and an inference about the population distribution is so important that one chapter is devoted to each.

7.6 Graphing Frequency Distributions

A better visual impression of a frequency distribution may be obtained from a graph. Moreover, statistical data presented by means of graphs and charts will reach people who would not read a table of figures. Such graphs and charts can be quite dramatic in their appeal if drawn by astute draftsmen using colors and pleasingly balanced lines or bars.

Equal Class Intervals. Frequency distributions may be graphed as *histograms* by erecting bars over the *class bounds* and selecting the heights of the bars so that the areas of the bars are proportional to the frequencies in the class intervals. The scale of the variable may be labeled at the class bounds or the mid-points, but in either case the bars are erected *at the bounds*. There are no gaps along the scale of the variable. With

equal class intervals, the bases of the bars will be equal. Since area equals base times height, the height of each bar as well as its area will be proportional to the frequency in the class interval it represents.

Frequency distributions are graphed in the form of *frequency polygons* by plotting frequency on the ordinate (vertical scale) against the mid-point of the class interval on the abscissa (horizontal scale). Histograms

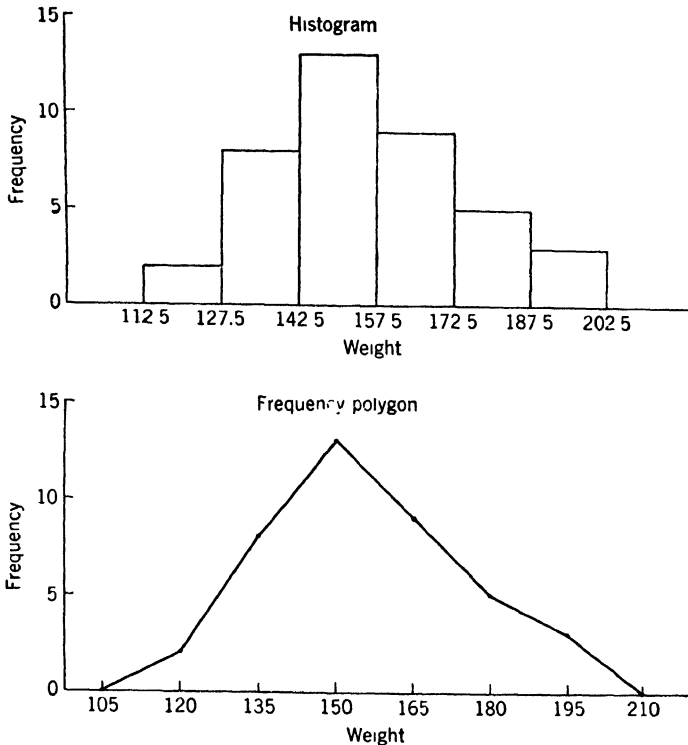


FIG. 7.1. Histogram and frequency polygon of weights of Table 7.3.

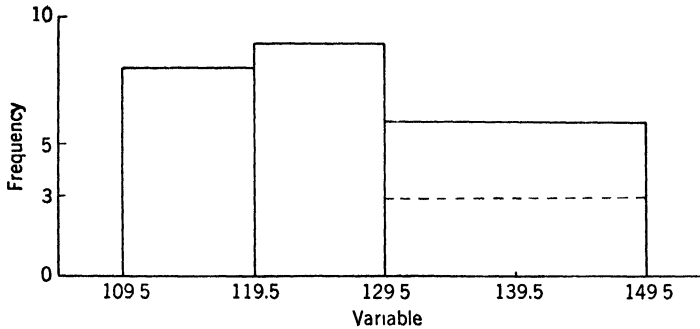
and frequency polygons were drawn in Sec. 2.6 for the number of red beads in samples from the bead box. The data of Table 7.3 are shown in these forms in Fig. 7.1.

Unequal Class Intervals. An example of a frequency distribution with unequal class intervals is the following:

Class bounds	Frequency
109.5-119.5	8
119.5-129.5	9
129.5-149.5	6

Although equal class intervals are generally preferred, unequal intervals

sometimes cannot be avoided. They are often graphed incorrectly, as shown below:

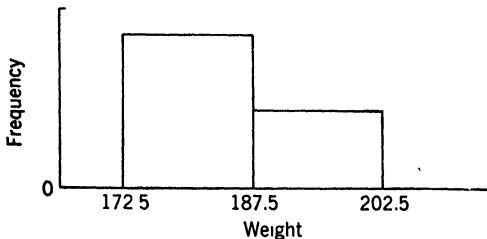


This method is incorrect because the *area* of the interval 129.5–149.5 represents 12 frequencies, and not 6. The area may be made proportional to 6 frequencies by averaging the number of observations in the two intervals (6) over the number of intervals (2). (The interval 129.5–149.5 is twice as long as the other equal intervals.) The average is 3 observations per interval and is shown correctly graphed by the dotted line.

Open-end Intervals. A distribution with an open-end interval,

$$\frac{202.5 \text{ and over}^*}{350 \text{ pounds}} \quad 1$$

may be graphed as follows:



*One observation is 350 pounds.

Scaling Histograms and Frequency Polygons. The visual impression given by a histogram or a frequency polygon may be radically altered by a change in the scales used to plot the frequencies and the values of the variables. It is imperative to start the ordinate (vertical scale) at zero and to avoid the temptation in graphing very large frequencies to break the vertical scale and start at a value higher than zero in order to enlarge the differences between successive bars or points. This practice leads to a distortion illustrated by Fig. 7.2a and b. When the scale is broken,

the second bar appears to be twice the length of the first bar, and the third bar three times the length of the first bar. This is misleading! Some statisticians will defend this practice if the scale break is clearly labeled, but even a label will not remedy the distortion for those who make comparisons by measuring the lengths of the bars from the abscissa. The human eye does not make allowance for the scale break. Anyone reading the frequencies from the ordinate scale will, of course, get the correct impression, but graphs are drawn to catch the eye and to reach people who do not carefully read tables or scales: they want a picture at

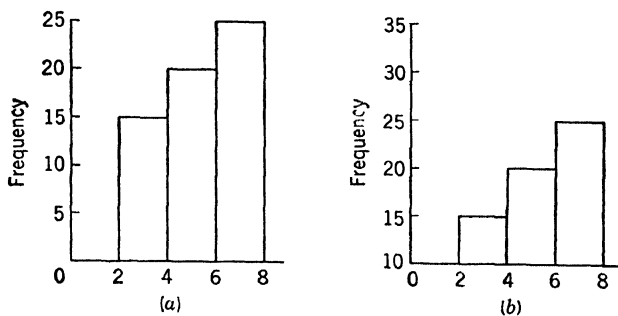


FIG. 7.2. Example of a distortion by a vertical scale break: (a) Full scale. (b) Broken scale.

one glance, but the picture from a graph with a broken scale is incorrect. There is a certain arbitrariness in the choice of scales on which to plot frequency distributions and polygons, just as there was in the choice of a class interval. Obvious distortions due to broken scales can and should be avoided in good statistical work.

EXERCISES

7.1. Using the data on height, weight, and age gathered in Exercise 2.24, make (a) an array, (b) a frequency distribution, (c) a histogram, and (d) a frequency polygon.

7.2. Write a brief summary of these data, commenting on the shapes of the distribution, skewness, extreme values, etc.

7.7 Measures of Central Position and Dispersion: Raw Data

Although tabular and graphic methods of summarizing the information in a sample of observations of a variable indicate some of the numerical characteristics of the distribution, for the most part they give only a visual picture. Quantitative measures of the “average” value and the dispersion are needed for an adequate description of the whole distribution.

Averages and measures of dispersion are studied together. An aver-

age is simply a way of describing a group of figures by a single number which is easy to grasp and is in some sense representative of the whole group. The very nature of an average implies a variability or a dispersion of individual values about the average. In fact, sets of numbers which are interesting to study usually have several values, and it is often because of the variation that they are studied in the first place. The variation about the average is just as important as the average itself. The emphasis in statistics is changing in accordance with this idea; in the beginning statistical methods were concerned primarily with the study of averages, but more recently they have been concerned with the study of variation or dispersion.

The notion of variability or variation or dispersion should be a familiar one. The variation in the weather is, in many parts of the country, a constant source of idle chitchat. Lovers of baseball are often concerned with the variations in their favorite batter's hitting -his long trips to the plate without a hit, etc. There are important areas where variation affects the decisions of everyday life. For example, the choice of an occupation or lifework may be affected by variation in income. Some prefer the security of a civil service job with a steady salary and automatic raises in pay. Others prefer a more variable income, with riches one year and the risk of near poverty the next. The average income for 10 years might be the same in both cases; the variation is very different. Two occupations may pay an average of \$5,000 per year for 10 years. One job will pay \$5,000 each year; the other will pay as high as \$10,000 in one year, but as low as \$2,500 in others. The total income (barring the progressive income tax, etc.) is the same in each case; the variability is an important consideration in making a real-life decision between the two occupations.

This chapter will emphasize the description of the sample by several measures of "average" and several measures of dispersion, present efficient computing methods, and interpret the results. A later chapter will take up the problem of drawing conclusions about population characteristics from a random sample.

The Median. The median is a positional average depending upon the order of the values of the variable. It is the middle of the distribution in the sense that it is the value which divides the total frequency into two *approximately equal* parts. In other words, roughly half the observations have values greater than the median and half have values less than the median. The median is symbolized by $\bar{x}$.

Although difficult to locate in the raw data, the median may easily be computed from the array. When the data are discontinuous, as in the array, the median is somewhat indeterminate, but it is conventionally defined as follows:

1. If there are an even number of observations in the array, the median is halfway between the $(n/2)$ th and $(n/2 + 1)$ st observation;

2. If there are an odd number of observations, the median is the $[(n - 1)/2 + 1]$ st observation.

The median is not affected by the size of the extreme observations. The largest observation could be increased many times and the median would remain the same.

Example 7.2. Given an array of 5 numbers: 1, 3, 6, 7, 10. The middle value is 6 and the median is $\bar{x} = 6$. With only a few observations, the median is the middle value although only $\frac{2}{5}$ of the values lie above the median and $\frac{2}{5}$ below the median.

Example 7.3. Given an array of 6 numbers: 1, 3, 6, 7, 10, 11. By convention, the median is the value midway between 6 and 7, or $\bar{x} = 6.5$. The median could be assigned any other value between 6 and 7; it is merely convention to split the difference in half.

Example 7.4. The median of the array of 40 weights (see Sec. 7.1) is computed by the first convention. $n/2 = 20$. The value of the 20th observation is 155. The value of the $(n/2 + 1) = 21$ st observation is also 155. $\bar{x} = 155$. In the same manner, the median age is determined as 21 years and the median height as 69.5 inches.

Example 7.5. Given an array of 5 numbers: 1, 3, 6, 7, 10, with median 6. If the largest observation were changed to a value of 200, making the array: 1, 3, 6, 7, 200, the median would still be 6. This property often renders the median a useful "average" when extreme observations make other measures atypical.

Quartiles. The *quartiles* of a distribution are measures very much like the median. Whereas the median divides the total frequency into approximately equal parts with one half of the frequencies above and one half below the median, the quartiles quarter the total frequency. So:

Q_1 , the first quartile, divides the distribution so that approximately $\frac{1}{4}$ of the observations lie below Q_1 and $\frac{3}{4}$ lie above Q_1 .

Q_2 , the second quartile, is equal to the median and divides the distribution so that $\frac{2}{4}$ of the observations lie below and $\frac{2}{4}$ above Q_2 .

Q_3 , the third quartile, divides the distribution so that approximately $\frac{3}{4}$ of the observations lie below Q_3 and $\frac{1}{4}$ lie above Q_3 .

Example: Given an array of numbers:

$$\begin{array}{ccccccc} 1, & 3, & 4, & 5, & 6, & 8, & 9, & 10, & 13, & 16, & 17, & 20 \\ & & Q_1 & & Q_2 & & Q_3 & & & & & \\ Q_1 = & 4.5 & & Q_2 = & 8.5 & & Q_3 = & 14.5 & & & & \end{array}$$

Percentiles. Medians and quartiles are special kinds of percentiles, which are the most general positional measure. The p th percentile divides the distribution so that approximately p per cent of the observations are below it and $1 - p$ per cent are above it. Thus, the 90th per-

centile ($X_{.90}$) divides the distribution so that approximately 90 per cent of the observations lie below and 10 per cent above it. The 80th percentile divides the distribution at the 80-20 percentage point. Q_2 and the median are the 50th percentile; Q_1 is the 25th percentile; and Q_3 is the 75th percentile.

Example: From the array of weights in Sec. 7.1, the following values are calculated:

$$\begin{aligned}Q_1 &= 143.5 = X_{.25} \\Q_2 &= 155 = X_{.50} \\Q_3 &= 165.5 = X_{.75} \\X_{.90} &= 185 \\X_{.10} &= 133.5\end{aligned}$$

The Mode. Another measure of central position is the mode. It is commonly regarded as the "typical" item in a distribution in the sense that more observations assume the modal value than any other value. In some distributions, a mode may not be well defined because there is no tendency to cluster; other distributions may have two or even three modes. In small samples, a mode may not appear in the array even though it appears in the frequency distribution.

Example: A good mode is defined only in the array of ages. The modal age is twenty-one years. No one weight or height can be considered a mode because not enough observations occur at any one value.

The Arithmetic Mean. The arithmetic mean is the most commonly used average of everyday conversation. It is simply the sum of the values of the observations divided by the number of observations. The mean weight of students in a class is the total weight of all students divided by the number of students. Symbolically, the arithmetic mean of a sample is $\bar{X} = \Sigma x/n$, a notation used earlier in this text.

Example: The arithmetic mean weight, height, and age, of the data in Table 7.1 are:

Weight: 157.2 pounds
Height: 69.25 inches
Age: 22.5 years

The arithmetic mean has one property which influences its use as a descriptive average in many practical situations. It is affected by the value of every observation in the sample. Extremely large or extremely small values may seriously affect its magnitude and render it atypical of the average of the distribution. The mean is widely used in inference problems because it can be manipulated algebraically and because the probability distribution of sample means has been investigated under a

wide variety of circumstances. Accordingly, it has also been used traditionally as an average which is descriptive of a sample. Its algebraic form of expression is convenient, but students should be wary of interpreting it as describing a sample set of data. The median, the mode, or the quartiles and percentiles provide much more useful and interpretable information in very many descriptive problems.

The Range. The range is a measure of variation based upon the extreme values in a set of numbers. It is customarily reported in one of two ways: (1) the range of weights (Table 7.1) is from 120 to 195 pounds; (2) the range of weights is 75 pounds. The preferable method, however, is to combine these two by saying the range of weights is 75 pounds, from 120 to 195. The figure 75 is more meaningful when the two extreme values (120 and 195 in this case) are indicated.

Although most people know what the range means, they often overlook an important property and thus fall into statistical fallacy. The range depends upon sample size: the larger the sample, the larger the expected range. Failure to take into account differences in sample size frequently leads to fallacious conclusions about the variability in two samples. A small experiment will demonstrate the relationship between the range and the sample size. Ask two persons in a group for their weights, and compute the range. Then, find the range of the weights of the whole group. It will probably exceed the range of the two weights because there is a greater chance of observing an extreme value with a larger sample.

The range is a very useful measure of dispersion in inference problems. The range of small samples is used quite widely in statistical quality-control work to measure the variability of a product characteristic.¹

Interquartile Range. Another measure of dispersion is the difference between the first and third quartiles: $Q_3 - Q_1$. By definition, this interval contains approximately 50 per cent of the observations. It is a measure of dispersion in the middle of a distribution. The longer the *interquartile range* the more values of the variable are necessary to account for half the total number of observations and the more the dispersion in the data. This range, of course, gives no information about the extreme values in the tails of the distribution.

A related measure is the *semi-interquartile range* or the *quartile deviation*, defined by $(Q_3 - Q_1)/2$. It is the average of the deviations of Q_3 and Q_1 from the median. If the distribution is symmetrical about the

¹ The literature of elementary statistics often scoffs at the range as a poor measure of variation because it fluctuates widely from sample to sample, because it takes account of only the extremes and not the middle observations of a distribution, and particularly because such literature has paid more attention to description than to inference problems.

median, the median plus or minus one quartile deviation will include approximately 50 per cent of the values of the variable.

The Variance and Standard Deviation. The most important measure of dispersion is the *standard deviation*. It measures the variation of observations about their arithmetic mean. The fundamental definition is

$$s_x = \sqrt{\frac{\Sigma(x - \bar{X})^2}{n - 1}} \quad (7.1)$$

Each observation is measured as a deviation from the mean, and the deviation for each value is squared. The sum of these squares is computed and divided by $n - 1$ (one less than the sample size); the square root of this quotient is the standard deviation, designated by s_x in the sample.

A measure of variation related to the standard deviation is the *variance*. It is the square of the standard deviation and will be designated by s_x^2 for the sample. The standard deviation is important for some statistical problems and the variance for others. The variance is used in the next sections because the troublesome radical sign can be omitted.

$$s_x^2 = \frac{\Sigma(x - \bar{X})^2}{n - 1} \quad (7.2)$$

The fundamental definitions of the variance and standard deviation must be learned. A simple example will illustrate their computation. What are the variance and standard deviation of the numbers 1, 2, 3, 4, 5? Their mean is $\bar{X} = 3(15/5)$. Each number may be expressed as a deviation from the mean and the subsequent computations arranged as follows:

x	$x - \bar{X}$	$(x - \bar{X})^2$
1	-2	4
2	-1	1
3	0	0
4	+1	1
5	+2	4
Sum	.	10

$$s_x^2 = \frac{\Sigma(x - \bar{X})^2}{n - 1} = \frac{10}{4} = 2.5 \quad s_x = \sqrt{2.5} = 1.58$$

No one (except students in elementary statistics) will actually compute deviations from the mean, square them, sum them, etc., to find either the standard deviation or the variance. This process entails an unnecessary amount of tedious work, unless the arithmetic mean happens

to be a whole number. A satisfactory computational formula is derived from an expansion of the basic formula for the variance.

$$s_x^2 = \frac{\Sigma(x - \bar{X})^2}{n - 1} = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1} \quad (7.3)$$

The basic data in this expanded formula are Σx and Σx^2 . Both of these terms appeared in the discussion of notation and computational methods (see Sec. 2.1 and the exercises). The sum and the sum of the squares of the original observations are easily computed with an adding machine and a table of squares, or by cumulating the squares in a computing machine.

Example:

x	x^2
1	1
2	4
3	9
4	16
5	25
15	55

$$s_x^2 = \frac{55 - 15^2/5}{4} = \frac{55 - 45}{4} = \frac{10}{4} = 2.5$$

$$s_x = \sqrt{2.5} = 1.58$$

Example: From the data of Table 7.1,

Weight:

$$\Sigma x = 6288 \quad \Sigma x^2 = 1,002,264 \quad n = 40$$

$$s^2(\text{wt.}) = \frac{1,002,264 - 6288^2/40}{39} = 353.6 \quad s(\text{wt.}) = 18.8$$

Age:

$$\Sigma x = 899 \quad \Sigma x^2 = 21,105$$

$$s^2(\text{age}) = \frac{21,105 - 899^2/40}{39} = 23.1 \quad s(\text{age}) = 4.8$$

Height:

$$\Sigma x = 2770 \quad \Sigma x^2 = 192,291$$

$$s^2(\text{ht.}) = \frac{192,291 - 2770^2/40}{39} = 12.0 \quad s(\text{ht.}) = 3.5$$

7.8 Measures of Central Position and Dispersion: Grouped Data

When data are grouped together in a frequency distribution, the methods of computing the various measures of average and dispersion vary from those used with the raw data. These methods are given in this section.

Arithmetic Mean. What is the mean of the series of numbers 1, 1, 2, 2, 2, 3, 4, 5, 5, 5, 6, 7, 7, 8? The sum of these numbers could be computed on an adding machine, but suppose they were grouped as follows:

x	f	fx
1	2	2
2	3	6
3	1	3
4	1	4
5	3	15
6	1	6
7	2	14
8	1	8
Total	14	58

This very simple frequency distribution groups together and records the frequency of those numbers which are alike. The total of these 14 numbers may be obtained by adding the products of each value of x and its associated frequency. This sum is

$$2 \cdot 1 + 3 \cdot 2 + 1 \cdot 3 + 1 \cdot 4 + 3 \cdot 5 + 1 \cdot 6 + 2 \cdot 7 + 1 \cdot 8 = 58$$

The total is shown in the table, and the mean is $58/14 = 4.1$.

Here is a model of a frequency distribution without the class intervals. Class intervals can be ignored because each observation in an interval, since its original value is unknown, will be given the value of the *mid-point* of the interval. The general formula is

$$\bar{X} = \frac{\sum f \cdot x}{n} \quad (7.4)$$

Example: The arithmetic mean of the weights given in Table 7.3 is computed in Table 7.7.

TABLE 7.7. COMPUTATION OF MEAN WEIGHT

Class interval	Mid-point x	Frequency f	$f \cdot x$
112 5-127 5	120	2	240
127 5-142 5	135	8	1080
142 5-157 5	150	13	1950
157.5-172 5	165	9	1485
172 5-187 5	180	5	900
187 5-202.5	195	3	585
Total	..	40	6240

$$\bar{X} = \frac{6240}{40} = 156$$

Computing an arithmetic mean from a frequency distribution with an open-end interval may be bothersome. If the exact values of the observations in these open-end intervals are footnoted (as the author has suggested), they may be added into the total found from the closed intervals by the method given above. Too often, however, the values of such observations are not explicitly given, and they must then be assumed. Values thus arbitrarily assigned will, of course, affect the mean.

The Median. The median is customarily computed from the frequency distribution by using linear interpolation since the exact values of the observations in any interval are unknown. The formula will be omitted because it is unduly complicated to write, whereas the basic method is simple to learn. The median weight of Table 7.3 is found in Table 7.8.

TABLE 7.8. COMPUTATION OF THE MEDIAN WEIGHT

Class bounds	f	Cumulative $f(<)$
112.5-127.5	2	2
127.5-142.5	8	10
142.5-157.5	13	23
157.5-172.5	9	32
172.5-187.5	5	37
187.5-202.5	3	40

$$\bar{x} = 142.5 + \frac{10}{13}(15) = 154$$

The data are first *cumulated* on a *less-than* basis; i.e., the frequencies are successively added together. Thus, there are 2 observations less than 127.5; 10 observations less than 142.5; 23 observations less than 157.5, etc. Since $n = 40$, $n/2 = 20$, and the median lies somewhere in the interval from 142.5 to 157.5. It is at least 142.5 pounds. Of the 13 observations in the median class interval, 10 are needed to make the cumulative frequency equal to 20 ($n/2$). On the assumption that the 13 observations in the median class are equally spaced within the interval (linear interpolation), 10/13 of the length of the class interval (15 pounds) is added to 142.5 to obtain the median. The symbol $\bar{x}$ is used for the median, and $\bar{x} = 142.5 + (10/13)15 = 154$. This value compares with the median of 155 obtained from the array in Sec. 7.7.

The median can readily be computed from a frequency distribution with an open-end interval because it is not affected by extreme values. It is computed from the order of the observations. A median weight of 154 pounds means that approximately half the students weigh more than 154 pounds and half weigh less than 154 pounds.

Quartiles and Percentiles. The quartiles and percentiles of a frequency distribution are obtained in the same manner as the median.

Interpolation is used to divide the distribution into quarters or percentages depending upon the measure desired. Thus, for Table 7.8, the following measures are computed:

$$Q_1: n/4 = 10; Q_1 = 142.5 \text{ (no interpolation is necessary)}$$

$$Q_2 = \text{median} = 154$$

$$Q_3: 3n/4 = 30; Q_3 = 157.5 + (7/9)15 = 169.2$$

$$X_{.90}: 9n/10 = 36; X_{.90} = 172.5 + (4/5)15 = 181.5$$

$$X_{.10}: n/10 = 4; X_{.10} = 127.5 + (2/8)15 = 131.3$$

The Mode. The modal value of a frequency distribution is affected by the choice of the class interval and is, therefore, an approximate measure at best. The mode may not be well defined in a distribution if the observations do not tend to concentrate in one interval. Other distributions may have more than one modal interval.

Elaborate formulas exist for computing the mode, but for most purposes a crude mode may be defined and used. In a frequency distribution, the modal class is usually enough to indicate the modal value. Thus, there were more students who weighed between 142.5 and 157.2 pounds than there were students of any other weight (Table 7.3); there were more students who were 21 and under 22 years old than there were students of any other age (Table 7.5); there were more students who were between 68.25 and 69.75 inches tall than there were students of any other height (Table 7.6).

The Range and Interquartile Range. The range is approximated in a frequency distribution by the lower limit of the first interval and the upper limit of the last interval. The range will be known in an open-end distribution only if exact values are footnoted.

The interquartile range and quartile deviation are determined from the quartiles. For the data of Table 7.8, the interquartile range is from 142.5 to 169.2 pounds. This range includes approximately 50 per cent of all the weights in the sample. The quartile deviation is

$$(169.2 - 142.5)/2 = 8.3 \text{ pounds}$$

The Variance and Standard Deviation. The methods of computation used with original observations may also be employed for data in the form of frequency distributions. The values assigned to the observations are the mid-points of the class intervals.

Since some values of x occur more than once, each deviation of x from $\bar{X}$ must be multiplied by the appropriate frequency. The basic formulas are

$$s_x^2 = \frac{\sum f(x - \bar{X})^2}{n - 1} \quad \text{and} \quad s_x = \sqrt{\frac{\sum f(x - \bar{X})^2}{n - 1}} \quad (7.5)$$

Again, the expanded form is the efficient computational method.

$$s_x^2 = \frac{\Sigma fx^2 - (\Sigma fx)^2/n}{n - 1} \quad s_x = \sqrt{\frac{\Sigma fx^2 - (\Sigma fx)^2/n}{n - 1}} \quad (7.6)$$

As an example of the use of Eq. (7.6), the variance and standard deviation of the weight data of Table 7.3 are computed in Table 7.9.

TABLE 7.9 VARIANCE AND STANDARD DEVIATION OF WEIGHTS

Mid-point x	Frequency f	$f \cdot x$	$f \cdot x^2$
120	2	240	28,800
135	8	1080	115,800
150	13	1950	292,500
165	9	1485	245,025
180	5	900	102,000
195	3	585	144,075
Total	40	6210	988,200

$$s_x^2 = \frac{988,200 - 6240^2/40}{39} = 378.46 \quad s_x = \sqrt{378.46} = 19.45$$

7.9 Relationships among the Mean, Median, and Mode

In a symmetrical distribution, the mean and median coincide; if the distribution is unimodal like the normal distribution, they also coincide with the mode. In asymmetrical or skewed distributions, the mean, median, and mode are not equal. In a positively skewed distribution,

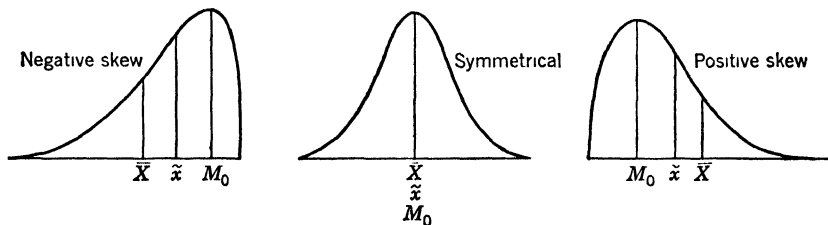


FIG. 7.3 Relationships among the mean ($\bar{X}$), the median ($\tilde{x}$), and the mode (M_0).

the mean is the largest, the median the next largest, and the mode the smallest of the three. In a negatively skewed distribution, the mean is the smallest, the median the next smallest, and the mode the largest of the three. These relationships are shown in Fig. 7.3.

A statistician who knows only the mean, median, and mode of a distribution can infer something about the general shape of the distribution—the direction of its skewness—from the relationships among these three measures.

7.10 Final Remark

The author believes that the computation of means, variances, and standard deviations from frequency distributions involves unnecessary expenditure of effort when the raw data are available. There is, of course, no choice when the data are available only in the form of a frequency distribution. But when the raw data are at hand, these computations can usually be completed in less time than it takes to decide upon the class intervals for the frequency distribution.

When computing equipment, even the desk calculator, was not widely available, grouping observations together into class intervals did simplify the arithmetic involved in many calculations. With even so simple a device as an adding machine, the original observations can be used in efficient computing. This is not to say that a frequency distribution is not an important tool in summarizing data. It is very useful when a tabular or graphic picture of the distribution is wanted, but it seldom pays to make a frequency distribution for the sole purpose of further computations. This is true even with as many as 10,000 observations, for example. Any sampling problem on so large a scale will make use of modern punch-card methods by which sorting, tabulating, and calculating can be done very rapidly.

EXERCISES

7.3. Use the height, weight, and age data (Exercise 2.21) in this exercise.

- Find the median from the array and the frequency distribution, and interpret it.
- Find the three quartiles from the array and the frequency distribution, and interpret them.
- Is a mode well-defined in the array? What is the modal interval of the frequency distribution? Interpret your modal value.
- Find the mean from the raw data.
- What is the range of the data?
- What is the interquartile range? Interpret it.
- Find the variance and the standard deviation from the raw data.
- Examine the relationship among the mean, median, and mode relative to the apparent skewness in the frequency distribution.
- If only one average could be used to summarize the central position of these variables, which average would you select and why? Would you prefer to report more than one? Why?

7.11 Comparison of the Variability in Two or More Samples

The standard deviations of weight, height, and age of the sample of 40 students for which the data were reported in Table 7.1 are

$$s(\text{wt.}) = 18.8 \quad s(\text{ht.}) = 3.5 \quad s(\text{age}) = 4.8$$

where s is the sample standard deviation. Because the standard deviation of weights is larger than the standard deviation of either age or height, is it valid to conclude that weights are more variable than the ages or heights of the 40 students? These standard deviations are not directly comparable because the units of measurement differ. Weights are measured in pounds, ages in years, and heights in inches. To compare validly the variability in these three distributions, a relative measure which does not depend upon the unit of measure should be used. One such measure is the *coefficient of variation*, symbolized by C.V. The coefficient of variation is the ratio of the standard deviation to the arithmetic mean,

$$\text{C.V.} = \frac{s}{\bar{X}} \quad (7.7)$$

and is commonly expressed as a percentage by multiplying by 100. Thus, the coefficients of variation for the three distributions, weight ($\bar{X} = 157.2$), age ($\bar{X} = 22.5$), and height ($\bar{X} = 69.25$), are

$$\text{C.V.}(\text{wt.}) = \frac{18.8}{157.2} = 12.0\%$$

$$\text{C.V.}(\text{ht.}) = \frac{3.5}{69.25} = 5.1\%$$

$$\text{C.V.}(\text{age}) = \frac{4.8}{22.5} = 21.3\%$$

On this basis, the age distribution is the most variable and the height distribution the least variable. Relative to the mean age, the standard deviation of the ages is larger than the standard deviation of either of the other variables relative to their respective mean value.

The coefficient of variation is a useful measure of relative variability when two variables with different units of measure are to be compared, or when the same variable with two different averages is to be compared, say, the weights of human beings and the weights of elephants. The absolute amount of variability as measured by the standard deviation is expected to be larger for elephants than for human beings because the magnitude of the weight is so much larger than for human beings. On a relative basis, it might very well be that human beings have the more variable weight.

7.12 Describing Populations with the Mean and Standard Deviation

Thus far in this chapter the descriptions have been concerned with samples. The standard deviation has real meaning in many problems in its use in describing populations.

Suppose that a variable x is *normally* distributed in the population with mean $\mu = 100$ and standard deviation $\sigma = 5$. The population is described by these three characteristics. In Sec. 4.4 the standard deviation was given real meaning in terms of its use in estimating the fraction of observations within given limits of the mean. For example, the fraction of a normal distribution between the limits $\mu \pm 3\sigma$ is approximately .997; the fraction of a normal distribution between the limits $\mu \pm 2\sigma$ is approximately .95. Knowing only the mean and standard deviation and the fact that the distribution is normal, one can describe the population distribution in terms of the fraction of observations within limits of the mean.

For example, with $\mu = 100$ and $\sigma = 5$, the fraction of the observations of x within $100 \pm (3)(5)$, or from 85 to 115, is approximately .997. The fraction of observations between $100 \pm (2)(5)$, or from 90 to 110, is approximately .95. The fraction of observations within 100 ± 5 , or from 95 to 105, is approximately .68. When it is known that the population is a normal distribution, the mean and standard deviation can be combined to make meaningful descriptive statements about that population.

What if it cannot be assumed that the population is a normal distribution? What then can be said about the fraction of observations within limits of the mean from a knowledge of μ and σ ? *Without any assumption about the form of the population*, the mean and standard deviation can be combined to describe the population according to the following inequality: *the fraction of observations inside the limits $\mu \pm k \cdot \sigma$ is at least $1 - 1/k^2$* . This inequality is known as Tchebycheff's inequality. Thus, the fraction of observations inside the limits $\mu \pm 3\sigma$ ($k = 3$) is *at least* $1 - (1/3)^2$, or $8/9$, which is approximately 89 per cent. The fraction of observations inside the limits $\mu \pm 4\sigma$ ($k = 4$) is *at least* $1 - (1/4)^2$, or $15/16$, which is approximately 94 per cent. For values of $k > 1$, the fraction of observations for any value of k may be determined from the inequality.

The actual fraction of observations inside given limits may be much more than that determined from the application of Tchebycheff's inequality. This inequality only states that the fraction is *at least* a certain amount. If the actual form of the distribution is known (e.g., that it is normal), a more accurate statement about the fraction can be made. Tchebycheff's inequality makes no assumption whatever about the form of the population. Its application, as well as that of the normal distribution, if appropriate, depends upon knowledge of the population standard deviation σ . When σ is unknown, the inequality is not applicable.

7.13 Coding Data for Simplified Computation

The arithmetic mean has important properties that are useful in simplifying its computation. How would you average the numbers 10, 14, 15, and 17?

1. Some of you may average the numbers as follows:

$$10 + 14 + 15 + 17 = 56 \quad \bar{X} = \frac{56}{4} = 14$$

2. Some of you may do the following:

$$0 + 4 + 5 + 7 = 16 \quad \frac{16}{4} = 4 \quad \bar{X} = 10 + 4 = 14$$

3. Others may do this:

$$-4 + 0 + 1 + 3 = 0 \quad \bar{X} = 14$$

The property of the mean illustrated by these methods may be formalized in an algebraic equation. Let x be a random variable with mean $\bar{X} = \Sigma x/n$. Suppose a constant a is added to each value of x to form a new variable y , where $y = x + a$. What is the mean of y ? As an example, suppose that the mean weight of a sample is 156 pounds and that a constant of 2 pounds is added to each weight. What is the new mean weight? It is 158 pounds. This result is derived in general as shown below.

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{\Sigma(x + a)}{n} = \frac{\Sigma x}{n} + \frac{\Sigma a}{n} = \bar{X} + a$$

The trick here is to realize that a is a constant and that Σa (over n observations) is $n \cdot a$. As an example, $\Sigma 2$, when $n = 5$, is

$$2 + 2 + 2 + 2 + 2 = 5 \cdot 2 = 10$$

A general statement of this property of the arithmetic mean is that if the same constant is added to each observation in a set of numbers, that constant is also added to the arithmetic mean of the numbers; or, in a slightly different form, the addition of a constant to every observation increases the mean by that constant.

What effect does subtracting a constant from every observation have on the mean?

$$\begin{aligned} y &= x - a \\ \bar{Y} &= \bar{X} - a \end{aligned}$$

Subtracting a constant from every observation subtracts that constant from the mean. What is the effect on the mean of multiplying every observation by a constant?

$$y = c \cdot x \quad (c \text{ a constant})$$

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{\Sigma(c \cdot x)}{n} = c \left(\frac{\Sigma x}{n} \right) = c \cdot \bar{X}$$

Multiplying every observation by a constant multiplies the arithmetic mean by that constant.

If $y = x/c$,

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{\Sigma x/c}{n} = \frac{1}{c} \cdot \frac{\Sigma x}{n} = \frac{\bar{X}}{c}$$

Dividing each observation by a constant divides the arithmetic mean by the same constant.

The arithmetic of computing variances and standard deviations may also be lightened by similarly transforming the variable; a constant may be added to or subtracted from each observation, or each observation may be multiplied or divided by a constant. The development is necessarily algebraic, but the final results are more important than the derivations. Let x be a variable with variance s_x^2 and standard deviation s_x . If x is transformed into the variable y , where $y = x - a$ and a is a constant, what are s_y^2 and s_y ? First, $\bar{Y} = \bar{X} - a$.

$$\begin{aligned} s_y^2 &= \frac{\Sigma(y - \bar{Y})^2}{n - 1} = \frac{\Sigma[(x - a) - (\bar{X} - a)]^2}{n - 1} = \frac{\Sigma(x - a - \bar{X} + a)^2}{n - 1} \\ &= \frac{\Sigma(x - \bar{X})^2}{n - 1} = s_x^2 \end{aligned}$$

$$s_y = s_x$$

That is to say, subtracting a constant a from every value of x does not change the variance or standard deviation. In similar fashion it may be shown that adding a constant a to each value does not change the standard deviation or the variance. This is an important property which may be illustrated as follows. Suppose the variable x is distributed about $\bar{X}$, as shown in Fig. 7.4b, with standard deviation s_x . The distribution of $y = x - a$ is shoved along the axis to a new mean, $\bar{X} - a$, but its variation is not changed ($s_y = s_x$). The whole distribution in its original shape is moved down the scale to a new mean, as shown in Fig. 7.4a.

Suppose that every observation is multiplied by a constant c .

$$\begin{aligned} \bar{Y} &= c \cdot \bar{X} \\ s_y^2 &= \frac{\Sigma(y - \bar{Y})^2}{n - 1} = \frac{\Sigma(c \cdot x - c \cdot \bar{X})^2}{n - 1} = \frac{c^2 \Sigma(x - \bar{X})^2}{n - 1} \\ &= c^2 \cdot s_x^2 \\ s_y &= c \cdot s_x \end{aligned}$$

The variance and standard deviation change. The variance is multiplied by c^2 and the standard deviation by c . Figure 7.4c shows the dis-

tribution of variable x . The new distribution is shifted along the axis to a new mean $\bar{Y} = c \cdot \bar{X}$, and at the same time the standard deviation is changed to $c \cdot s_x$. It is larger than s_x if $c > 1$. The spread of the distribution has changed as well as its position on the scale (Fig. 7.4d).

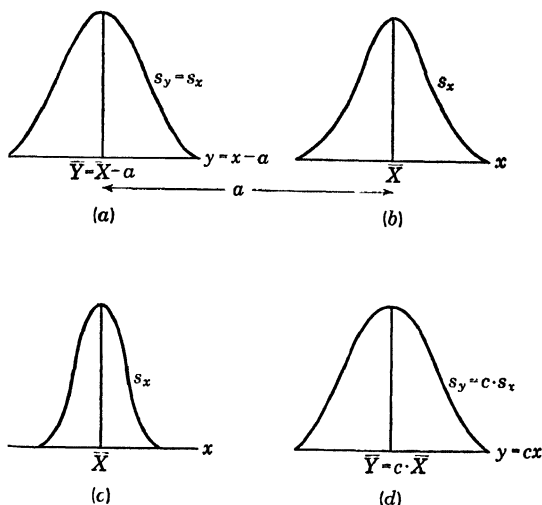


Fig. 7.4. Illustrations of the effects of subtracting and multiplying by constants on the mean and standard deviation.

Likewise it may be shown that dividing by a constant c changes the variance and standard deviation. If $y = x/c$,

$$s_y^2 = \frac{s_x^2}{c^2} \quad s_y = \frac{s_x}{c}$$

The spread of the distribution will be less if $c > 1$.

Occasionally a combination of adding (or subtracting) a constant and multiplying (or dividing) by a constant may be used to good advantage. An example is the use of the coding, $y = (x - a)/c$, for computing the variance and standard deviation of the height data of Table 7.1. Heights were measured to the nearest half inch so that each observation is reported to one decimal place and three digits. The squares of the original observations have six digits and two decimal places. If the value of 60.0 inches (the smallest height) is chosen for a , and the value .5 (the unit of measure) for c , numbers much smaller than the original observations may be used in the computation. With these values, the coded variable is

$$y = \frac{x - 60.0}{.5} \quad \text{or} \quad y = 2(x - 60.0)$$

From Table 7.1, values of x may be immediately coded to values of y , as

shown in the table below; these y values may be readily squared because they are whole numbers.

x	$y = 2(x - 60)$	y^2
70.0	20	400
64.5	9	81
64.0	8	64
70.0	20	400
71.0	22	484
Total ..	742	15,636

These results are quickly obtainable by using a table of squares and an adding machine.

$$\bar{Y} = \frac{742}{40} = 18.55 \quad y = \frac{x - a}{c}$$

Therefore

$$\bar{Y} = \frac{\bar{X} - a}{c}$$

$$\bar{X} = c \cdot \bar{Y} + a = (.5)(18.55) + 60 = 69.3 \text{ in.}$$

From the expanded form of Eq. (7.3),

$$s_y^2 = \frac{15,636 - 742^2/40}{39} = 41.997, \text{ or } 45.0 \quad s_y = \sqrt{45.0} = 6.7$$

The variance and standard deviation of height may be found by solving the relationship between s_y^2 and s_x^2 , depending upon the coding operation used. Subtraction of the constant a has no effect on the variance; division by the constant c has the effect of dividing the variance by c^2 . Therefore, the relationship is

$$s_y^2 = \frac{s_x^2}{c^2} \quad \text{or} \quad s_y = \frac{s_x}{c}$$

Since s_y^2 and s_y are computed from the sample and since c is known, these equations may be solved for s_x^2 and s_x :

$$s_x^2 = c^2 \cdot s_y^2 \quad s_x = c \cdot s_y$$

$$s_x^2 = (.5)^2(45.0) = 11.25 \quad s_x = (.5)(6.7) = 3.35$$

The property that the variance and standard deviation are not changed when a constant is added to or subtracted from each observation is of great computational usefulness. The magnitudes of the observations may be greatly reduced but without affecting the measures of dispersion.

EXERCISES

7.4. Compute $\bar{X}$ and s_x for height and weight (Exercise 2.24) by coding weights to a variable $(x - 100)$ and heights to a variable $2(x - 60)$. Is the standard deviation the largest for weight or height? For which is the variability the greatest? Why?

7.5. From the 1950 Census of Population report for the state in which you now live, copy out the income distribution of families and unrelated individuals for the urban place where you live (or the nearest urban place if you live in a rural area). Plot this distribution in the form of a histogram taking care with unequal and open-end intervals. Is there a central tendency in these data? If the distribution is skewed, in what direction is the skewness?

What average income is reported? Why is this average used? Is the modal income readily apparent? What do both of these averages mean?

7.6. How does your income compare with the average of the income distribution of Exercise 7.5? How did you define your income? Check your definition of income with that given in the introduction to the Census report. Recompute your income if your definition differs from that of the Census and again compare it with the average.

7.7. Upon entering a college or university, students are generally given a battery of tests in various subject matter fields. The scores on these tests are usually reported in percentiles. What useful information is contained in the following report of the percentiles for one student: mathematics, 95; English, 75; foreign language, 15; social science, 90; physical science, 90?

7.8. An article in a medical journal recently reported the results of an experiment measuring the temperature of children of different ages. The results are given in the table below. What conclusions do you draw about the variability in temperatures among children of different ages?

Age of child	Number of children	Range of temperature	Average temperature
2	645	26	99.72
2½	589	.24	99.71
3	1398	.38	99.71
4	1037	.34	99.69
5	1244	35	99.68

7.9. A survey of households within a 50-mile radius of Ames, Iowa, was undertaken in November, 1951, to estimate the size and characteristics of television audiences in this area. The data shown at the top of page 166 are taken from this survey.

What are the mean, median, and modal size screen? Interpret each of these averages in the terms of this problem. Which average is best for this problem?

7.10. Comment on the statement that the mean, median, and mode become identical when the whole population and not just a sample is examined.

HOUSEHOLD TV SETS BY SCREEN SIZE

Screen size, in.	Number of sets	Screen size, in.	Number of sets	Screen size, in.	Number of sets
7 0	2	10 5	5	16 0	142
7 5	1	11 0	1	17 0	285
8 0	1	12 0	17	17 5	1
8 5	1	12 5	41	19 0	16
9 0	2	14 0	34	20 0	30
10 0	14				

7.11. Comment on the conclusion that someone is clearly misrepresenting the case in a labor dispute because the representative of the employer claims that the average hourly earnings in the shop are 156.5, while the representative of the employees claims that the prevailing wage in the shop is but 125 cents.

7.12. The mean grade on an examination in statistics is reported by the instructor as 78. What else would you like to know about the grades on this examination in order better to assess your own score?

7.13. An instructor has decided to grade on the "curve"—the curve being the normal distribution. He has decided to give 7 per cent A's, 20 per cent B's, 46 per cent C's, 20 per cent D's, and 7 per cent F's. If $\mu = 75$ and $\sigma = 8$, what numerical grades are the dividing lines among the five letter grades? What are the conditions for the valid application of grading on the curve? Would you want to be graded on the curve in, say, this course? Why or why not? What kinds of tests would lend themselves to grading on the curve?

7.14. Secure the batting averages of the leading 20 batters in both the American and the National League. Find $\bar{X}$ and s_x . In which league are the averages more variable?

7.15. In 1946 the mean monthly employment in a particular plant was 23,000 with a standard deviation of 1000. In 1954 the mean monthly employment was 15,000 with a standard deviation of 750. How would you measure employment stability? According to this measure, in which period was employment more stable?

REFERENCES

- Haemer, Kenneth (ed.): Presentation Problems, *The American Statistician* (Washington). This is the news publication of the American Statistical Association, one of its features is a section on presenting statistics.
- Modley, Rudolf, and Dyno Lowenstein: "Pictographs and Graphs," Harper & Brothers, New York, 1952.
- Myers, John H.: "Statistical Presentation," Littlefield, Adams & Co., Ames, Iowa, 1950.
- Spear, Mary Eleanor: "Charting Statistics," McGraw-Hill Book Company, Inc., New York, 1952.
- U.S. Department of Commerce, Bureau of the Census, "Manual of Tabular Presentation," Government Printing Office, Washington, 1949.

CHAPTER 8

STATISTICAL INFERENCE: FREQUENCY DISTRIBUTIONS

As more and more observations are taken in a sample—that is, as the sample size becomes larger and larger—the class intervals of a frequency distribution are often made smaller so as to represent the underlying distribution accurately. As the bars of the corresponding histogram become narrower and narrower, an attempt to approximate the distribution by a smooth curve seems natural. Even when the variable is discrete (i.e., can assume only definite integral values), a smooth curve may be sought as an approximation. For example, the discrete binomial distribution has already been approximated by the normal distribution—a smooth, continuous curve (see Sec. 4.3 and Fig. 4.3). Variables like weight, age, and height are continuous variables which can assume any value within their range. As the sample size becomes larger, the sample distribution approaches the population distribution, and the continuous nature of these variables leads to a search for a continuous mathematical curve to characterize their distributions.

The normal distribution has historically played an important role in describing the actual distribution of many measurements. It was discovered in the eighteenth century by the mathematician De Moivre during his investigations of the properties of the binomial distribution. He found that the normal distribution is a useful approximation to the binomial. Another eighteenth-century mathematician, Gauss, investigated the theory of errors of measurement. On certain very plausible hypotheses (mainly that the errors are due to many small, independent, chance causes), errors of measurement ought to be normally distributed. Gauss found very close agreement between the normal curve and the distribution of actual measurements. This led to the belief that almost all measurements are normally distributed, an unfounded belief which many students of elementary statistics unfortunately retain after completing their course.

Although not all measurements are normally distributed, much con-

fidence can still be placed in the normal distribution and its use in statistics. Now, however, its greatest use lies in the area of sampling statistics and not in the description of the distribution of observations of natural phenomena. Especially in business and economic statistics, distributions tend to be skewed. A great variety of distributions do exhibit the familiar bell shape, but even these vary widely as to symmetry, peakedness, flatness, and length of tails. Rectangular, J-shaped, and U-shaped distributions occur in practice. The intention in this chapter is to demonstrate how a frequency distribution of data may be summarized by the use of a simple mathematical formula. The general topic is called curve fitting in the sense that the task is to see how closely a mathematical curve fits the observed data.



FIG. 8.1. Examples of curves which are good and bad fits to actual distributions.

Some mathematical curves which do or do not fit actual distributions of data are shown in Fig. 8.1. It is fairly obvious which are good fits and which are bad fits. Section 8.1 presents a measure of how well a particular curve fits a given set of data, and a method of judging the sizes of this measure.

8.1 Testing Goodness of Fit

The method of measuring the discrepancy between an observed and a theoretical distribution and of deciding when the discrepancy is so large that the theoretical distribution is not a good fit or does not adequately explain the observed distribution is developed in a very simple problem where all the parameters are known in advance. This digression from the main problem is necessary in order to provide tools for the later analysis.

An "honest" die is one with a probability of $1/6$ for each face. If an honest die is rolled 60 times, each face is expected to occur in $1/6$ of the trials, or 10 times. Table 8.1 shows the results of actually rolling a die

60 times. None of the observed frequencies are exactly equal to 10, but some faces come up more often and some less often than expected. The hypothesis to be tested is that the die rolled is honest, or that the rolls are a sample from a population with a known probability of $1/6$ for the occurrence of each face on one roll.

TABLE 8.1. OBSERVED AND EXPECTED FREQUENCIES FOR 60 ROLLS OF AN HONEST DIE

Face	Observed frequency O	Expected frequency E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
1	6	10	-4	16	1.6
2	15	10	5	25	2.5
3	7	10	-3	9	.9
4	4	10	-6	36	3.6
5	17	10	7	49	4.9
6	11	10	1	1	1
Total	60	60	0		13.6

A not very obvious but perfectly valid measure of the discrepancy between an observed and expected frequency is the quantity $(O - E)^2/E$, where O is the observed frequency and E is the expected frequency. For example, for face 1, the value of this measure is $(6 - 10)^2/10 = 1.6$. For other faces, the values are given in Table 8.1. The sum of these quantities for all faces is an index of the discrepancy of all faces from expectation. This sum, called χ^2 (chi-square), serves as an index of probability. The logic of this procedure is well worked out but is omitted here. In equation form,

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (8.1)$$

where O = observed frequency in a particular group or cell

E = expectation of the frequency in a group or cell

k = number of groups or cells in the distribution

$\sum^k$ = summation taken over all k groups or cells

In Table 8.1, the value of χ^2 is 13.6. Is this an indication that the die is honest? Or that it is loaded?

If the agreement between the observed and expected frequencies is perfect, $\chi^2 = 0$ because each $(O - E)$ is zero. Larger values of χ^2 correspond to increasingly poor agreement. Is 13.6 in the present example larger than can be expected for 60 rolls of an honest die? Or, what is the probability that a χ^2 as large as or larger than 13.6 would occur in a

random experiment of 60 rolls of an honest die? To answer this question, the probability distribution or the sampling distribution of the measure of discrepancy χ^2 is needed.

On the assumption that the hypothesis is true— that the die is honest or that the data come from a population with the specified probability of 1/6 for each face— χ^2 has a probability or sampling distribution known as the χ^2 distribution. This probability distribution is characterized by *one constant or parameter* which, for lack of a better name, is called *degrees of freedom*.

Remark. You may remember that the binomial probability distribution has one parameter P , the proportion of the population in a given category, and that the normal probability distribution has two parameters, the mean μ and the standard deviation σ .

When the population probability distribution is completely known, i.e., when there are no unknown parameters, the number of degrees of freedom is equal to $k - 1$, where k is the number of groups or cells in the frequency distribution. Different probability levels of χ^2 for several degrees of freedom are given in Table A.9 in the Appendix. The statistical test is formally set out as follows:

H_0 : Die is honest, or probability of each face is 1/6.

H_1 : Die is not honest.

$k = 6$ groups $k - 1 = 5$ degrees of freedom

$\alpha = .05$ (the arbitrary probability of rejecting the hypothesis that the die is honest when, in fact, it is honest)

R : Upper tail, because large values of χ^2 are consistent with H_1 .

From Table A.9 the value of χ^2 which is exceeded with probability .05 for 5 degrees of freedom is 11.070. This is written $\chi_{.05}^2(5) = 11.1$.

Decision Rule. Accept H_1 if $\chi^2 \geq 11.1$; otherwise accept H_0 .

Decision. Accept H_1 and conclude that the die is not fair because $\chi^2 = 13.6$ falls in the region R . The probability of observing a greater discrepancy than appeared in Table 8.1 is too small (less than $\alpha = .05$) to conclude that the discrepancy is due to chance fluctuations. The conclusion is that the theory is inadequate to fit the facts, or that the die is dishonest.

Remark. If you grasped the meaning of the first few chapters of this book, the following question should now be on the tip of your tongue: What are the operating characteristics of this test? In other words, if the die is dishonest, what is the probability (β) of deciding to accept H_0 , that it is honest? Only the probability (α) of accepting H_1 when, in fact, H_0 is true is specified in the test procedure given above.

This question is a good one and, in principle, merits an answer. In practice, the answer is very difficult because there are many probability distributions for which a die could be called dishonest. For example, one such distribution might specify that the probability of face 6 is 1/3, the probability of face 1 is 1/3, and the probabilities of the remaining faces are each 1/12. The sampling distribution of χ^2 must

then be calculated, using this probability distribution for the die, and the probability that $\chi^2 \leq 11.1$ (the value which leads to the acceptance of H_0) computed. The fact is that the value of β for well-defined alternatives of this kind is not known, and subsequent statistical decision rules will be based upon a choice of α alone. The value of β is determined largely by the sample size, and one should be aware that in small samples, the statistical decision may be to accept H_0 because of the large value of β even for serious departures of the true distribution from H_0 .

EXERCISES

8.1. Compare the observed frequencies of Table 1.5 with the expected frequencies given by the binomial distribution with $n = 10$ and $P = .5$. What conclusions do you draw about this bead-box experiment?

8.2. Compare the observed frequencies of Exercise 2.23 with the binomial distribution with $n = 10$ and $P = .2$. What conclusions do you draw about your bead-box experiment? (See Remark of Sec. 8.3.)

8.3. Roll two dice 108 times and record the total number of dots on each roll. Make a frequency distribution of the number of dots. Compare the observed frequencies of this experiment with the expected frequencies given by the probability distribution derived in Exercise 2.3. What conclusions do you draw about the dice?

8.2 Fitting a Normal Distribution

Davis has presented some very interesting data on the failures of different kinds of systems: bank-statement and ledger-card errors, student typing errors, lost identification badges, payroll-check errors, bus motor failures, and the like.¹ Table 8.2 is the frequency distribution of the number of miles to first failure of bus motors operated by a large city bus company. The "failure was either abrupt, in which some part broke and the motor would not run; or, by definition, when the maximum power produced, as measured by a dynamometer, fell below a fixed percentage of the normal rated value. Failures of accessories which could be easily replaced were not included in these data."

Davis postulates a *normal theory of failure* for some systems, so named because the normal probability distribution is the mathematical curve which fits the observed data. This theory is postulated for the first bus motor failure data. The expected values of initial bus motor failures given in Table 8.2 are found by fitting a normal distribution to the observed data. The random variable is designated by t (in place of the usual x). The two parameters of the normal distribution, μ_t and σ_t , are unknown, but may be estimated from the sample by $\bar{t}$ and s_t , the mean

¹ See D. J. Davis, An Analysis of Some Failure Data, *Journal of the American Statistical Association*, vol. 47, pp. 113-150, June, 1952. Tables, figures, and discussions from this article are reproduced in this chapter with the kind permission of the author and the *Journal*.

and standard deviation, respectively, of the sample. The standardized normal variable Z , which is $Z = (t - \mu_t)/\sigma$ for the population, becomes $Z = (t - \bar{t})/s_t$ for the sample. Fitting a normal distribution consists of computing $\bar{t}$ and s_t , transforming the class limits of the frequency distribution into Z values, and reading from a table of areas of the normal

TABLE 8.2 INITIAL BUS MOTOR FAILURES*

Distance interval, thousands of miles	Observed number of failures	Expected number of failures
Less than 20	6	3.6
20-40	11	8.2
40-60	16	18.5
60-80	25	31.3
80-100	34	39.9
100-120	46	38.3
120-140	33	27.6
140-160	16	15.0
160-180	2	6.1
180 up	2	2.5
Total	191	191.0

* SOURCE: D. J. Davis, An Analysis of Some Failure Data, *Journal of the American Statistical Association*, vol. 47, p. 145, June, 1952.

distribution (Table A.3) the areas between the intervals of Z . These areas are the expected relative frequencies of occurrence of t under the assumption of a normal distribution; they are converted into expected frequencies by multiplying by the total number of observations in the sample.

The mean and standard deviation of initial bus motor failures are computed in Table 8.3. Mid-points of the class intervals are used for values of t ; they are arbitrarily assigned to the open-end intervals. Since these arbitrary values affect both the mean and standard deviation, there will be some discrepancy between the expected values computed here and those given in Table 8.2. The discrepancy should not, however, be large (see Table 8.4).

The computations for fitting the normal distribution are given in Table 8.4.

Example: For the interval of t from 120 to 140,

$$\begin{aligned}\bar{t} &= 96.7 & s_t &= 37.5 \\ t &= 120 & Z &= \frac{120 - 96.7}{37.5} = .62 \\ t &= 140 & Z &= \frac{140 - 96.7}{37.5} = 1.15\end{aligned}$$

Area between $Z = .62$ and $Z = 1.15$ is .1425 from Table A.3. This area is the expected relative frequency in the interval 120 to 140 for a normal distribution with mean 96.7 and standard deviation 37.5. Since $n = 191$, the expected frequency is n times the expected relative frequency, or $191(.1425) = 27.2$.

TABLE 8.3. COMPUTATION OF THE MEAN AND STANDARD DEVIATION OF INITIAL BUS MOTOR FAILURES

Interval of t	Mid-point t	f	$f \cdot t$	$f \cdot t^2$
Less than 20	10	6	60	600
20-40	30	11	330	9,900
40-60	50	16	800	40,000
60-80	70	25	1,750	122,500
80-100	90	34	3,060	275,400
100-120	110	46	5,060	556,600
120-140	130	33	4,290	557,700
140-160	150	16	2,400	360,000
160-180	170	2	340	57,800
180 and up	190	2	380	72,200
Total	.	191	18,470	2,052,700

$$\bar{t} = \frac{18,470}{191} = 96.7^* \quad s_t = \sqrt{\frac{2,052,700}{190} - \frac{18,470^2}{191}} = 37.5$$

* This value compares with a mean of 97 reported by D. J. Davis, An Analysis of Some Failure Data, *Journal of the American Statistical Association*, vol. 47, pp. 113-150, June, 1952.

TABLE 8.4. FITTING A NORMAL DISTRIBUTION TO INITIAL BUS MOTOR FAILURES

Interval of t	Interval of Z^*	Expected relative frequency	Expected frequency n (relative frequency)	Expected frequency (Table 8.2)
Less than 20	Less than -2 045	.0205	3 9	3 6
20-40	-2 045 - -1 51	.0450	8 6	8 2
40-60	-1 51 - -0 98	.0980	18 7	18 5
60-80	-0 98 - -0 44	.1665	31 8	31 3
80-100	-0 44 - +0 09	.2059	39 3	39 9
100-120	0 09 - 0 62	.1965	37 5	38 3
120-140	0 62 - 1 15	.1425	27 2	27 6
140-160	1 15 - 1 69	.0796	15 2	15 0
160-180	1 69 - 2 22	.0323	6 2	6 1
Over 180	Over 2 22	.0132	2 5	2 5
Total	.	1 0000	190 9	191 0

$$* Z = \frac{t - \bar{t}}{s_t} = \frac{t - 96.7}{37.5}$$

The observed and expected number of failures differ from each other in Table 8.2. χ^2 is a measure of the discrepancy between them for the whole distribution, and the procedure for testing the normal theory of error as applied to the first bus motor failure is a variation on the χ^2 test of Sec. 8.1. The details are given in Sec. 8.3.

8.3 Testing the Fit of a Normal Distribution

Usually the population probability curve is not known completely, and certain parameters must therefore be estimated from the sample. For example, in fitting a normal distribution to sample data, the mean and standard deviation are estimated by $\bar{X}$ and s_x of the sample. The measure of the discrepancy between the observed and expected frequencies is still χ^2 as defined by Eq. (8.1). But χ^2 is now distributed according to the χ^2 distribution with $k - c - 1$ *degrees of freedom*, where c is the number of parameters estimated from the sample. In other words, the number of degrees of freedom is reduced by 1 for each constant estimated from the sample. For a normal distribution, the mean and standard deviation are estimated from the sample; $c = 2$ and the number of degrees of freedom is $k - 2 - 1 = k - 3$.

A test of the adequacy of fit of the normal distribution to the first bus motor failures is shown in Table 8.5.

TABLE 8.5. χ^2 TEST FOR GOODNESS OF FIT OF NORMAL DISTRIBUTION TO FIRST BUS MOTOR FAILURES*

Interval	Observed frequency O	Expected frequency E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
Less than 20	6	3 6			
20-40	11 } 17	8 2 } 11 8	5 2	27 04	2 29
40-60	16	18 5	-2 5	6.25	34
60-80	25	31 3	-6 3	39 69	1 27
80-100	34	39 9	-5 9	34 81	87
100-120	46	38 3	7 7	59 29	1 55
120-140	33	27 6	5 4	29 16	1 06
140-160	16	15 0	1 0	1 00	07
160-180	2	6 1	-4 6	21 16	2 46
180 up	2 } 4	2 5 } 8 6			
Total. . .	191	191 0	0		$\chi^2 = 9.91$

* Data from Table 8.2.

The observed and expected frequencies are taken from Table 8.2. The computed value of χ^2 is 9.91.

Remark. In applying the χ^2 test of goodness of fit, each cell or group of the frequency distribution should have an expected number of 5 or more. When the

expected number is less than 5, cells should be lumped together so that their total expected value is greater than 5. Thus, in Table 8 5 (and in Table 8 2), the first and last pairs of cells are lumped together to make their expected values total more than 5.

The statistical test follows:

H_0 : Normal distribution with $\bar{t} = 96.7$ and $s_t = 37.5$ is adequate fit to the first bus motor failure data.

H_1 : Normal distribution is not adequate fit to data.

$\alpha = .05$

R : Upper tail.

$k = 8$ $k - 3 = 5$ degrees of freedom $\chi_{.05}^2(5) = 11$ (Table A.9)

Decision Rule. Accept H_1 if $\chi^2 \geq 11$; otherwise accept H_0 .

Decision. Accept H_0 because $\chi^2 = 9.91$ falls in the region A . The conclusion is that the normal theory of error is adequate to describe the first bus motor failures.

8.4 Another Theory of Error

Table 8.6 gives the frequency distribution of the distance to the *fifth* bus motor failure. The shape of this distribution is by no stretch of the

TABLE 8.6. FIFTH BUS MOTOR FAILURES*

Distance interval, thousands of miles	Observed number of failures	Expected number of failures
0-20	29	36 0
20-40	27	20 7
40-60	14	12 0
60-80	8	6 9
80 up	7	9 4
Total	85	85 0

* SOURCE: J. D. Davis, An Analysis of Some Failure Data, *Journal of the American Statistical Association*, vol. 47, p. 147, June, 1952.

imagination normal. The mode is in the first interval and each succeeding frequency is less than the one preceding it. Davis postulates another theory of error, an *exponential theory*, for failures in many systems. The fifth bus motor failure is one example of it.¹

¹ The expected frequency of the exponential theory of error is given by

$$nf(t) = \left(\frac{n}{\bar{t}}\right) e^{-t/\bar{t}}$$

where t is the distance to failure and $\bar{t}$ is the mean distance to failure. The expected frequencies of Table 8.6 are computed from this equation. Only the results, and not the mathematics of fitting this particular type of distribution, are pertinent here.

A test of the adequacy of the exponential theory to explain the fifth failures of the bus motors is given in Table 8.7 with data taken from Table 8.6. The value of χ^2 is computed as 4.40. Only one constant is estimated from the sample in fitting the exponential distribution—the

TABLE 8.7. χ^2 TEST FOR GOODNESS OF FIT OF EXPONENTIAL DISTRIBUTION TO FIFTH BUS MOTOR FAILURES*

O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
29	36 0	-7 0	49 00	1 36
27	20 7	6 3	39 69	1 92
14	12 0	2 0	4 00	33
8	6 9	1 1	1 21	18
7	9 4	-2 4	5 76	0 61
85	85 0	0 0		4.40

* Data from Table 8.6

arithmetic mean $\bar{l}$. The number of degrees of freedom is therefore $k - 2$. The statistical test follows.

H_0 : Exponential distribution is a good fit to distribution of fifth failures.

H_1 : Exponential is not a good fit.

$\alpha = .05$

R : Upper tail.

$k = 5$ $k - 2 = 3$ D.F. $\chi_{.05}^2(3) = 7.8$ (from Table A.9)

Decision Rule. Accept H_1 if $\chi^2 \geq 7.8$; otherwise accept H_0 .

Decision. Accept H_0 that the exponential distribution is a good fit to fifth bus motor failures because $\chi^2 = 4.40$ falls in the region A . The exponential theory of error is an adequate explanation of the fifth bus motor failures.

8.5 An Example of No Theory of Error

The last data on bus motor failures taken from the article by J. D. Davis are those in Table 8.8 on second failures. This table gives the observed distribution and the expected number of failures for both the normal and the exponential theory of error. The computed values of χ^2 are also given at the bottom of each table. The statistical tests of these theories of error follow.

Normal theory:

H_0 : Normal theory is adequate for second failures.

H_1 : Normal theory is not adequate for second failures.

$\alpha = .05$ $k = 7$ $k - 3 = 4$ D.F. $\chi_{.05}^2(4) = 9.5$

Decision Rule. Accept H_1 if $\chi^2 \geq 9.5$; otherwise accept H_0 .

Decision. Accept H_1 because $\chi^2 = 16.57$ falls in the region R . The normal theory of error is not an adequate explanation of second failures.

Exponential theory:

H_0 : Exponential theory is adequate for second failures.

H_1 : Exponential theory is not adequate for second failures.

$\alpha = .05 \quad k = 8 \quad k - 2 = 6 \text{ D.F.} \quad \chi_{.05}^2(5) = 12.6$

TABLE 8.8. SECOND BUS MOTOR FAILURES*

Distance interval, thousands of miles	Observed number of failures	Expected number of failures† (normal theory)	Expected number of failures (exponential theory)
0-20	19	9.2	25.4
20-40	13	12.9	19.8
40-60	13	16.9	14.5
60-80	15	18.7	11.1
80-100	15	17.2	8.2
100-120	18	13.1	6.0
120-160	7	13.0	8.3
160 up	4	3.0	10.7
Total	104	104.0	104.0
$\chi^2 \dots$	.	16.57	39.53

* SOURCE: J. D. Davis, An Analysis of Some Failure Data, *Journal of the American Statistical Association*, vol. 47, pp. 145-146, June, 1952.

† From truncated normal distribution.

Decision Rule. Accept H_1 if $\chi^2 \geq 12.6$; otherwise accept H_0 .

Decision. Accept H_1 because $\chi^2 = 39.53$ falls in the region R . The exponential theory of error is not an adequate explanation of the second bus motor failures.

The results of the study of bus motor failures, with the accompanying discussion, are quoted directly from the article by Davis.¹

Result. The normal curve appears to approximate the distribution of the first bus motor failures, while the distributions of the third, fourth, and fifth bus motor failures appear to be approximated by the exponential distribution.

The distribution of the second failures of the bus motors is not described by either theoretical curve but looks like a combination of both. [See Fig. 8.2.]

Discussion of Results. Consideration of the circumstances attendant upon the first and subsequent motor failures makes the peculiarities of the distribution curves [Fig. 8.2] appear reasonable. A normal distribution of mileage to first failure can be expected, for the moving parts of the motor slowly abrade each other until, when a sufficient amount of metal is worn away, they either break or

¹ See Davis, *op. cit.*, pp. 144, 147.

no longer perform their function with satisfactory efficiency. This expectation is confirmed by first failure being caused singly or in combination by worn cylinders, pistons, piston rings, valves, camshafts, connecting rod or crankshaft bearings, etc. There were, however, a considerable number of failures in which moving parts fractured at low mileages, indicating manufacturing errors or inadequate or improper maintenance or repair. These latter conditions were exponential type

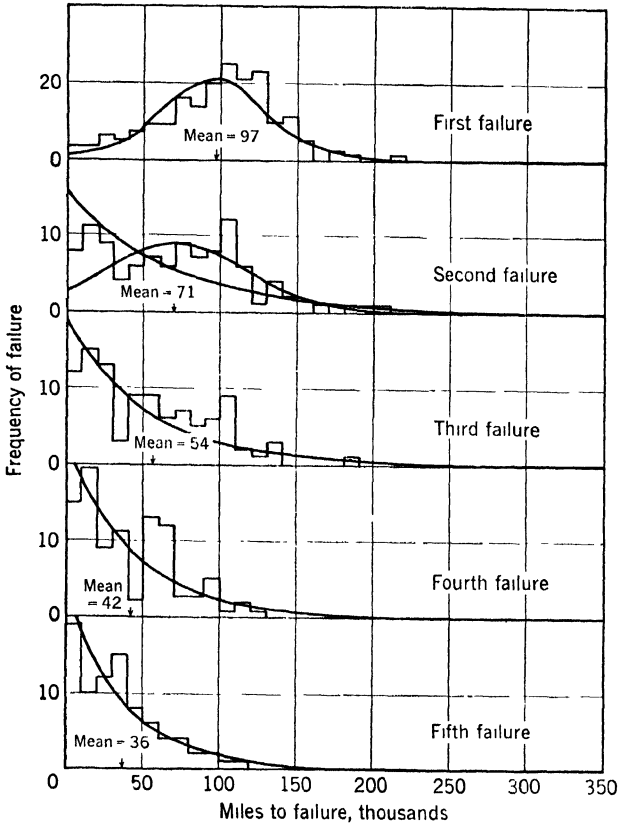


FIG. 8.2. Frequency distributions of bus motor failure with miles driven. (SOURCE: D. J. Davis, *An Analysis of Some Failure Data*, *Journal of the American Statistical Association*, vol. 47, p. 146, June, 1952.)

occurrences which should follow a different distribution. Unfortunately these data on the cause of failure were so fragmentary and incomplete that a segregation of the two types of failure could not be accomplished . . .

An exponential distribution of the third and subsequent failures appears logical in that by the time the motor has been overhauled twice it consists of components in a scattered state of wear. The individual parts may each exhibit a normal distribution of failure with operating time, but with components in random stages of wear a motor has an equally likely chance of failing during any period of opera-

tion. This does not mean that the chance of failure during equal operating periods is the same after each overhaul. The mean mileage to failure after the first overhaul is less than that of first failure, and after each subsequent overhaul, a shorter mean mileage to failure is exhibited.

The second failures are not predominantly due to either exponential or normal theory causes and can therefore be expected to exhibit a distribution characteristic of neither cause.

8.6 Estimating a Population Distribution without Specifying Its Form

The second bus motor failures analyzed in Sec. 8.5 were described inadequately by both the normal and exponential theory of errors. Even when no mathematical form of the distribution may be specified, an inference about the population distribution may still be made. This inference takes the form of estimating a confidence band for the *cumulative relative frequency distribution*.

The frequency distribution must first be cumulated, usually on a *less-than* basis, and then the cumulative frequencies expressed as percentages of the total number of observations. The distribution of second failures

TABLE 8.9 ESTIMATING CUMULATIVE DISTRIBUTION OF SECOND BUS MOTOR FAILURES

(1) Distance interval, thousands of miles	(2) f	(3) Cumulative f	(4) Cumulative per cent	(5) Cumulative per cent (rounded)	(6) 90 per cent confidence limits
0-20	19	19	18.27	18	6-30
20-40	13	32	30.77	31	19-43
40-60	13	45	43.27	43	31-55
60-80	15	60	57.69	58	46-70
80-100	15	75	72.12	72	60-84
100-120	18	93	89.12	89	77-100
120-160	7	100	96.15	96	84-100
160 up	4	104	100.00	100	88-100
Total	104				

is given in Table 8.9. The frequencies are first cumulated on a *less-than* basis from top to bottom in column 3 of the table. Thus:

1. There are 19 failures in less than 20 thousand miles;
2. There are 32 failures in less than 40 thousand miles;
3. There are 45 failures in less than 60 thousand miles, etc.

The cumulative frequencies are then expressed as percentages of the total frequency ($n = 104$). These are given (rounded to the nearest whole number) in column 5. The cumulative relative frequency distribution is read:

1. Eighteen per cent of the failures occurred in less than 20 thousand miles.

2. Thirty-one per cent of the failures occurred in less than 40 thousand miles.

3. Forty-three per cent of the failures occurred in less than 60 thousand miles, etc.

The confidence bands are formed by *adding* and *subtracting* $100d_{c.c}$ from the cumulative relative frequencies of the sample, where $d_{c.c}$ is the deviation between the cumulative distributions of population and sample and C.C. is the probability that the maximum deviation between the cumulative distributions does not exceed $d_{c.c}$. Values of $d_{c.c}$ for different sample sizes and confidence levels are given in Table 8.10.

TABLE 8.10. TABLE OF THE PROBABILITY DISTRIBUTION OF $d_{c.c}$ FOR ESTIMATING A CUMULATIVE FREQUENCY DISTRIBUTION*

Sample size	C.C.†				
	80	.85	.90	95	.99
5	45	17	51	56	67
10	.32	34	37	41	49
20	.23	25	.26	29	36
25	21	22	24	27	32
30	19	20	22	24	29
35	18	19	20	23	27
40	17	18	19	21	25
45	16	17	18	20	24
50	.15	16	17	.19	.23
Larger n	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

* This table is reproduced from Wilfrid J. Dixon and Frank J. Massey, Jr., "Introduction to Statistical Analysis," McGraw-Hill Book Company, Inc., New York, 1951, Table 21, p. 348, with the permission of the authors and the publisher.

† The confidence coefficient (C.C.) is the probability that the maximum deviation between the cumulative distribution of the population and of the sample does not exceed the tabled value of $d_{c.c}$. Linear interpolation may be used for values of n which are not given in the table.

For the second bus motor failures, $n = 104$. With C.C. = .90, $d_{c.c.} = 1.22/\sqrt{n} = 1.22/\sqrt{104} = .1196$, or approximately 12 per cent. The 90 per cent confidence limits are found by adding and subtracting 12 per cent to each of the cumulative percentages in the distribution. The results are shown in column 6 of Table 8.9. The probability is .90 that the band overlaps the true cumulative relative frequency distribu-

tion of second bus motor failures, meaning that if numerous samples of second bus motor failures are taken and confidence bands similarly computed for each sample, 90 per cent of the statements that each band includes the true distribution will be true. Or, in other words, the fraction of the bands that trap the true distribution is 90 per cent. This estimate of the population cumulative relative frequency distribution does not depend upon any specification of the mathematical form of the distribution.

The confidence band obtained for 104 observations of second failures has a width of plus or minus 12 per cent, or 24 per cent. Suppose that it is desired to estimate the cumulative relative frequency distribution of second failures to within plus or minus 1 per cent at the 99 per cent confidence level. How large a sample is necessary? Table 8.10 provides a rough estimate of the sample size necessary to meet these specifications. $100d_{c.c.} = 1$ per cent; $d_{c.c.} = .01$. With C.C. = .99, from Table 8.10, $d_{c.c.} = 1.63/\sqrt{n}$. Therefore,

$$\frac{1.63}{\sqrt{n}} = .01 \quad n = \left(\frac{1.63}{.01}\right)^2 = 163^2 = 26,569$$

Approximately 25,000 observations are necessary to estimate the cumulative relative frequency distribution to within plus or minus 1 per cent at the 99 per cent confidence level. If the estimate must be within plus or minus 5 per cent with 90 per cent confidence, $100d_{c.c.} = 5$ per cent; $d_{c.c.} = .05$. With C.C. = .90, from Table 8.10,

$$\frac{1.22}{\sqrt{n}} = .05 \quad n = 595$$

Without specifying the mathematical form of the distribution, an experiment may be planned to estimate its cumulative relative frequency distribution with any desired accuracy and confidence coefficient. This is a very useful procedure when all attempts to write mathematical functions for empirical distributions fail.

EXERCISES

8.4. Lifetimes in hours of four hundred seventeen 40-watt 110-volt internally frosted incandescent lamps taken from 42 weekly quality-control forced-life test samples are given on page 141 of the article by Davis. (See Sec. 8.2 for the complete reference.) Make a frequency distribution of these data using 25-hour class intervals. Plot a histogram of this distribution. Do the lifetimes appear normally distributed?

Find the mean and standard deviation of this distribution. Make another distribution with 6 class intervals by expressing the lifetimes in hours as devia-

tions from the mean life in standard deviations. The first interval will be less than -2 , the next from -2 to -1 , then -1 to 0 , etc., to more than $+2$. Count the observed number of frequencies in these intervals by fitting a normal distribution to the data. Test the goodness of fit of the normal distribution with the χ^2 test.

8.5. Davis reports an investigation of student typing errors for nine typists during definite intervals throughout one week. The mean number of strokes per error for each typist defines the basic interval of inspection for errors. The observed distribution of the number of errors per interval is compared with the expected theoretical frequency determined from an exponential theory of error. Test the adequacy of this theory.

Number of errors per interval	Observed frequency	Expected frequency
0	135	135 0
1	140	135 0
2	58	67 5
3	29	22 5
4 or more	5	7 0
Total .	367	367 0

8.6. The number of aircraft accidents at a given training base are given for each day of the week. Test to see if accidents are uniformly distributed over the days of the week.

Sunday.		12
Monday		14
Tuesday.		6
Wednesday..		10
Thursday		9
Friday..		7
Saturday .		12
Total. .		70

Lump together Saturday, Sunday, and Monday, and test to see if social activities of the week end affect accident rates.

CHAPTER 9

STATISTICAL INFERENCE: THE ARITHMETIC MEAN AND RELATED PROBLEMS

Example 9.1. You are the purchasing agent for a machine shop which buys iron castings from a foundry. The crucial specification is weight, for a casting that is too heavy will not lend itself easily to further processing. From its detailed statistical records, the foundry reports that the weights of its castings are almost normally distributed with a mean of 20.0 pounds and a standard deviation of 1.0 pound. From every purchase of several hundred castings, you take a small sample and check the weights against the average reported by the foundry. You find that the weights of a sample of 10 are 19.8, 20.3, 20.6, 21.1, 19.3, 19.6, 20.1, 20.8, 21.1, 21.3. As the purchasing agent, would you accept the foundry's claim that the average weight of castings is 20.0 pounds? What average weight of castings will you allow and still accept the foundry's claim?

Example 9.2. The quality-control department of a food-manufacturing corporation is given the responsibility of seeing that 12-ounce cans of soup are neither overweight nor underweight. The filling machines can be adjusted within narrow limits to regulate the amount of soup poured into each can. The inspector for this operation follows a standardized procedure in his quality determination: (1) he selects 10 cans of soup randomly from 1 hour's production of one filling machine; (2) he weighs each of these cans on a scale and records the individual weights; (3) he finds the average weight of the 10 cans; (4) he decides to accept or reject the hour's production by comparing the average of the 10 cans with a standard weight. (If he rejects the hour's production, all the cans filled during that hour must be weighed individually and accepted or rejected individually.)

EXERCISE

9.1. Judge: Will you accept the hour's production if the 10 cans of soup have an average weight of 11 ounces? 11.5 ounces? 12 ounces? Will you accept it if the average is 13 ounces? 12.5 ounces? 12 ounces?

Planner: What average weight will you allow for the 10 cans and still accept them as not underweight? 11 ounces? 11.5 ounces? 12 ounces? What average weight will you allow and still accept the cans as not overweight? 13 ounces? 12.5 ounces? 12 ounces?

The correct numbers for the statistical decision rules in these two problems are not 20 pounds and 12 ounces. The correct numbers must allow for sampling variability in the arithmetic mean. All the ideas necessary to the solutions of Examples 9.1 and 9.2 have been presented in the first three chapters; the only new material required is the appropriate probability distribution of the arithmetic mean.

9.1 General Description of the Decision Problem

The basic concepts of decision problems are reviewed here in terms of Example 9.1. The *elementary unit* is a casting. The *characteristic* of the elementary unit is its weight. An *observation* of this characteristic is the number of pounds an individual casting weighs. The *population* is the distribution of the weights of the individual castings poured by the foundry and sent out in the shipment to the machine shop. The *sample* is 10 castings selected randomly from the population. The *random variable* is the *arithmetic mean* weight of the sample castings in pounds.

The solution to the problem, i.e., the decision rule, depends upon the *probability distribution*, or the *sampling distribution*, of the mean $\bar{X}$. Once the sampling distribution is known, the decision process consists of setting up a hypothesis H_0 and an alternative hypothesis H_1 , deciding upon a value for the probability of accepting H_1 when H_0 is correct (α), and selecting a region R (R leads to the acceptance of H_1) which meets the criterion of α and minimizes β , the probability of accepting H_0 when H_1 is correct. Although the mechanical details differ from those of the binomial problems in Chaps. 3 to 5, the basic ideas are the same.

EXERCISE

9.2. Answer these questions in terms of Example 9.1.

- What is the hypothesis to be tested? (H_0)
- What is the alternative hypothesis? (H_1)
- What action is to be taken if H_0 is accepted? If H_1 is accepted?
- What is the meaning of the α error? Of the β error?
- What are the consequences of these errors?
- Which one of these errors is of most concern to the foundry? To the machine-shop purchasing agent?
- What is the meaning of alpha? Of beta?
- What does it mean to fix alpha at some preassigned value?
- What is the sample space?
- Where will the region R (which leads to the acceptance of H_1) be placed in the sample space? Why?

9.2 An Example of a Sampling Distribution of the Arithmetic Mean

The decision rule for Example 9.1 (and 9.2) depends upon a knowledge of the probability distribution or the sampling distribution of the arith-

metic mean. In a very simple case, the sampling distribution of $\bar{X}$ can be obtained by *enumerating all possible samples* and making a distribution of the means of these samples. Suppose that the population consists of four values of a variable x : 1, 3, 5, and 7. The mean of this population is $\mu_x = 4$; the variance of this population is

$$\sigma_x^2 = \frac{\Sigma(x - \mu)^2}{N} = \frac{(1 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 + (7 - 4)^2}{4} = \frac{20}{4} = 5$$

and the standard deviation of the population is

$$\sigma_x = \sqrt{5} = 2.24$$

All possible samples of 2 observations ($n = 2$) selected with replacement from this population are enumerated and the results given in Table 9.1.

TABLE 9.1. ENUMERATION OF ALL POSSIBLE SAMPLES OF $n = 2$ SELECTED WITH REPLACEMENT FROM A POPULATION CONSISTING OF THE FOUR NUMBERS: 1, 3, 5, 7

Sample values	Sum Σx	Mean $\bar{X}$	Sample values	Sum Σx	Mean $\bar{X}$
1, 1	2	1	5, 1	6	3
1, 3	4	2	5, 3	8	4
1, 5	6	3	5, 5	10	5
1, 7	8	4	5, 7	12	6
3, 1	4	2	7, 1	8	4
3, 3	6	3	7, 3	10	5
3, 5	8	4	7, 5	12	6
3, 7	10	5	7, 7	14	7

TABLE 9.2. PROBABILITY DISTRIBUTION OF $\bar{X}$

Mean $\bar{X}$	Samples f	Probability of $\bar{X}$ $\Pr(\bar{X}) = f/N$
1	1	1/16
2	2	2/16
3	3	3/16
4	4	4/16
5	3	3/16
6	2	2/16
7	1	1/16
Total	16	1.00

Sixteen different samples are possible. The probability distribution of the means of these 16 samples is given in Table 9.2. Thus, the probability is 1 in 16 that the sample mean of a randomly selected sample of

2 observations will be 1. The probability is 1 in 4 that the sample mean will be 4. Other probabilities are readily obtained from Table 9.2.

The probability distribution of the mean has a mean, a variance, and a standard deviation. Since $\bar{X}$ is the symbol for the mean, it is attached as a subscript to each of the constants or parameters of the probability distribution of the mean. The basic computations are given in Table 9.3.

$$\begin{aligned}\mu_{\bar{X}} &= \Sigma \text{Pr}(\bar{X}) \cdot \bar{X} = 4 \\ \sigma_{\bar{X}}^2 &= \Sigma \text{Pr}(\bar{X}) \cdot (\bar{X} - \mu_{\bar{X}})^2 = 2.5 \\ \sigma_{\bar{X}} &= \sqrt{2.5} = 1.58\end{aligned}$$

TABLE 9.3. MEAN, VARIANCE, AND STANDARD DEVIATION OF THE PROBABILITY DISTRIBUTION OF $\bar{X}$

$\bar{X}$	$\text{Pr}(\bar{X})$	$\text{Pr}(\bar{X}) \cdot \bar{X}$	$(\bar{X} - \mu_{\bar{X}})$	$(\bar{X} - \mu_{\bar{X}})^2$	$\text{Pr}(\bar{X}) \cdot (\bar{X} - \mu_{\bar{X}})^2$
1	1/16	1/16	-3	9	9/16
2	2/16	4/16	-2	4	8/16
3	3/16	9/16	-1	1	3/16
4	4/16	16/16	0	0	0
5	3/16	15/16	+1	1	3/16
6	2/16	12/16	+2	4	8/16
7	1/16	7/16	+3	9	9/16
Total ..	1 00	64/16 = 4		.	$\sigma_{\bar{X}}^2 = 40/16 = 2.5$

$$\mu_{\bar{X}} = \Sigma \text{Pr}(\bar{X}) \cdot \bar{X} = 4 \quad \sigma_{\bar{X}}^2 = \Sigma \text{Pr}(\bar{X}) \cdot (\bar{X} - \mu_{\bar{X}})^2 = 2.5$$

The mean, variance, and standard deviation of the probability distribution of the mean are related to the mean, variance, and standard deviation of the population. The mean of the probability distribution of the mean equals the mean of the population. This statement is expressed by

$$\mu_{\bar{X}} = \mu_x \quad (9.1)$$

The variance of the sampling distribution of the mean is the variance of the population divided by the sample size, or

$$\sigma_{\bar{X}}^2 = \frac{\sigma_x^2}{n} \quad (9.2)$$

The standard deviation of the sampling distribution of the mean is the standard deviation of the population divided by the square root of the sample size, or

$$\sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}} \quad (9.3)$$

These relationships are verified by the experiment just performed with $n = 2$.

$$\begin{aligned}\mu_x &= 4 & \mu_{\bar{x}} &= 4 \\ \sigma_x^2 &= 5 & \sigma_{\bar{x}}^2 &= 2.5 \\ \sigma_x &= \sqrt{5} & \sigma_{\bar{x}} &= \sqrt{2.5}\end{aligned}$$

By Eqs. (9.1 to 9.3),

$$\begin{aligned}\mu_{\bar{x}} &= \mu_x = 4 \\ \sigma_{\bar{x}}^2 &= \frac{\sigma_x^2}{n} = \frac{5}{2} = 2.5 \\ \sigma_{\bar{x}} &= \frac{\sigma_x}{\sqrt{n}} = \frac{\sqrt{5}}{\sqrt{2}} = \sqrt{\frac{5}{2}} = \sqrt{2.5} = 1.58\end{aligned}$$

The relationships expressed by Eqs. (9.1) to (9.3) are true for the sampling distribution of the arithmetic mean from any population when the sampling is with replacement.

Remark. The sampling distribution of the mean when the sampling is *without replacement* has a variance and standard deviation which differ from those given in Eqs. (9.2) and (9.3) when the sampling is with replacement. The mean of the sampling distribution of the mean still equals the population mean [Eq. (9.1)]. The variance and standard deviation are given by

$$\sigma_{\bar{x}}^2 = \frac{N-n}{N-1} \frac{\sigma_x^2}{n} \quad \sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \frac{\sigma_x}{\sqrt{n}}$$

N is the size of the population and the other symbols retain their same meaning. If the population is many times larger than the sample, the factor which allows for sampling without replacement is negligible

In practical problems the populations are usually so large that complete enumeration of all possible samples is impossible and the probability distribution must be derived by mathematical methods. Knowledge of the mean, variance, and standard deviation of the sampling distribution of $\bar{X}$ is supplemented by a *mathematical equation* for the *form* of the probability distribution. Probabilities are then computed from this equation and tabled for easy reading. In general, the mathematical equation or form depends upon the kind of population sampled. Without showing derivations, the following sections give appropriate sampling distributions of the mean when various assumptions about the population are made.

EXERCISES

9.3. The incomes of a hypothetical population of 10 individuals are given below. For this population find μ_x , σ_x^2 , and σ_x . Enumerate all possible samples of $n = 3$, sampling with replacement. Compute $\bar{X}$ for each sample, and make a probability distribution of $\bar{X}$. Find $\mu_{\bar{x}}$, $\sigma_{\bar{x}}^2$, and $\sigma_{\bar{x}}$.

Individual	Income	Individual	Income	Individual	Income
A	\$ 900	E	\$2200	H	\$3100
B	1500	F	3600	I	5900
C	2700	G	2000	J	1300
D	1800				

9.4. Given a population with $\mu_x = 10$ and $\sigma_x = 2$. If samples of 100 are selected at random with replacement from this population, what are $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$?

9.5. Based upon $n = 100$, $\sigma_{\bar{X}} = 10$. What sample size is necessary to reduce $\sigma_{\bar{X}}$ to 2.5? To 5?

9.6. Under what circumstances will $\sigma_{\bar{X}} = \sigma_x$? = 0?

9.3 Sampling a Known Normal Population

Under the first set of assumptions for studying the sampling distribution of the mean, the population is a *normal distribution with a known mean and standard deviation* (or variance). In Example 9.1, the foundry reports that the population of weights of individual castings is approximately normally distributed with a mean weight $\mu_x = 20.0$ pounds and a standard deviation $\sigma_x = 1.0$ pound. Given this set of assumptions, what is the probability distribution of the mean or, in other words, what is the sampling distribution of $\bar{X}$ in samples of n observations selected at random from this population? It is impossible to enumerate all samples because the population consists of a large number of castings, but the sampling distribution has been derived mathematically with the results given below.

Equations (9.1) to (9.3) still hold for these assumptions. The sampling distribution of the mean $\bar{X}$ of samples of n observations selected randomly from a normal population with mean μ_x and standard deviation σ_x has the following properties:

1. Has a mean equal to the mean of the population: $\mu_{\bar{X}} = \mu_x$.
2. Has a standard deviation equal to the standard deviation of the population divided by the square root of the sample size: $\sigma_{\bar{X}} = \sigma_x / \sqrt{n}$.
3. Is *normally* distributed when the population is normally distributed.

The relationship between the population of x values and the sampling distribution of $\bar{X}$ is shown in Fig. 9.1. Both distributions are normal and both have the same mean. The variation in the distribution of means is less than the variation in the population from which the samples were selected, as shown by $\sigma_{\bar{X}} = \sigma_x / \sqrt{n}$.

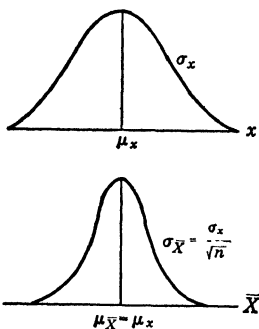


FIG. 9.1. Relationships between the population distribution of x values and the probability distribution of the mean $\bar{X}$.

Since the distribution of the means is normal, areas and probabilities associated with different values of $\bar{X}$ are found from the table of normal areas (Table A.3). The relationship between Z and a random variable which is normally distributed is generally of the form $Z = (\text{random variable} - \text{mean}) / \text{standard deviation}$. The random variable is the mean $\bar{X}$, and

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad (9.4)$$

where $\mu_{\bar{X}} = \mu_x$ and $\sigma_{\bar{X}} = \sigma_x / \sqrt{n}$. The use of a table of normal areas (Table A.3) to compute probabilities has been previously discussed in detail.

EXERCISES

9.7. Review Exercise 1.6. In this exercise samples of $n = 10$ observations were drawn at random from a population which is approximately a normal distribution with a mean $\mu_x = 3.0$ and a standard deviation $\sigma_x = 1.0$. Review Exercise 1.7 by comparing the mean of the distribution of sample means with the population mean, the variability of the means with the variability in the population, and the form of the distribution of the means with the form of the population.

9.8. In Example 9.1, the distribution of individual casting weights is approximately normal with $\mu_x = 20.0$ and $\sigma_x = 1.0$.

a. What is the probability that an individual casting will have a weight between 19.0 and 21.0? Between 18.0 and 22.0? Between 17.0 and 23.0?

b. Within what weights will about 99.7 per cent of all castings lie?

9.9. The distribution of means of random samples of 10 observations from the population of castings will be normal with a mean $\mu_{\bar{X}} = 20.0$ and a standard deviation $\sigma_{\bar{X}} = \sigma_x / \sqrt{n}$, or $1.0 / \sqrt{10} = .316$.

a. What is the probability that the mean weight of a random sample of 10 castings will be between 19.0 and 21.0?

b. Within what weights will approximately 99.7 per cent of the mean weights of a large number of random samples of 10 observations lie?

c. How do the answers to (a) and (b) compare to the comparable parts of Exercise 9.8?

9.4 Decision Rule for Foundry Castings

The decision process in Example 9.1 is formally set out below.

$H_0: \mu_0 = 20.0$ pounds (foundry's claim)

$H_1: \mu_1 > 20.0$ pounds (overweight castings for machine shop)

$n = 10$ $\bar{X}$ (random variable)

$\alpha = .05$ $R: \bar{X} \geq \bar{X}_r$ (upper tail region for accepting H_1)

The value of Z which cuts off .05 in the upper tail of a normal distribu-

tion is 1.645. The region of rejection in the $\bar{X}$ distribution is found from the relationship

$$Z = \frac{\bar{X}_r - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

These regions are shown in Fig. 9.2.

$$\begin{aligned} \mu_x = \mu_{\bar{X}} = 20.0 \quad \sigma_x = 1.0 \quad \sigma_{\bar{X}} &= \frac{\sigma_x}{\sqrt{n}} = \frac{1.0}{\sqrt{10}} = .316 \\ 1.645 &= \frac{\bar{X}_r - 20}{.316} \quad \bar{X}_r = 20.52 \end{aligned}$$

Decision Rule. Accept H_1 , and do not buy the shipment if $\bar{X} \geq 20.52$; otherwise accept H_0 , and buy the shipment.

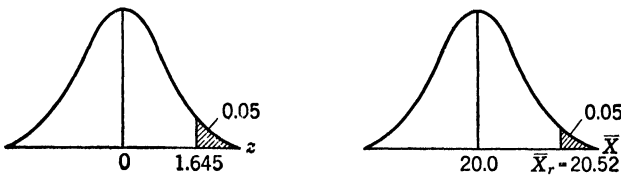


FIG. 9.2. Regions of rejection in the z and $\bar{X}$ distributions.

Decision. Accept H_0 because 20.4 is less than 20.52. The purchasing agent will buy the castings because he has no reason to believe that their mean weight is more than 20.0 pounds.

Operating-characteristic Curve. With the statistical decision rule to buy the shipment of castings if the average weight of 10 castings is 20.52 or less, the probability of buying depends upon the true average weight of the castings in the population. This average is unknown, but if it is more than 20.0, the correct decision is not to buy the shipment. The probability that the shipment will be bought if μ_1 is greater than 20.0 pounds depends upon what that value of μ_1 happens to be. Values of β , the probabilities of buying the shipment for various values of μ_1 , form the *operating-characteristic curve*. The OC curve for the present statistical test is given in Table 9.4 and Fig. 9.3.

The OC curve is most important in tests of statistical hypotheses. A foundry sells iron castings with a specific claim about their mean weight. A scientific attitude dictates a test of the claim by selecting a random sample, weighing the sample castings, and deciding to buy a particular shipment only after observing the results of the test. The actual average weight is unknown. What if it should be 20.5 pounds? With a rule to accept H_0 and to buy the shipment if the average of 10 castings is less than 20.52 pounds, what is the probability of deciding to buy the shipment? From the OC curve of Table 9.4, this probability is approxi-

mately .5. If the average weight is as much as 21.0 pounds, the probability of buying the shipment is .06. The graph dramatically shows that the probability of buying the shipment becomes smaller as the true,

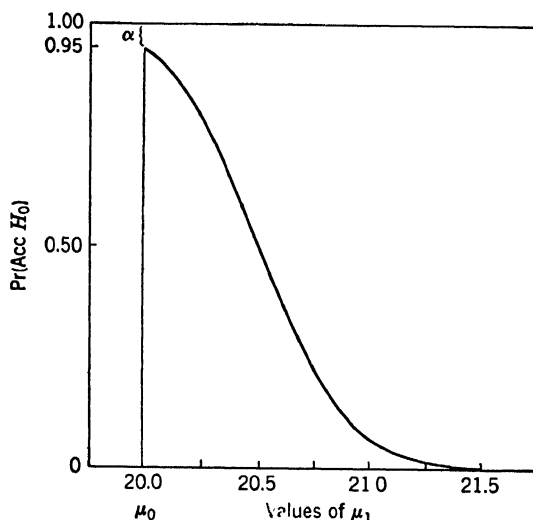


FIG. 9.3. Operating-characteristic curve of the foundry-casting problem.

but unknown, average weight of castings becomes larger. Finally, if the average is 21.5 pounds, the probability of buying the shipment is very small (.001).

TABLE 9.4. OPERATING-CHARACTERISTIC CURVE OF THE FOUNDRY-CASTING PROBLEM

$H_0: \mu_0 = 20$		$\alpha = .05$
$H_1: \mu_1 > 20$		$R: \bar{X} \geq 20.52$
$n = 10$		$\sigma_x = 1, \sigma_{\bar{X}} = .316$

μ	$Z = \frac{20.52 - \mu}{.316}$	$\beta = \Pr(\text{Acc } H_0 \mu)$ $\beta = \text{area below } Z$
20.0 (= μ_0)	+1.645	0.95 = $1 - \alpha$
20.2	+1.0	0.84
20.5	+0.06	0.52
21.0	-1.52	0.06
21.5	-3.10	0.001

The OC curve is the plan of operation stressed in Chap. 3. It is not enough to make a rule based upon a claim and a more or less arbitrary value of α . It is true that β will be minimized in such a procedure, but "minimum" is not synonymous with "small." The probability of deciding to accept H_0 when H_1 is the correct hypothesis is known only through

the OC curve. The plan of operation should cover a large number of alternative possibilities for the value of μ . The OC curve does just this.

It should again be emphasized that statistical tests are probability tests. The region in the sample space which leads to the acceptance of H_1 depends upon the arbitrarily selected value of α —the probability of accepting H_1 when H_0 is true. To this extent, then, the decision rule is arbitrary. Two different individuals—for instance, the sales manager of the foundry and the purchasing agent of the machine shop—may very well arrive at two different decision rules because they evaluate differently the seriousness of this error. This is not a problem for the statistician, who can simply point out that statistical tests hinge upon these probabilities and ask for a selection of α according to individual judgment.

It must also be emphasized again that *the work of practical statistics is not the acceptance of hypotheses*, despite the fact that this textbook presents the subject in terms of accepting H_0 or H_1 . Statements about hypotheses can readily be translated into statements about actions to be taken in a wide variety of situations. Hence, the real emphasis in statistics is placed upon statements about *hypotheses as guides to particular actions*.

Numerical Computation of Operating-characteristic Curve. The operating-characteristic curve shows the probability of accepting H_0 for various values of μ consistent with the alternative hypothesis. This is also β , the probability of making an error of the second kind. In general,

$$\beta = \Pr(\text{Acc } H_0 | H_1 \text{ true})$$

In the foundry-castings problem, the hypothesis H_1 specifies values of $\mu_1 > \mu_0$. Consequently, the region R is placed in the upper tail of the distribution and is of the form $R: \bar{X} \geq \bar{X}_r$. The hypothesis H_0 is accepted only if the value of $\bar{X}$ is less than $\bar{X}_r$. The probability β may then be written in general terms as

$$\beta = \Pr(\bar{X} < \bar{X}_r | \mu_1)$$

which, in the iron-castings problem, is

$$\beta = \Pr(\bar{X} < 20.52 | \mu_1)$$

The probability associated with a given value of μ_1 may be read from a table of normal areas (Table A.3) by solving for Z in the equation

$$Z = \frac{\bar{X}_r - \mu_1}{\sigma_{\bar{X}}} = \frac{20.52 - \mu_1}{.316}$$

and finding the area *less than* Z .

Example: Let $\mu_1 = 20.5$.

$$Z = \frac{20.52 - 20.5}{.316} = +.06 \quad \Pr(Z < +.06) = .52$$

If $\mu_1 = 21.0$,

$$Z = \frac{20.52 - 21.0}{.316} = -1.52 \quad \Pr(Z < -1.52) = .06$$

Other values are shown in Table 9.4.

Remark. In the computation of the OC curve, the value $\bar{X}_r = 20.52$ is kept constant and the value of μ_1 is varied.

Sample Size in Testing Hypotheses. The OC curve of the statistical test in the iron-castings problem shows that if the true average weight of the castings is 20.0 pounds, the probability of accepting H_0 and buying the shipment of castings is .95 ($= 1 - \alpha$). If the average weight is 20.5 pounds, the probability of buying the shipment is .52. Although the β probabilities have been minimized, they are still rather large. However, they may be reduced by taking a larger sample. How large a sample should be selected?

The number of observations depends upon the specification of four quantities: μ_0 and α , and μ_1 and β . For example, let the specifications of a proposed sampling plan for the iron castings be:

$$\begin{array}{ll} \mu_0 = 20.0 & \alpha = .05 \\ \mu_1 = 20.5 & \beta = .01 \end{array}$$

These four quantities have the following meaning: If the average weight of castings in the shipment sent out by the foundry is 20.0 pounds, the probability of accepting H_1 and not buying the shipment is controlled at .05. If the average weight is 20.5 pounds, the probability of accepting H_0 and buying the shipment is fixed at .01. Both α and β are fixed at small values. The test specifications are two points on an OC curve, as shown in Fig. 9.4. The sample size necessary to make the OC curve pass through these two points is derived from the basic definitions of α and β .

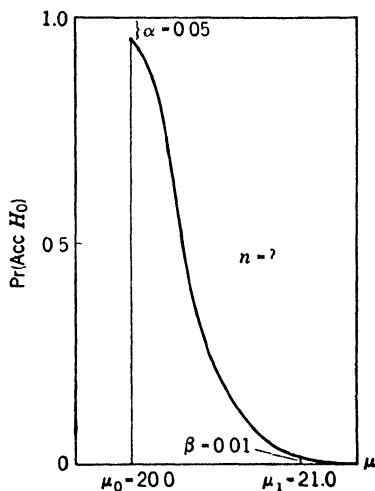


FIG. 9.4. Two points on the OC curve to determine sample size.

Remark. The derivation given here is for a one-tailed test with $\mu_1 > \mu_0$.

$$\alpha = \Pr(\bar{X} \geq \bar{X}_r | \mu_0) \quad \text{and} \quad \beta = \Pr(\bar{X} < \bar{X}_r | \mu_1)$$

Since $\bar{X}$ is normally distributed, the values of Z_α and Z_β are

$$Z_\alpha = \frac{\bar{X}_r - \mu_0}{\sigma_{\bar{X}}} \quad \text{and} \quad -Z_\beta = \frac{\bar{X}_r - \mu_1}{\sigma_{\bar{X}}}$$

But, $\sigma_{\bar{X}} = \sigma_x / \sqrt{n}$. Therefore,

$$Z_\alpha = \frac{\bar{X}_r - \mu_0}{\sigma_x / \sqrt{n}} \quad \text{and} \quad -Z_\beta = \frac{\bar{X}_r - \mu_1}{\sigma_x / \sqrt{n}}$$

These are two equations in two unknowns: n and $\bar{X}_r$. The solution of the two equations follows:

$$\sqrt{n} (\bar{X}_r - \mu_0) = Z_\alpha \cdot \sigma_x \quad (9.5)$$

$$\sqrt{n} (\bar{X}_r - \mu_1) = -Z_\beta \cdot \sigma_x \quad (9.6)$$

Subtracting Eq. (9.6) from Eq. (9.5),

$$\sqrt{n} (\mu_1 - \mu_0) = \sigma_x (Z_\alpha + Z_\beta) \quad (9.7)$$

$$\begin{aligned} \sqrt{n} &= \frac{\sigma_x (Z_\alpha + Z_\beta)}{\mu_1 - \mu_0} \\ n &= \left[\frac{\sigma_x (Z_\alpha + Z_\beta)}{\mu_1 - \mu_0} \right]^2 \end{aligned} \quad (9.8)$$

The value of $\bar{X}_r$ is found by substituting the value of n in either Eq. (9.5) or Eq. (9.6).

Remark 1. Z_α and Z_β are always positive in Eq. (9.8) as derived.

Remark 2. If the alternative hypothesis is $H_1: \mu_1 < \mu_0$, the denominator of the formula for n becomes $\mu_0 - \mu_1$, and the signs of Z_α and Z_β are changed in Eq. (9.5) and Eq. (9.6).

Remark 3. In a two-tailed test, n is first determined for $\mu_1 > \mu_0$ and then for $\mu_1 < \mu_0$. If the two values of n differ, either the average or the larger of the two may be used.

For the proposed test,

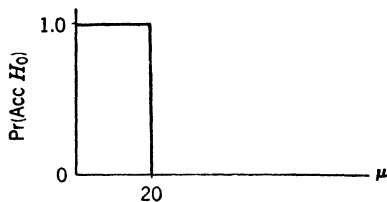
$$\begin{aligned} \mu_0 &= 20.0 & \alpha &= .05 & Z_\alpha &= 1.645 & \sigma_x &= 1.0 \\ \mu_1 &= 20.5 & \beta &= .01 & Z_\beta &= 2.33 \\ n &= \left[\frac{1.0(1.645 + 2.33)}{(20.5 - 20.0)} \right]^2 = 63 \end{aligned}$$

With $n = 63$ and the region of rejection in the upper tail, $\bar{X}_r$ is determined from Eq. (9.5).

$$\sqrt{63} (\bar{X}_r - 20.0) = 1.645(1) \quad \bar{X}_r = 20.21$$

Decision Rule. Accept H_1 and do not buy the shipment of castings if the average weight of a random sample of 63 castings is more than 20.2 pounds; otherwise accept H_0 and buy the shipment.

The ideal OC curve for the statistical test of the average weight of iron castings is shaped like this:



For all values of μ greater than 20.0, the probability of accepting H_0 and buying the shipment is zero; for all values of μ equal to or less than 20.0, the probability of accepting H_0 and buying the shipment is 1. Theoretically, this OC curve exists only when the whole population is examined. The larger the sample size, the sharper is the discrimination between acceptable and unacceptable values of μ .

One of the determinants for choosing the proper sample size for testing hypotheses is a value of μ_1 under the alternative hypothesis. The claim that the mean weight of castings is $\mu_0 = 20.0$ implies that any mean weight more than 20.0 is unacceptable to the machine shop. As a *practical matter*, what mean weight of castings is on the border line between a good and a bad casting? In other words, what is the smallest mean weight at which almost none of the shipments will be bought by the machine shop? Is the weight of the castings so critical that a very small probability of accepting a shipment with a mean weight of 20.01 pounds will be desired? If not, is the weight so critical that the probability of accepting a shipment with a mean weight of 20.1 pounds should be small? To repeat, what is the smallest mean weight at which almost none of the shipments will be bought by the machine shop, or at which the probability of buying the shipment is small? This is a *practical matter*. A statistical sampling plan can be devised for any values of μ_1 , however close to 20.0 or μ_0 , although it may be very expensive because it requires a large sample. The possibility of formulating such a plan is no reason to insist that the unacceptable lower limit for μ_1 should be very close to μ_0 . This value is selected on practical operating rather than statistical grounds. In the present example, the unacceptable value for the mean weight is set at 20.5, and the probability of accepting or buying a shipment with a mean weight that high is set at only .01.

Shipments with mean weights between 20.0 and 20.5 are bought or returned with probabilities higher than α or β . The foundry is more or

less indifferent to which action is taken. In fact, for one average weight, 20.21 (the value of $\bar{X}_r$), the decision is a tossup, i.e., the probability of accepting the shipment is .5 and the probability of rejecting it is .5. For shipments which are definitely good ($\mu \leq \mu_0$), the probability of not buying is at most α ; for shipments which are definitely bad ($\mu \geq \mu_1$), the probability of buying is no more than β . The probabilities of making errors in statistical decision rules cannot be made equal to zero, but they can be controlled at small values through the proper choice of a sample size.

How to Judge Sample Results. To this point, the statistical decision rule for accepting or rejecting castings made by the foundry is a plan of operation set up before the sample is selected. This rule decides which values of the random variable will lead to acceptance of H_0 and which will lead to acceptance of H_1 . This section will explore the problem of judging sample results after they have occurred. The two methods are equivalent, but the mechanical details differ slightly.

The arithmetic mean is normally distributed when the population is a known normal distribution. The first step in selecting the region R —those values of $\bar{X}$ which accept H_1 —is to find the appropriate value of Z in the standard normal distribution $f(Z)$. This region is placed in one or both tails, depending upon the hypothesis H_1 , and the value or values of Z are selected according to the arbitrary probability α . Thus, in a one-tailed test with all of the region R in the upper tail ($H_1: \mu_1 > \mu_0$), and with $\alpha = .05$, the appropriate value is $Z_\alpha = 1.645$. The region $R: \bar{X} \geq \bar{X}_r$ is determined from the relationship between $\bar{X}_r$ and Z ; namely,

$$Z_\alpha = \frac{\bar{X}_r - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

This relationship is shown in Fig. 9.5a.

The characteristic of the sample can also be defined in terms of the variable Z and the decision rule stated in terms of Z . The range of values of Z is divided into two parts, R and A ; the decision is to accept H_1 if the sample Z falls in R , and to accept H_0 if it falls in A . In the example given above, $Z = 1.645$ is the dividing line between R and A . The decision rule is stated: If $Z \geq 1.645$, accept H_1 ; otherwise accept H_0 . The sample value of Z is found by inserting the sample value of $\bar{X}$ in the equation

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

The location of the regions R and A in the Z distribution is the same as in the distribution of $\bar{X}$ for the same alternative hypothesis H_1 . These

regions are shown in Fig. 9.5*b*. The value of Z which divides R and A is labeled Z_α to correspond to the probability level of α . That is to say,

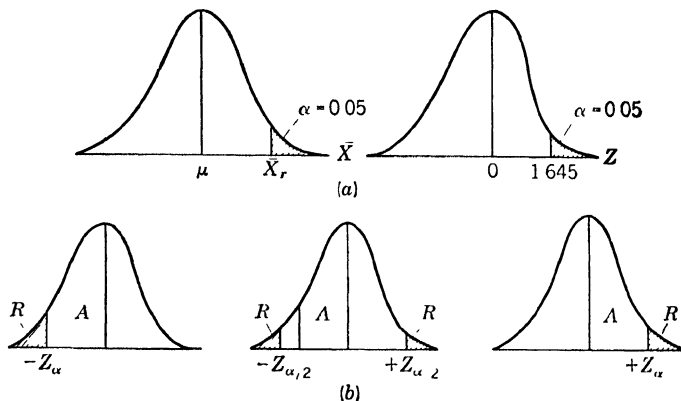


FIG. 9.5. Decision rules stated in terms of Z : (a) Relationship between Z and $\bar{X}_r$. (b) Regions R and A in the Z distribution

Z_α is the value of Z which cuts off the probability α (or $\alpha/2$) in the appropriate tail of the normal distribution. The three statistical decision rules for the three alternative hypotheses are:

1. $\mu_1 < \mu_0$: Accept H_1 if $Z \leq Z_\alpha$; otherwise accept H_0 .
2. $\mu_1 > \mu_0$: Accept H_1 if $Z \geq Z_\alpha$; otherwise accept H_0 .
3. $\mu_1 \neq \mu_0$: Accept H_1 if $Z \geq Z_{\alpha/2}$ or $Z \leq -Z_{\alpha/2}$; otherwise accept H_0 .

Example: In the problem of the weight of the foundry castings,

$$\begin{aligned}
 H_0: \mu_0 &= 20.0 & \alpha &= .05 & Z_\alpha &= 1.645 \\
 H_1: \mu_1 &> \mu_0 & n &= 10 & \bar{X} &= 20.4 \text{ (observed)} \\
 \sigma_x &= 1.0 & \sigma_{\bar{x}} &= \frac{1.0}{\sqrt{10}} = .316
 \end{aligned}$$

Decision Rule. Accept H_1 if the sample Z is equal to or greater than 1.645; otherwise accept H_0 .

From the sample, $Z = (20.4 - 20.0)/.316 = 1.27$.

Decision. Accept H_0 and buy the shipment of castings because $Z = 1.27$ falls in the region A .

This is the same decision reached in the previous analysis; it should be the same because the methods are equivalent. All the ideas on planning a statistical decision rule before a sample is taken are applicable to the formulation of the proper decision rule for judging a sample after it has been selected.

Remark. The literature of statistics often refers to this method of judging sample results as the *critical ratio*.

EXERCISES

9.10. Suppose that the weights of the cans of soup in Example 9.2 are normally distributed with $\sigma_x = 1.0$ ounces.

a. Draw a picture of the population distribution. What is its mean? What is its standard deviation?

b. Within what limits above and below the mean will approximately 95 per cent of the weights lie? What fraction of weights will be between $\mu_x \pm 3\sigma_x$?

c. In a sample of $n = 10$, what are $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$? Draw a picture of the probability distribution of $\bar{X}$. What is its form? What assumptions are necessary to ensure this form? How does this distribution compare with the population distribution?

d. Within what limits above and below the mean will about 99.7 per cent of all sample means lie if $n = 10$? What fraction of all sample means will be within range $\mu_{\bar{X}} \pm 1.96 \sigma_{\bar{X}}$?

e. Set up the statistical decision rule in Example 9.2, using your own value of α . Specify exactly H_0 , H_1 , and R . If the mean of a random sample of 10 cans is 11.8 ounces, what decision do you reach?

f. What is the general shape of the OC curve in this problem? What do points on this curve mean?

g. With $\alpha = .10$, one point on the OC curve is $\beta = .5$ if $\mu_1 = 12.52$. What do these values of α and β mean? If $\mu_1 = 11.48$, $\beta = .5$. Interpret this value of β .

9.11. For your own value of α specified in Exercise 9.10, find several points on the OC curve, and draw a figure of this curve.

9.12. Suppose that the quality-control supervisor submits the following specifications for sampling cans of soup:

$$\mu_0 = 12 \quad \alpha = .10 \quad \mu_1 = 12.5, \text{ or } 11.5 \quad \beta = .01$$

a. Draw the OC curve specified by these four values. What sample size is necessary to make the OC curve go through these four points?

b. How are values of μ_1 determined?

9.13. A random sample of 10 cans of soup has a mean $\bar{X} = 12.6$ ounces. Make the decision to accept H_0 or H_1 by judging this value of $\bar{X}$ in terms of Z .

9.5 Sampling a Normal Population with Unknown Standard Deviation

Example 9.1 is now restated in slightly different form. You are the purchasing agent for a machine shop which buys iron castings from a foundry. The crucial specification is weight, for a casting which is too heavy will not easily lend itself to further processing. From its detailed statistical records, the foundry reports that the weights of its castings are almost normally distributed with a mean of 20.0 pounds. From every purchase of several hundred castings, you take a small sample and check the weights against the average reported by the foundry. You find that the weights in a sample of 10 are 19.8, 20.3, 20.6, 21.1, 19.3, 19.6, 20.1, 20.8, 21.1, 21.3. As the purchasing agent, would you

accept the foundry's claim that the average weight of the castings is 20.0 pounds?

This problem is more realistic than the previous one because the standard deviation of the population is not known. It must be estimated from the sample by the formula

$$s_x = \sqrt{\frac{\sum(x - \bar{X})^2}{n - 1}} \quad (9.9)$$

Under this set of assumptions—*normal population with known mean but unknown standard deviation*—the probability distribution of the mean has the following properties:

1. Its mean is the mean of the population.

$$\mu_{\bar{X}} = \mu_x \quad (9.10)$$

2. Its *estimated standard deviation* is the standard deviation of the sample divided by the square root of the sample size.

$$s_{\bar{X}} = \frac{s_x}{\sqrt{n}} \quad (9.11)$$

where s_x is defined by Eq. (9.9).

3. The means are normally distributed, but the quantity

$$\frac{\bar{X} - \mu_x}{s_{\bar{X}}} = t \quad (9.12)$$

is distributed according to another *probability distribution* called the *t distribution*.

The *t* distribution is a continuous, symmetrical, bell-shaped probability distribution. Its mathematical function contains only one *parameter* which is called *degrees of freedom*. In the present problem, the number of degrees of freedom is one less than the sample size: $n - 1$. Degrees of freedom will be abbreviated D.F., so that

$$\text{D.F.} = n - 1 \quad (9.13)$$

The *t* distribution has been partly tabled in Table A.4 of the Appendix. The degrees of freedom are shown in the first column; the entries in the body of the table are the positive values which *t* will exceed with the probabilities given in the column headings. For example, with D.F. = 9, the entry for a probability of .025 is $t = 2.262$. This means that for D.F. = 9, the probability that *t* will exceed 2.262 is .025. Since the distribution is symmetrical, the probability that *t* will be less than -2.262 for 9 D.F. is also .025. A picture of the tabled values of *t* is shown at the top of the table as a constant reminder of the probabilities

given in this table of the t distribution. With this new sampling distribution of the mean, the problem of the mean weight of castings may be solved under the assumption that the standard deviation of the population is unknown and must be estimated from the sample.

The sample mean is 20.4 pounds. The sample standard deviation is computed from the 10 observations: 19.8, 20.3, 20.6, 21.1, 19.3, 19.6, 20.1, 20.8, 21.1, 21.3. From these data: $\Sigma x = 204.0$; $\Sigma x^2 = 4165.90$

$$s_x^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1} = \frac{4165.90 - 204^2/10}{9} = .48 \quad s_x = .69$$

The statistical test is formally set out below.

$H_0: \mu_0 = 20.0$ pounds (foundry's claim)

$H_1: \mu_1 > 20.0$ pounds (overweight castings for machine shop)

$n = 10$ $\alpha = .05$ $\bar{X}$: random variable

$n - 1 = 9$ D.F. $t_{\alpha}(9) = 1.833$ $R: \bar{X} \geq \bar{X}_r$ (upper tail)

The value of t which cuts off .05 in the upper tail of the t distribution for

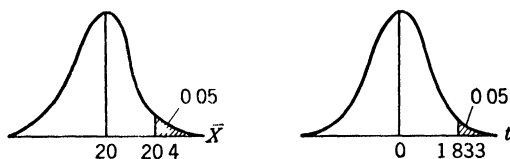


FIG. 9.6. Regions of rejection in the $\bar{X}$ and t distributions.

9 D.F. is 1.833. The region of rejection in the $\bar{X}$ distribution is found from the relationship,

$$t = \frac{\bar{X}_r - \mu_0}{s_{\bar{X}}}$$

These regions are shown in Fig. 9.6.

$$\mu_{\bar{X}} = \mu_x = 20.0 \quad s_x = .69 \quad s_{\bar{X}} = \frac{.69}{\sqrt{10}} = .22$$

Therefore,

$$1.833 = \frac{\bar{X}_r - 20}{.22} \quad \bar{X}_r = 20.4$$

Decision Rule. Accept H_1 if $\bar{X} \geq 20.4$; otherwise accept H_0 .

Decision. The sample mean 20.4 is at the border line of significance, i.e., at the border line of deciding to buy the shipment.

In summary, under the assumption that the population is a normal distribution with known mean but unknown standard deviation, which is estimated from the sample standard deviation, the appropriate prob-

ability distribution of the mean is the t distribution with $n - 1$ degrees of freedom. Table A.4 gives probabilities for many different degrees of freedom.

How to Judge Sample Results with Unknown Standard Deviation. A variation on the method of judging sample results presented in Sec. 9.4 is used when the standard deviation of the population is unknown and must be estimated from the sample. The characteristic of the sample is now the sample value of t computed from

$$t = \frac{\bar{X} - \mu_x}{s_x}$$

where s_x is computed from Eq. (9.9) and $s_x = s_x/\sqrt{n}$. The distribution of t is divided into two regions, A and R , so that A includes values of t which accept H_0 and R values of t which accept H_1 . The dividing line between R and A is the value of t for $n - 1$ degrees of freedom which cuts off the probability α (or $\alpha/2$), depending upon H_1 , in the appropriate tail (or tails) of the distribution. The decision rules for the three possible alternative statements of H_1 parallel the rules given for Z in Sec. 9.4.

Example: Decision rule for weight of foundry castings.

$$\begin{array}{lll} H_0: \mu_0 = 20.0 & \alpha = .05 & s_x = .69 \\ H_1: \mu_1 > 20.0 & t_{\alpha}(9) = 1.833 \text{ (upper tail)} & s_x = .22 \\ n = 10 & n - 1 = 9 \text{ D.F.} & \bar{X} = 20.4 \text{ (observed)} \end{array}$$

Decision Rule. Accept H_1 and refuse the shipment if $t \geq 1.833$; otherwise accept H_0 and buy the shipment.

From the sample, $t = (20.4 - 20.0)/.22 = 1.82$.

Decision. The sample value of $t = 1.82$ is approximately at the border line between accepting H_0 and accepting H_1 . The result differs from the decision based on the rule in terms of $\bar{X}$ only because of rounding errors in the computation of t and $\bar{X}_r$.

Operating-characteristic Curve. When the appropriate probability distribution is the t distribution, the evaluation of β in terms of the operating characteristic curve is beyond the scope of this book. But the idea of the OC curve is not beyond the scope of the book, and the OC curve for the decision rule of Sec. 9.5 is given in Table 9.5. Points on the OC curve are the probabilities of accepting H_0 and buying the shipment if the mean weight of the shipment is the value of μ in column 1. Thus, the probability of buying a shipment with a mean weight of approximately 20.4 pounds is .5, and the probability of buying a shipment with a mean weight of approximately 21 pounds is .01. As the true mean weight of shipments becomes larger and larger, the probability of accepting H_0 and buying them becomes smaller. Without the

OC curve, the probability of buying shipments with mean weights greater than $\mu_0 = 20$ would be unknown. The probability of accepting H_1 and not buying is fixed at .05 when $\mu = 20.0$. Errors of the second kind may be made with high probability because of the small sample size, but this fact would never be known without the OC curve.

Remark 1. The t distribution is commonly called "Student's t " after the pseudonym of the man who first derived it.

Remark 2. The table of the t distribution (Table A.4) lists selected values of the degrees of freedom up to 120. For over 120 D.F., the last row of the table for infinite (∞) degrees of freedom is used.

Remark 3. For infinite degrees of freedom the value of t associated with a probability of .05 is 1.645 and a probability of .025, 1.96. These are the same as the values of the normal variable Z for the same probabilities. As n becomes larger and larger, the t distribution approaches the normal distribution.

TABLE 9.5. OPERATING-CHARACTERISTIC CURVE*			
$H_0: \mu_0 = 20.0$	$\alpha = .05$	$s_x = .69$	$R: \bar{X} \geq 20.4$
$H_1: \mu_1 > 20.0$	$n = 10$	$s_{\bar{X}} = .22$	
Value of $\mu =$		$Pr(Acc H_0)$	
20 00		.95 = $1 - \alpha$	
20 09		.90	
20 19		.80	
20 27		.70	
20 33		.60	
20 39		.50	
20 45		.40	
20 52		.30	
20 59		.20	
20 70		.10	
20 79		.05	
20 95		.01	

* Computed from Table I-PII ($\alpha = .05$) in Errors of the Second Kind in Testing "Student's" Hypothesis, by J. Neyman and B. Tokarska, *Journal of the American Statistical Association*, vol. 31, p. 322, June, 1936.

EXERCISE

9.14. Assume that the 10 cans of soup in Example 9.2 weigh the following amounts: 12.0, 12.5, 13.0, 11.0, 12.5, 12.0, 11.5, 13.5, 13.0, 13.0 ounces. Test the hypothesis that the mean weight of cans in the filling process is 12 ounces. Use decision rules stated in terms of $\bar{X}$ and t . What assumptions are necessary to the valid application of these rules? Does the sample lead to the acceptance of H_0 or H_1 ?

9.6 Sampling an Arbitrary Population

Each test in the preceding sections rests upon the explicit assumption that the population is normally distributed. How does the fact that

many populations—even those with bell-shaped distributions—are not exactly normal in practice affect the probability distribution of the mean? One of the most important theorems assigning a central place in statistics to the normal distribution is the *central limit theorem*. It states that, for a population with mean μ_x and finite standard deviation σ_x , the probability distribution of

$$\frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} = Z$$

approaches the normal distribution as n becomes large. In other words, sample means are approximately normally distributed with mean μ_x and standard deviation $\sigma_x/\sqrt{n}$ in large samples from nonnormal populations.

The theorem is stated in terms of the population parameters μ_x and σ_x , but if σ_x is unknown, the sample standard deviation is used as an estimate. If the population is approximately normal, even though the sample standard deviation is used to estimate the standard deviation of the population, the sampling distribution of the mean is approximately normal for large sample sizes.

Confidence in the assumption of normality for the distribution of means of large samples comes both from the central limit theorem and from sampling practice. Failure of the normal approximation occurs generally under two conditions: (1) the distribution of the population is very peaked; or (2) the distribution of the population contains some extreme observations which dominate the average. Distributions of economic data, like sizes of towns, farms, and certain business enterprises, are characterized by many small units and a few very large ones. A moderate amount of skewness will not destroy the validity of the normality assumption, but a few extreme items that cause marked skewness should be segregated and treated separately. The removal of the extremes modifies the skewness and improves the normal approximation.

The sampling distributions of the means from some nonnormal populations have been worked out and recorded by statisticians. When the exact sampling distribution is not known, it is customary to assume normality, after exercising the precaution noted above, if the samples are large.

Example 9.3. For the past several years a personnel director of a large company has been giving all employees an aptitude test. The compiler of the test has recorded results for several thousand people scattered widely over the United States and has found that the average score is 80. The personnel director wishes to determine whether his employees differ from the population in general, i.e., whether they tend to have higher or lower aptitude test scores, on the average. He selects 400 employees at random from all employees and finds that $\bar{X} = 82$

and $s_x = 10$. The decision procedure is set out below:

$$\begin{aligned} H_0: \mu_0 &= 80 & n &= 400 \\ H_1: \mu_1 &\neq 80 & \bar{X} &= 82 \text{ (observed in sample)} \\ \alpha &= .05 & \frac{\alpha}{2} &= .025 & s_x &= 10 \text{ (computed from sample)} \\ R: \text{two-tailed} & & s_{\bar{X}} &= \frac{10}{\sqrt{400}} = .5 \end{aligned}$$

The values of Z which cut off .025 in each tail of the *normal* distribution are -1.96 and 1.96 . The critical values of $\bar{X}$, which define the region of rejection are

$$-1.96 = \frac{\bar{X}_r - 80}{.5} \quad \bar{X}_r = 79.2$$

and

$$1.96 = \frac{\bar{X}_r - 80}{.5} \quad \bar{X}_r = 80.8$$

Decision Rule. Accept H_1 if $\bar{X}$ is greater than 80.8 or less than 79.2.

Decision. Accept H_1 because the sample mean is 82. The personnel director has reason to believe that his employees differ in average aptitude test score from the whole population, and that, on the average, they have higher scores.

Operating-characteristic Curve. The operating characteristic curve is shown in Table 9.6. Each point on the OC curve is the probability of accepting H_0 for a given value of μ . In general,

$$\beta = \Pr(79.2 \leq \bar{X} \leq 80.8 | \mu_1)$$

since the region of acceptance includes values of $\bar{X}$ from 79.2 to 80.8. This probability is found by converting each critical value of $\bar{X}$ into a Z score and finding the area *between* the two values of Z .

TABLE 9.6. OPERATING-CHARACTERISTIC CURVE FOR THE TEST OF THE MEAN APTITUDE TEST SCORE

$\mu =$	Region of acceptance in Z scores	$\Pr(\text{Acc } H_0 \mu)$ (area between Z values)
78 0	2 4 - 5 6	01
78 5	1 4 - 4 6	08
79 0	0 4 - 3 6	36
79 5	-0 6 - +2 6	72
80 (= μ_0)	-1 96 - +1 96	95 = 1 - α
80.5	-2 6 - +0 6	72
81 0	-3 6 - -0 4	.36
81 5	-4 6 - -1.4	.08
82 0	-5 6 - -2 4	01

Example 9.4. Let $\mu_1 = 81$.

For $\bar{X} = 79.2$,

$$Z = \frac{79.2 - 81}{.5} = -3.6$$

For $\bar{X} = 80.8$,

$$Z = \frac{80.8 - 81}{.5} = -.4$$

$$\Pr(-3.6 \leq Z \leq -.4) = .36 \quad (\text{from Table A.3})$$

The probability of accepting H_0 becomes smaller with both large and small values of μ because the alternative hypothesis specifies values of μ both larger and smaller than $\mu_0 = 80$. The values of β in this test are not a minimum with respect to all alternative values of μ_1 , but the test is a *best unbiased test* because the probability of accepting H_0 is greater for $\mu_0 = 80$ than for any other value of μ_1 .

Sample Size with Unknown Standard Deviation. The sample size necessary to control both α and β is a variation on Eq. (9.8) of Sec. 9.4. The standard deviation is estimated from the sample by s_x , and

$$n = \left[\frac{s_x(Z_\alpha + Z_\beta)}{\mu_1 - \mu_0} \right]^2 + \frac{Z_\alpha^2}{2} \quad (9.14)$$

The number of observations is slightly more than that determined from Eq. (9.8), if s_x equals σ_x , and the increase depends only upon the value of α .

Example 9.5. Equation (9.14) is derived for a one-tailed test. The aptitude score problem involves a two-tailed test because of H_1 . An approximation to the necessary sample size can be determined by treating the problem as a one-tailed test with $\alpha/2$ probability in that tail. Suppose the specifications of two points on a desired OC curve are as follows:

$$\begin{array}{llll} H_0: \mu_0 = 80.0 & \alpha = .05 & \frac{\alpha}{2} = .025 & Z_{\alpha/2} = 1.96 \\ H_1: \mu_1 = 81.0 & \beta = .025 & & Z_\beta = 1.96 \end{array}$$

From the sample, $s_x = 10$. By Eq. (9.14),

$$n = \left[\frac{10(1.96 + 1.96)}{81 - 80} \right]^2 + \frac{1.96^2}{2} = 1534$$

In practice, a sample of approximately 1500 observations is necessary to ensure that, when the true mean aptitude test score is 80, the probability of accepting H_1 (that it is either more or less than 80) is approximately .05; that, if the true mean aptitude test score is 81, the probability of accepting H_0 (that the mean score is 80) is approximately .025. The choice of the critical value for $\mu_1 = 81$ rests upon the practical considerations of the prospective use of the test and of the seriousness of accepting that $\mu_0 = 80$ when it is, in fact, as large as 81.

9.7 Summary of Sampling Distributions of the Arithmetic Mean

The several sets of assumptions and the sampling distribution of the mean appropriate to each are reviewed here for easy reference.

Normal population with known mean μ_x and known standard deviation σ_x (Sec. 9.3): $\bar{X}$ is normally distributed with mean $\mu_{\bar{X}} = \mu_x$ and standard deviation $\sigma_{\bar{X}} = \sigma_x/\sqrt{n}$. $Z = (\bar{X} - \mu_{\bar{X}})/\sigma_{\bar{X}}$.

Normal population with known mean μ_x but unknown standard deviation (Sec. 9.5): The sample standard deviation s_x is used as an estimate of the population standard deviation. The quantity $t = (\bar{X} - \mu_{\bar{X}})/s_{\bar{X}}$ is distributed according to the t distribution with $n - 1$ degrees of freedom, where $s_{\bar{X}} = s_x/\sqrt{n}$ and $s_x = \sqrt{\sum(x - \bar{X})^2/(n - 1)}$.

Arbitrary population with mean μ_x and standard deviation σ_x (Sec. 9.6): $\bar{X}$ is approximately normally distributed with mean $\mu_{\bar{X}} = \mu_x$ and standard deviation $\sigma_{\bar{X}} = \sigma_x/\sqrt{n}$. The larger the sample size, the better is the approximation. If σ_x is unknown, it is customary to assume approximate normality for the distribution of $\bar{X}$ if the samples are large and s_x is used to estimate the standard deviation of the population. $Z = (\bar{X} - \mu_{\bar{X}})/s_{\bar{X}}$.

EXERCISES

9.15. On the basis of 100,000 measurements, the average height of draftees into the armed services in 1942 is estimated at 66.0 inches with a standard deviation of 2 inches. A random sample of 400 draftees from a given state has an average height of 66.3 inches. Is it reasonable to assume that these draftees came from a population with the same mean as all draftees?

9.16. Past records of a filling station indicate that the average number of gallons per customer is 8.5 with a standard deviation of 1.2 gallons. A new procedure of the station is to have each attendant ask each customer buying gasoline, "Fillerup?" At the end of the first week, the average sale to 900 customers is 8.7 gallons. Is it worthwhile to insist that the attendants use this new technique? If you accept H_1 , is the increase in the average sales practically significant?

9.8 Estimation of the Arithmetic Mean

Many practical problems in statistics involve estimation of the arithmetic mean because there is no preconceived notion about the value of the population mean from which to form a hypothesis. Examples of such estimations are:

1. The average number of persons per dwelling unit for real estate research
2. The mean income of the residents of a city for census purposes
3. The mean value of a characteristic of a product to guide purchasers of that product

In all these problems the unknown mean must be estimated, since no claim or hypothesis about its value is set forth. The estimate can be made with sufficient precision from samples, and will cost less than a complete count of the population.

Point Estimates. Suppose that a random sample of 30 is selected from all charge accounts at a department store. The mean balance of the sample is \$30 and the standard deviation is \$20. What is the estimate, based upon the sample, of the unknown mean balance of all charge accounts at this store? Ordinarily, the sample mean is used as an estimate of the unknown population mean. Is this a justifiable procedure? Whether it is justifiable depends upon the properties of the estimate so obtained. What properties are desirable for estimates of this kind? Two are suggested here.

In the first place, the estimate, if computed for a large number of random samples of the same size from a given population, should be such that the average of all the sample estimates equals the true population value. Such an estimate is said to be *unbiased*. In the second place, the estimate should fluctuate as little as possible from sample to sample. If one method leads to an estimate with a standard deviation of 2, for example, and another to an estimate with a standard deviation of 4, the first method is generally preferable to the second. The method of estimation which results in estimates with the smallest or minimum sampling variation (as measured by the standard deviation) is said to be *efficient*.

An estimate which is sometimes both unbiased and efficient is the *maximum-likelihood estimate*. This is a single number, or a *single-valued estimate*. In random samples from a known normal population, the mean is also normally distributed; in random samples from a normal population with unknown standard deviation, the mean is distributed according to the t distribution. The maximum-likelihood estimate of μ , called $\hat{\mu}$, is the value which if substituted for μ in either the normal probability distribution or the t distribution, depending upon the assumptions about the population, makes the probability of occurrence of the estimate a maximum. The exact determination of maximum-likelihood estimates is a problem for the calculus; only the final answer is given here. The maximum-likelihood estimate of the population mean is $\hat{\mu} = \bar{X}$, the sample mean. This is the best single-valued estimate of μ according to the criterion of maximum likelihood. The common practice of using the sample mean to estimate the unknown population mean thus has statistical justification.

The definition of the maximum-likelihood estimate may be confusing because it is a conditional statement. Restated in somewhat different terms, a maximum-likelihood estimate of an unknown parameter is that

value computed from the sample which, if it were the parameter value, would maximize the probability of occurrence of the observed sample. In the simple case of the arithmetic mean, the sample mean is the maximum-likelihood estimate of the unknown population mean.

Confidence-interval Estimates of the Arithmetic Mean. Although the use of the sample mean as a single-valued estimate of the unknown population mean is a common and statistically justifiable practice, it does not follow that the population mean is *equal* to the sample mean $\bar{X}$. Randomness in sample selection will cause $\bar{X}$ to differ from μ . The difference between $\bar{X}$ and μ is often called the *random error*. It is not an error in the sense of a mistake, but merely a difference that occurs because the population is judged from only a sample of observations. It seems reasonable, then, to make allowance for this random error by adding something to the sample $\bar{X}$. A recommended procedure is to make an interval of the form $\bar{X} \pm$ "something" and to state that this interval includes the true value of μ . It will never be known with certainty whether the interval includes μ ; the statement that it does include μ could be wrong. Therefore, a probability is also attached to the interval—the probability that the interval does include the true value of μ .

The longer the interval, the greater is the probability that it contains μ ; the shorter the interval, the smaller is the probability that it contains μ . But, the longer the interval, the less precise is the estimate of μ ; the shorter the interval, the more precise is the estimate of μ , that is, the narrower the range which attempts to locate μ . It is desirable to have a long interval for a high probability that the interval contains μ and a short interval for greater precision of the estimate. The inherent conflict between the probability and the precision is resolved in Sec. 9.9.

The problem of establishing a confidence-interval estimate of the unknown population mean μ may be stated in this way. First, a random sample is selected. Second, on the basis of this sample lower and upper limits, designated by $\underline{\mu}$ and $\bar{\mu}$, are set to form an interval $\underline{\mu}$ to $\bar{\mu}$, with $\underline{\mu}$ called the *lower confidence limit* and $\bar{\mu}$ the *upper confidence limit*. The interval may not contain μ because of random sampling variation. The probability that the interval will contain μ must, therefore, be set in advance. This probability is called the *confidence coefficient* and is written C.C. It is an arbitrarily selected probability that the interval contains μ . It may be a probability of .9, .95, .5, or any other value.

When it is assumed that the population is normal with unknown standard deviation, the arithmetic mean is distributed according to the t distribution with $n - 1$ degrees of freedom and the standard deviation is estimated by s_x from the sample (see Sec. 9.5). Let $\bar{X}$ be the observed sample mean (\$30) and $\bar{\mu}$ be the upper confidence limit in the charge account problem. Now, what value of $\bar{\mu}$ will give a probability of .025

of observing a sample mean equal to or less than $\bar{X} = 30$? The value of t associated with a probability of .025 in the lower tail of the t distribution for $n - 1 = 29$ degrees of freedom is -2.045 (Table A.4). The relationship between $\bar{X}$ and t is

$$-2.045 = \frac{\bar{X} - \underline{\mu}}{s_{\bar{X}}}$$

where $s_{\bar{X}} = s_x/\sqrt{n}$. Let $\underline{\mu}$ be the lower confidence limit. What value of $\underline{\mu}$ will give a probability of .025 of observing a sample mean equal to or greater than $\bar{X} = 30$? The value of t associated with a probability of .025 in the upper tail of the t distribution for $n - 1 = 29$ degrees of freedom is 2.045 . The relationship between t and $\bar{X}$ is

$$2.045 = \frac{\bar{X} - \underline{\mu}}{s_{\bar{X}}}$$

These two equations are solved for $\underline{\mu}$ and $\bar{\mu}$.

$$\bar{\mu} = \bar{X} + 2.045(s_{\bar{X}}) \quad \underline{\mu} = \bar{X} - 2.045(s_{\bar{X}})$$

In the present example, $s_x = \$20$ and $n = 30$. $s_{\bar{X}} = 20/\sqrt{30} = 3.6$.

$$\bar{\mu} = 30 + 2.045(3.6) = 37.36 \doteq \$37$$

$$\underline{\mu} = 30 - 2.045(3.6) = 22.64 \doteq \$23$$

The 95 per cent confidence interval for the mean charge-account balance is from \$23 to \$37.

In general, the confidence limits for the arithmetic mean of a random sample from a normal population with unknown standard deviation are

$$\underline{\mu} = \bar{X} - t_{c.c.}(n-1) \cdot s_{\bar{X}} \quad (9.15)$$

$$\bar{\mu} = \bar{X} + t_{c.c.}(n-1) \cdot s_{\bar{X}} \quad (9.16)$$

where $\pm t_{c.c.}(n-1)$ = values of t which for $n-1$ degrees of freedom include a probability equal to the confidence coefficient selected and $s_{\bar{X}} = s_x/\sqrt{n}$. The interval is often written

$$\bar{\mu} = \bar{X} \pm t_{c.c.}(n-1) \cdot s_{\bar{X}} \quad (9.17)$$

When the form of the population is not specified, the sample mean is approximately normally distributed in large samples. The standard deviation may be estimated from the large sample (see Sec. 9.6). Then, the confidence limits become

$$\underline{\mu} = \bar{X} - Z_{c.c.} \cdot s_{\bar{X}} \quad (9.18)$$

$$\bar{\mu} = \bar{X} + Z_{c.c.} \cdot s_{\bar{X}} \quad (9.19)$$

$$\bar{\mu} = \bar{X} \pm Z_{c.c.} \cdot s_{\bar{X}} \quad (9.20)$$

where $s_{\bar{x}} = s_z/\sqrt{n}$ and $\pm Z_{C.C.}$ are the values of Z in the normal probability distribution which include a probability equal to the confidence coefficient.

Example 9.6. Let $n = 400$; $\bar{X} = \$30$; $s_z = \$20$.

With C.C. = .95,

$$Z_{C.C.} = \pm 1.96 \quad (\text{from Table A.3})$$

$$s_{\bar{x}} = \frac{20}{\sqrt{400}} = 1$$

From Eq. (9.20),

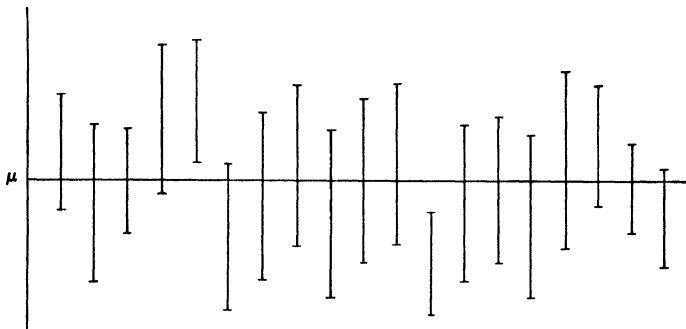
$$\bar{\mu} = 30 \pm 1.96(1) \quad \bar{\mu} = \$28 \text{ to } \$32$$

Several comments are in order at this point. First, the confidence coefficient is selected exactly as α is selected in testing a hypothesis about the value of μ . A value of $\bar{X}$ is observed in a random sample of n observations. What admissible values of μ would be accepted by this sample value of $\bar{X}$ and what values would be rejected? All the accepted values of μ constitute the interval $\bar{\mu}$ to $\bar{\mu}$. All the rejected values of μ fall outside the interval. The confidence coefficient C.C. is equal to $1 - \alpha$. The size of α depends upon the cost of error I: accepting H_1 when H_0 is true. The size of the C.C. depends upon the seriousness of forming an interval which does not include μ or, in other words, the importance of forming an interval which does include μ .

Second, what is the meaning of the statement, "The probability is .95 that the interval includes μ "? The confidence coefficient, a probability which may be used to predict the outcomes of a physical experiment, is interpreted in terms of the relative frequency of occurrence in the experiment. The physical experiment of an estimation problem consists of selecting a random sample, forming an interval, and making the statement that the interval contains the parameter. The confidence coefficient has meaning only in relation to this procedure. If the sampling process is repeated in exactly the same manner for a large number of samples, if an interval is computed for each sample in exactly the same way, and if the statement is made that each interval contains the parameter, the confidence coefficient is a prediction about the relative frequency of occurrence in a long series of random trials. With a confidence coefficient of .95, 95 per cent of the intervals formed from a large number of random samples in the same manner as in the example will include the true value of μ . Or, if the statement is made that each interval contains μ , 95 per cent of such statements will, on the average, be correct. The 95 per cent also holds true for samples selected from different populations in the course of a lifetime. Even if the technique is applied differently each time, 95 per cent of the statements are still expected to be true.

The meaning of a confidence coefficient is not always clearly understood. It is erroneous, but unfortunately popular, to interpret a confidence coefficient by saying that the probability is .95 that the true value of μ lies in a certain interval (such as \$28 to \$32). The population mean is a constant and therefore has no probability distribution. Probabilities are attached to variables, not to constants. The intervals fluctuate because they are computed from random samples, the means of which vary from sample to sample. It is therefore correct to speak of the probability that an interval will contain the mean, but not of the probability that the mean will lie within the interval. The intervals are subject to random sampling fluctuation; the true mean is not subject to this fluctuation. The true mean either does or does not lie within an interval, and hence there can be no statement as to the probability of its lying within the interval. Determination of whether it lies within a certain interval depends solely on examination of the whole population, a process which automatically dismisses the problem of estimation.

A diagram may clarify this discussion on the meaning of the confidence coefficient. Locate the true value of μ on a vertical scale and draw a straight line horizontally at this point. Take a series of samples, find the values of $\bar{X}$ in the samples, and compute the confidence intervals from these samples. Plot these intervals on the diagram, as follows:



The values of $\bar{X}$ and hence the confidence intervals computed from them vary. On the diagram, some of the intervals will overlap μ and others will not. The confidence coefficient is the proportion of these intervals that will overlap the parameter μ on the average or in the long run. (Reread the remark at the end of Sec. 5.2.)

Remark. Not only will the general level of the interval vary because of different means in different samples, but the lengths of the intervals will also vary because the sample standard deviation varies from one sample to the next.

Finally, why is an interval that misses the true parameter ever constructed? Why not make an interval with 100 per cent confidence that

it traps the true value of μ ? If the random variable is the mean charge-account balance, what is the 100 per cent confidence interval? The interval from minus infinity to plus infinity ($-\infty$ to $+\infty$) is absolutely certain to contain μ . In fact, the interval from $-\$10,000$ (a credit balance) to $\$100,000$ may absolutely contain μ . These are trivial answers. The interval is shortened to approximately \$4 in length (or to \$14, depending upon the sample size), and this interval is located along the scale of μ with a loss of only 5 per cent confidence.

9.9 Sample Size: Estimation of $\bar{X}$

The problem of determining the sample size in estimating an arithmetic mean is quite similar to that in estimating a proportion (Sec. 5.3). For example, the average income of families in a city is to be estimated from a random sample selected from a recent city directory. Is the average income to be estimated within $\pm \$30$, $\pm \$200$, $\pm \$500$? In other words, how long is the desired confidence interval? \$60? \$400? \$1000? Suppose that the average income is to be estimated within $\pm \$30$. Whatever the average income $\bar{X}$, the objective is to estimate it within an interval of length $\bar{X} \pm 30$.

With large samples, the confidence-interval estimate takes the form

$$\bar{\mu} = \bar{X} \pm Z_{c.c.} \cdot s_{\bar{X}} \quad (9.20)$$

If the interval is to have a length $\bar{X} \pm 30$, then

$$\bar{X} \pm Z_{c.c.} \cdot s_{\bar{X}} = \bar{X} \pm 30 \quad \text{or} \quad Z_{c.c.} \cdot s_{\bar{X}} = 30$$

Since $s_{\bar{X}} = s_x / \sqrt{n}$

$$Z_{c.c.} \cdot \frac{s_x}{\sqrt{n}} = 30 \quad (9.21)$$

With C.C. = .95, $Z_{c.c.} = \pm 1.96$. Equation (9.21) can be solved for n if s_x can be estimated. An estimate of s_x good enough for determining the approximate sample size can probably be obtained from previous experience, from someone else who has had such previous experience, or from a small "pilot" sample. In this problem, if $s_x = 200$,

$$1.96 \frac{(200)}{\sqrt{n}} = 30 \quad n = \frac{(200)^2(1.96)^2}{30^2} = 171$$

This is a rough idea of the sample size. In practice, 200 observations would be selected.

The sample size depends upon:

1. The length of the desired interval (d = half the interval)
2. The confidence coefficient ($Z_{c.c.}$)
3. An estimate of the standard deviation (s_x)

$$n = Z_{c.c.}^2 \cdot \frac{s_x^2}{d^2} \quad (9.22)$$

The length of the interval depends upon how precisely μ is to be located. The confidence coefficient depends upon how important it is to make an interval which includes the true value of μ . An estimate of the standard deviation can always be obtained in some manner. All the information necessary to design an experiment properly can never be known in advance. After the study is finished, pertinent information which would have been useful in designing the study is certain to appear. An essential part of any learning process is doing. An estimate of the standard deviation sufficiently good to determine the sample size can always be obtained in practice.

Just as in testing hypotheses, samples that are too large and therefore too costly mean either that the experiment must be abandoned or that the specifications of the estimate must be changed. An arbitrary sample size, chosen without previous planning, may lead to an estimate that is too vague (i.e., the interval is too long) to be of any practical value or to an interval that is much too short (i.e., it is more precise than is needed). In the first instance, the question is not answered; in the second, the question could have been answered more cheaply.

Suppose, for example, that it is necessary to estimate within $\pm \$30$ with C.C. = .9973. Then, $d = 30$, $Z_{c.c.} = \pm 3$, and $s_x = 200$.

$$n = \frac{(3)^2(200)^2}{30^2} = 400$$

If it is necessary to estimate within $\pm \$10$ with 99.73 per cent confidence,

$$n = \frac{(3)^2(200)^2}{10^2} = 3600$$

Roughly, therefore, around 3500 to 4000 observations are needed. (If one were thinking in terms of 100 observations, this is not large enough; if one were thinking of 10,000 observations, this is too many.) If a sample of this size is too costly, either the survey should be dropped until more funds are available, or the precision of the estimate and/or the confidence coefficient should be revised to allow a smaller sample size.

Most statisticians believe that determination of the sample size is an essential part of any statistical experiment. For this reason, the advice of a statistician, if sought at all, should be sought at the inception of an analysis or a study, and not after the data have been taken. Research is too often carried out with arbitrary sample sizes, and the investigator arrives at estimates not precise enough to answer his questions. Such research is better left undone and the money saved.

EXERCISES

9.17. In Exercise 1.7, you selected a random sample of 10 observations from a bowl which contained 998 chips, each labeled with an identifying number and a value of a meaningless variable X . As a part of the exercise, you were asked to find the average value, $\bar{X}$, and to compute what you now know is the variance, s_x^2 , and the standard deviation, s_x .

The distribution of the 998 values of X is approximately normal. It cannot be exactly normal because the values of X are discrete and the normal distribution is continuous, but the number of chips which were assigned each value of X was determined from probabilities taken from a normal distribution with $\mu_x = 3.0$ and $\sigma_x = 1.0$. With these parameters and the sample values, $\bar{X}$ and s_x , answer the following questions:

a. Using the known value of $\sigma_x = 1.0$, set up a test of the hypothesis that $\mu_x = 3.0$. On the basis of your own sample average $\bar{X}$, do you accept or reject H_0 ? What is the correct decision? What percentage of the samples selected by the class lead to the rejection of H_0 ? How does this value compare with the value of alpha used?

b. Using the sample estimate of the standard deviation, set up a test of the hypothesis that $\mu_x = 3.0$. On the basis of your sample average, do you accept or reject H_0 ? What is the correct decision? What percentage of the samples selected by the class lead to a rejection of H_0 , and how does this value compare with alpha?

c. Using your own sample mean and standard deviation, construct an 80 per cent confidence interval for μ . Does your interval contain μ ? How many of the intervals computed by the class do contain the true mean? How does this value compare with the confidence coefficient? Interpret the meaning of confidence in an interval.

9.18. Use the data of Exercise 9.14 and make an interval estimate of the average weight of the cans of soup.

9.19. If you accepted H_1 in Exercise 9.15, what estimate do you make of the average height of draftees in the given state from which the sample of 400 came?

9.20. From the data of Exercise 9.16, what is your estimate of the average number of gallons per customer under the new selling technique? Would you want to measure any other characteristic of the gasoline sales in this new situation before making such an estimate?

9.10 Estimating Totals or Aggregates

The mean charge-account balance in a department store based upon a random sample of 30 accounts with $\bar{X} = \$30$ and $s_x = 20$, was estimated in Sec. 9.8 as

$$\begin{aligned}\bar{\mu} &= \bar{X} \pm t_{c \cdot c} (n - 1) \cdot s_x \\ \bar{\mu} &= 30 \pm 2.045(3.6), \text{ or } \$23 \text{ to } \$37\end{aligned}\tag{9.17}$$

A problem may be stated so as to require an estimate of the *total* or *aggregate* amount instead of an estimate of the mean. For example, the

department store may desire to estimate the total amount due from charge-account customers. This total is determined exactly during the preparation of the profit and loss statement and the balance sheet at the end of the accounting period, but in the interim an estimate sufficiently precise for control or planning purposes may be found by means of a sample of charge accounts.

The sample mean may be used to estimate the total for the whole population by multiplying it by the total number of items in the population, i.e., the total number of charge accounts. If, for example, the store has 5000 charge-account customers, an estimate of the total receivables is 5000(\$30), or \$150,000. Since this estimate is subject to the same general sampling errors as the arithmetic mean, an interval estimate is more desirable. The standard deviation of the estimated total is related to the standard deviation of the mean. Let X' be the estimated total found from the mean $\bar{X}$. Then,

$$X' = N\bar{X} \quad (9.23)$$

and

$$s_{x'} = N s_{\bar{x}} = \frac{N s_x}{\sqrt{n}} \quad (9.24)$$

where N is the population size, n is the sample size, and s_x is the standard deviation of the sample. When the standard deviation is estimated from the sample, the mean is distributed according to the t distribution with $n - 1$ degrees of freedom. Since multiplying by a constant does not change the form of the distribution, the total X' is also distributed according to the t distribution with $n - 1$ degrees of freedom. An interval estimate of the total then follows from Eq. (9.17):

$$\text{Estimated total} = X' \pm t_{c.c.}(n - 1) \cdot s_{x'} \quad (9.25)$$

In the department-store problem, $\bar{X} = 30$, $s_x = 20$, $n = 30$, and $N = 5000$. Therefore, $X' = 5000(30) = 150,000$;

$$s_{x'} = \frac{5000(20)}{\sqrt{30}} = 18,000$$

With C.C. = .95 and $n - 1 = 29$ D.F., $t_{.95} = 2.045$.

$$\begin{aligned} \text{Estimated total} &= 150,000 \pm 2.045(18,000) \\ &= \$113,000 \text{ to } \$187,000 \end{aligned}$$

The 95 per cent confidence-interval estimate of total accounts receivable is from \$113,000 to \$187,000.

In large samples, both the arithmetic mean and the total are approxi-

mately *normally* distributed. The confidence-interval estimate is of the form

$$X' \pm Z_{C.C.} s_{x'} \quad (9.26)$$

Example 9.7. Let $n = 400$, $N = 5000$, $\bar{X} = 30$, $s_x = 20$. The estimated total is $X' = N\bar{X} = 5000(30) = 150,000$. The standard deviation of X' is $s_{x'} = 5000(20)/\sqrt{400} = 5000$. With C.C. = .95, $Z = \pm 1.96$. Then, the interval estimate of the total receivables is

$$150,000 \pm 1.96(5000) \doteq \$140,000 \text{ to } \$160,000$$

The probability is .95 that the interval from \$140,000 to \$160,000 traps the true total value of accounts receivable.

9.11 Estimating the Median

The average most widely used in inference problems has historically been the arithmetic mean because its sampling distribution is easily derived under different sets of assumptions and because more research has been devoted to the investigation of its properties. More recently, however, statisticians have been exploring the possibilities of sample statistics whose application depends on less detailed assumptions. Among the sampling distributions of other averages currently being investigated is that of the median. The sampling distribution of the median has the important property of requiring no assumption about the form of the population distribution. In almost all problems concerning the arithmetic mean, it is assumed that the population is normally or very nearly normally distributed, but this is not necessary for estimates of the median.

The array of weights in the sample of 10 foundry castings in Example 9.1 is given below, each labeled with an x value. Thus, x_1 is the smallest observation in the array, x_5 is the fifth from the lower end, and x_{10} is the largest.

19.3	19.6	19.8	20.1	20.3	20.6	20.8	21.1	21.1	21.3
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

The median lies between the $(n/2)$ th and $(n/2 + 1)$ th observations; here, since $n = 10$, it lies between the fifth and sixth observations. It is the mid-point between 20.3 and 20.6, or $\bar{x} = 20.45$. Since this estimate of the median is based upon a random sample from the whole shipment of foundry castings, it is subject to sampling variability. An interval estimate of the median may be obtained by using Table A.5.

Table A.5 gives values of k and $n - k + 1$ so that the interval from x_k to x_{n-k+1} traps the true median with confidence of *more than* .95 or .99. The confidence coefficient expressing the probability that the interval traps the true median is stated as an inequality. Its exact value, which varies with the sample size, is given in the table.

With $n = 10$ and C.C. $\geq .95$, $k = 2$ and $n - k + 1 = 9$, from Table A.5. The values of x_2 and x_9 are the confidence limits for the median, and the exact confidence coefficient is .98. $x_2 = 19.6$ and $x_9 = 21.1$. Thus, the probability is .98 that the interval from 19.6 to 21.1 traps the true median weight of foundry castings, i.e., the weight which is exceeded by half of all the castings made. The valid application of this method depends only on the assumptions that the variable is continuous and the sample is random.

The confidence interval for the median of a continuous distribution is based upon the binomial distribution with $P = .5$, because any sample item is equally likely to be greater or less than the population median. In other words, the probability that each sample item will fall either above or below the median of the population is .5. In a random sample of n observations, $\mu = nP$ and $\sigma = \sqrt{nPQ}$; in large samples, the normal distribution is a good approximation to the binomial distribution (see Chap. 4). Since $P = .5$, $\mu = .5n$ and $\sigma = .5\sqrt{n}$. With a confidence coefficient C.C., the value of $Z_{c.c.}$ which includes the fraction C.C. of the normal distribution may be read from the table of the normal distribution. To find the confidence interval, the procedure is to compute the quantity

$$Z_{c.c.} (.5) \sqrt{n} \quad (9.27)$$

and take the *next larger integer*. It indicates the number of observations to the right and to the left of the median that comprise the interval. For example, with $n = 100$ and C.C. = .95, $Z_{c.c.} = 1.96$, and Eq. (9.27) becomes $1.96(.5) \sqrt{100} = 9.8$. The next larger integer is 10. The median lies between the fiftieth and fifty-first observations. Counting 10 observations to the right and to the left of the median estimates the interval as extending from the fortieth to the sixty-first observations. The probability is .95 that this interval traps the true population median.

Example 9.8. A sample of 100 charge accounts is selected at random from all charge accounts at a department store. Part of the array of the balances of these sample accounts, each with its observation number, is given below.

9.50,	...	33.75,	...	35.00,	35.50,	...	37.50,	...	75.00
x_1	...	x_{40}	...	x_{50}	x_{51}	...	x_{61}	...	x_{100}

The median balance of the sample is $\bar{x} = 35.25$. With $n = 100$ and C.C. = .95, $Z_{c.c.} (.5) \sqrt{n} = 9.8$, or, rounded up, 10. The 95 per cent confidence interval for the median charge-account balance extends from the fortieth to the sixty-first observation, or from 33.75 to 37.50. The probability is approximately .95 that this interval traps the true median of the population, or that balance which is exceeded by half of the charge accounts.

This method of estimating the median is one of a class of *distribution-free* methods, so called because they make no assumption about the form of the population distribution. Distribution-free estimates are usually simple as well as generally applicable. The computations (especially of the standard deviation) necessary for estimating a mean are much more tedious than those necessary for estimating the median. The most time-consuming feature of median estimation is making the array. In small samples, the entire sample distribution is arrayed. In large samples, only the middle items need be arrayed; the observations in the tails are merely counted to locate the median.

9.12 Estimating Percentiles

In general, the technique described for finding a confidence interval for the median may be used for other percentiles. For example, the probability is .75 that a sample item will exceed the 75th percentile, and .25 that it will be less than the 75th percentile. Therefore, $\mu = nP = .75n$, and $\sigma = \sqrt{nPQ} = .43\sqrt{n}$. For other percentiles, the values of P and Q are changed accordingly. In large samples, the normal approximation to the binomial is used in the same general way as it is for estimating the median.

9.13 Tolerance Intervals

The problem of setting a range of values for a variable, to include a certain fraction of the observations in the population, is frequently met in many types of analysis. For example, a machine shop may want to estimate a range including 95 per cent of the weights of individual castings, or a department store may want to estimate a range including a certain fraction of all charge-account balances. This range is called a *tolerance interval*, and the limits of the interval are called *tolerance limits*. In general, tolerance limits are limits between which a certain fraction of the population may be expected to lie.

The only tolerance intervals discussed here are those formed symmetrically around the mean $\bar{X}$ and based upon random samples from a *normally* distributed population. This assumption may seriously restrict the practical application of tolerance intervals. A tolerance interval takes the form

$$\bar{X} \pm Ks_x \quad (9.28)$$

where $\bar{X}$ is the mean of a sample, s_x is the standard deviation of the sample, and K is a constant to be determined.

The fraction or proportion of observations within the limits $\bar{X} \pm Ks_x$ is a random variable. Different samples will have different means and different standard deviations, and different fractions will be included between the limits. The statement that one particular interval includes

the fraction P of the population could be wrong. Therefore, a probability that this statement is correct must be attached to the interval. It is possible to determine K so that the probability that the interval $\bar{X} \pm Ks_x$ will include the fraction P or more of the population is known.

Tolerance limits are computed with the aid of Table A.6, which gives values of K for C.C. = .95 and .99 and P = .75, .90, .95, and .99 for selected values of n .

Example 9.9. The mean weight of foundry castings in Example 9.1 is 20.4 pounds and the standard deviation is $s_x = .69$. $n = 10$. Let C.C. = .95 and $P = .90$. The meaning of C.C. = .95 and $P = .90$ is: What interval of the form $\bar{X} \pm Ks_x$ has the probability .95 of trapping at least 90 per cent of the weights of the foundry castings? From Table A.6, $K = 2.839$. The tolerance interval is

$$20.4 \pm 2.839(.69) \doteq 18.4 \text{ to } 22.3$$

Thus, the probability is .95 that the interval from 18.4 to 22.3 pounds traps at least 90 per cent of the weights of all foundry castings.

With $P = .99$, $K = 5.649$ and the tolerance interval is

$$20.4 \pm 4.433(.69) \doteq 17.3 \text{ to } 23.5 \text{ lb}$$

The probability is .95 that the interval from 16.5 to 24.3 pounds traps at least 99 per cent of the weights of all foundry castings.

Tolerance limits must not be confused with confidence limits. Although confidence limits for the population mean are similar in form to tolerance intervals, the confidence coefficient has a different meaning. In respect to confidence limits, it represents the fraction of intervals computed from a large number of samples that will trap the population mean. In respect to tolerance intervals, on the other hand, the probability coefficient is the fraction of intervals computed from a large number of samples that will trap at least the proportion P of the population. Two coefficients express the uncertainty in the tolerance interval: P , the proportion of the population trapped by the interval, and C.C., the probability that the interval traps at least this proportion.

EXERCISES

9.21. From the data of Exercise 9.16, make an interval estimate of the total number of gallons of gasoline which may be sold to an estimated 4000 customers in the next month.

With 95 per cent confidence, what limits will you set to include 95 per cent of future sales of gasoline to individual customers?

9.22. It is desired to estimate the average income in a given community within $\pm \$100$ at the 90 per cent confidence level. An estimate of the standard deviation of incomes in this community, based upon a pilot study, is $s_x = \$500$. How large a sample is needed to estimate with the required precision?

9.23. A random sample of 746 adults is selected from the Census returns of a city of approximately 100,000 inhabitants. The following data are tabulated for this sample.

Tabulation	Total	Male	Female
Number of persons	746	348	398
Number receiving wage and salary income	375	252	123
Number receiving wage and salary income of less than \$1000	200	90	110
Total income received	\$378,900	\$306,785	\$72,115
Sum of squares (Σx^2)	548,724,315	498,615,211	50,109,104

- Find the average income of all persons, all males, and all females.
- Find the average income of all persons, all males, and all females who receive wage and salary income.
- What is the proportion of males and females in the city?
- What is the percentage of all persons, of males, and of females in the city who receive wage and salary income?
- For what proportion of males and females who receive wage and salary income is this income less than \$1000?
- Compute the 95 per cent confidence limits for each of the estimates (a) to (d). What assumptions are made in attaching this coefficient to each of the intervals?

9.24. In a statistical report the statement is made that the 95 per cent confidence interval for the mean number of inhabitants per dwelling unit is 2.8 ± 1.5 , or from 1.3 to 4.3 persons. Which of the following interpretations of this statement are true and which are false?

- That 95 per cent of the sample dwelling units will have between 1.3 and 4.3 inhabitants.
- That 95 per cent of future dwelling units drawn at random will have between 1.3 and 4.3 inhabitants.
- The probability is .95 that the true mean number of inhabitants per dwelling unit is between 1.3 and 4.3.
- If confidence intervals are computed in the same way for each of a large number of samples of the same size, 95 per cent of such intervals will overlap the interval 1.3 to 4.3.
- The mean number of inhabitants will be within 1.5 of 2.8, 95 per cent of the time.
- That 95 per cent of the time we would get 2.8 ± 1.5 .
- If a large number of samples of the same size is taken and confidence intervals computed in the same way for each, 95 per cent of such intervals will capture the true mean number of inhabitants per dwelling unit.
- If a large number of samples is taken, the interval 1.3 to 4.3 will capture about 95 per cent of the sample means.
- The probability that a dwelling unit will deviate by more than 1.5 is .95.
- You have 95 per cent confidence in this interval, the exact interpretation varying from one observer to another.

9.25. Evaluate this statement: If the standard deviations of the populations are the same, a sample of 500 from a population of 200,000 will give almost as precise an estimate of the population mean as a sample of 500 from a population of 10,000.

Do you find this result hard to believe? Does it seem intuitively obvious that if information has been obtained about only a small fraction of the population, the sample mean cannot be very accurate? How would you demonstrate that this point of view is erroneous? (*Hint:* See the Remark of Sec. 9.2 on sampling without replacement.)

9.26. A survey is being designed to obtain information about radio and television listening habits. In addition to writing the questionnaire, it is necessary to plan the sample size. It has been decided that the two most important aspects of the survey are the percentage of TV ownership and the average length of listening time per week. The basic time interval of 1 week is chosen because it eliminates day-to-day variation. It is desired to estimate with 95 per cent confidence intervals the percentage of TV ownership within ± 3 per cent, and the mean listening time per week within ± 1 hour.

A previous survey, done on a small scale, indicates that the percentage of TV ownership is about 70 per cent, the mean number of hours listening time per week is about 20, with a standard deviation of 3 hours. How large a sample should be taken to estimate with the desired accuracy? Which question controls the sample size: TV ownership or the amount of listening time per week? What are the implications of this problem for designing a survey with many different parts to it?

What are the problems in this kind of survey? For example, what kind of questions would you ask to elicit information about the amount of time spent listening to TV or radio each week? What problems would arise in obtaining this information for the whole family or separate individuals, adults or children?

9.27. In Sec. 8.2, the mean and standard deviations of the initial bus motor failures were computed in Table 8.3 as $\bar{t} = 96.7$ and $s_t = 37.5$. Make an interval estimate of the mean length of time to initial bus motor failure. Also, estimate the quartiles and the median length of time to initial failure.

9.28. From the data of Table 8.6, estimate the median time to failure of the fifth bus motor failures. Interpret this median. How does this estimate compare with that for the initial bus motor failures?

9.29. Use the data of Exercise 8.4 for this exercise. Estimate the mean lifetime of the 40-watt incandescent lamps. Estimate the median lifetime of these lamps. Which is the easier estimate to make?

Find the limits in which you have 95 per cent confidence that they will trap at least 95 per cent of the future lifetimes of similar lamps.

9.30. Use the data of Exercise 8.5 in this exercise. Estimate the mean and median number of student typing errors per interval. Which average would you use in a problem of this kind? Make a tolerance interval with these data using your own probability coefficients.

SUPPLEMENTARY READING

Moroney, M. J.: "Facts from Figures," Chaps. 3, 4, 5, 6, 9, 10, 13, and 14, Penguin Books, Inc., Baltimore, 1953.

CHAPTER 10

SIMPLE LINEAR REGRESSION AND CORRELATION

As the manager of the cost analysis department of a large manufacturing concern, you are charged with the responsibility of estimating the number of man-hours of work required to make various products. From available data on the total number of man-hours and of units of product made in the past, you must estimate the man-hours required to fill future production orders.

As an example, the data available for 25 different job orders of a particular product made in one department are given in Table 10.1. The

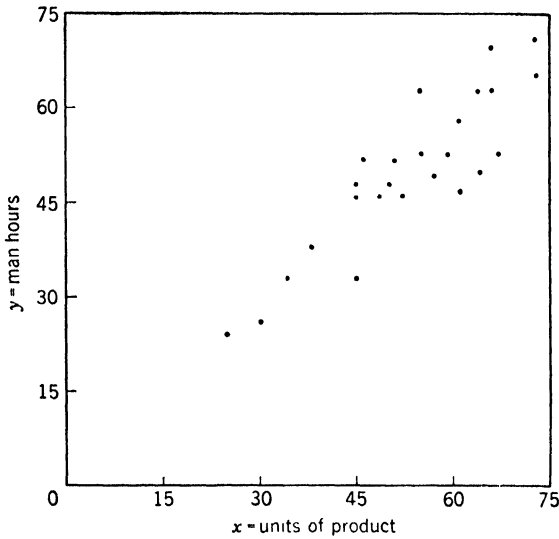


FIG. 10.1. Scatter diagram of units of product made and man-hours required to make them for 25 job orders.

total number of units of product is labeled x ; the total number of man-hours worked to make this number of units is labeled y . These data are plotted on Fig. 10.1. It is obvious from the figure that the more units of product made, the larger is the number of man-hours worked. In

other words, an apparent relationship exists between the number of units of product and the number of hours required to make them. The job of the cost analysis department is to express that relationship in mathematical terms and use it to estimate the man-hours of work required to fill future orders. Before solving this problem, a digression to some general ideas about the relationship between two variables is necessary.

TABLE 10.1. DATA ON UNITS OF PRODUCT AND TOTAL NUMBERS OF MAN-HOURS WORKED FOR 25 JOB ORDERS

Job	Units of product x	Total man-hours y	Job	Units of product x	Total man-hours y	Job	Units of product x	Total man-hours y
1	66	70	9	55	55	17	46	52
2	49	46	10	51	52	18	38	38
3	64	65	11	50	48	19	64	63
4	45	46	12	55	63	20	73	65
5	61	62	13	45	48	21	61	58
6	52	46	14	73	71	22	34	33
7	67	68	15	45	33	23	57	49
8	59	53	16	30	26	24	66	63
						25	25	24

10.1 Examples of Relationships between Two Variables

A relationship between two variables exists when the value of one depends upon the value of the other. For example, the total cost of soup depends upon the number of cans of soup purchased at the stated price. Figure 10.2a is a chart of the cost of several different purchases of soup. If one can of soup is purchased, the cost is 10 cents; if two cans are purchased, the cost is 20 cents. The cost of different numbers of cans can be read exactly from Fig. 10.2a because for each number of cans bought, one and only one price is paid. Generally, the relationship between two variables is not so exact as this. For example, Fig. 10.2b and c show data on grades in mathematics, statistics, and Latin. The data show that the higher the grade in mathematics, the higher the grade in statistics; but for any particular grade in mathematics, no one grade in statistics can be read from Fig. 10.2b. Similarly, the data of Fig. 10.2c show that the higher the grade in mathematics, the lower the grade in Latin. (These data are hypothetical, and it is not to be inferred that students who are good in mathematics are necessarily bad in Latin.) For any one grade in mathematics, no one grade in Latin can be read from the graph. Both figures suggest that grades of the two subjects compared are related, but that the relationship is different from that for the cost of cans of soup.

The cost of buying different numbers of cans of soup can be expressed in the form of a mathematical equation as well as read from Fig. 10.2a. If Y equals the cost and x equals the number of cans purchased at

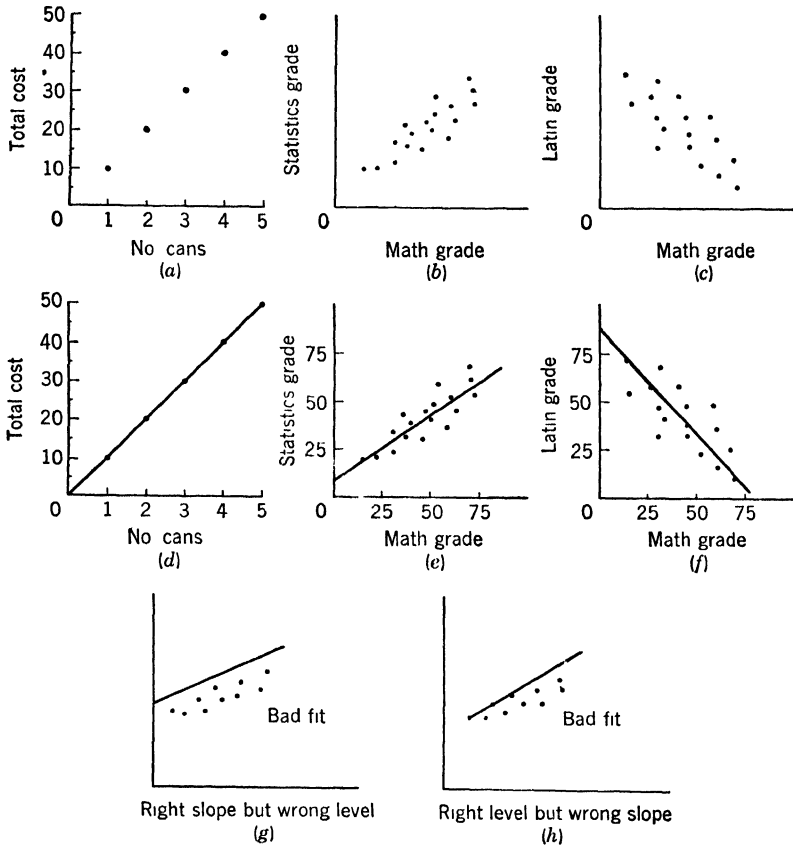


FIG. 10.2. Examples of relationships between two variables.

10 cents per can, the mathematical relationship between Y and x is given by the equation

$$Y = 10x$$

This equation is the mathematical function relating Y and x . It is the equation of the *straight line* which, if drawn on Fig. 10.2a, will pass through each of the points. This is shown in Fig. 10.2d. An equation of this kind presents the relationship between the cost and the number of cans purchased more simply than a chart.

Remark on Functional Relationships. The idea of a function involves the concepts of a relationship between the values of two (or more) variables and the dependence of one variable on the other (or others). For example, the quantity of a commodity

demand in the market is a function of (i.e., related to) the price of the commodity; costs are often a function of output, worker efficiency may be a function of the number of hours worked. Technically, Y is said to be a function of x if specific values of Y correspond to values of x . Some elementary functions studied in algebra are

$$Y = 2 + 2x \quad (\text{a straight line})$$

$$Y = 2x^2 \quad (\text{a parabola})$$

$$Y = \frac{3}{x} \quad (\text{a hyperbola})$$

Values of Y for different values of x for these elementary functions are given in the table below.

$Y = 2 + 2x$		$Y = 2x^2$		$Y = 3/x$	
x	Y	x	Y	x	Y
-3	-4	-3	18	-3	-1
-2	-2	-2	8	-2	-1.5
-1	0	-1	2	-1	-3
0	2	0	0	0	∞
1	4	1	2	1	3
2	6	2	8	2	1.5
3	8	3	18	3	1

In general, $Y = f(x)$ is the notation for a functional relationship between Y and x . It indicates that Y and x are related, but it does not specify the exact mathematical relationship. In functional relationships of the form $Y = f(x)$, x is often called the *independent* and Y the *dependent* variable.

The mathematical relationships between the grades in Fig. 10.2*b* and *c* are not so easily identified as the equation which fits in Fig. 10.2*a*. Writing a mathematical equation to express the relationship is a matter of finding some standard equation which seems to "fit" the data. Both figures show that the relationship is linear, i.e., a straight line. What straight line best represents these data? For the present, the line used is drawn freehand, but later it will be determined algebraically. The line should be neither too high nor too low, and it should have the same general slope as the data. Straight lines are drawn in by hand on Fig. 10.2*e* and *f*, so as to pass through the center of the scatter of points. Examples of lines which are bad fits to data are shown in Fig. 10.2*g* and *h*. One line has the right slope but starts at the wrong level; the other starts at the right level, but has the wrong slope.

From the lines drawn on the figures relating the grades in different subjects, an estimate of the *mean* statistics or *mean* Latin grade can be read for each grade in mathematics. For example, if the mathematics grade is 25, the estimated mean grade is 25 in statistics and 60 in Latin. If the mathematics grade is 75, the estimated mean grade is 60 in statis-

tics and 5 in Latin. The estimate is not the exact grade of any particular student in statistics or Latin, but the *mean* grade of all students with the stated grade in mathematics.

Relationships between two variables need not be in the form of a straight line. Figure 10.3a and b show the relationship between total cost and number of units produced, and price and number of articles sold. In neither figure does a straight line represent the relationship. A curve is more suitable, and a curve is drawn on each graph. Although

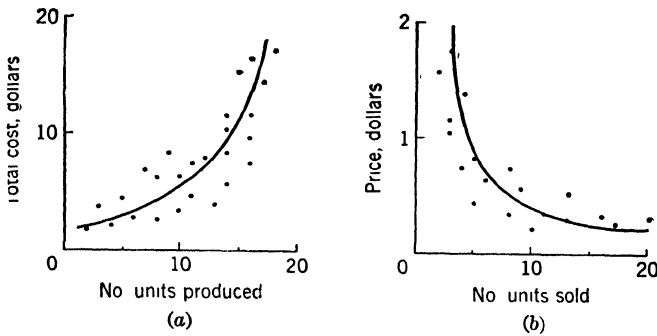


FIG. 10.3. Examples of nonlinear relationships.

curved lines are often necessary to express the relationship between two variables, this text will discuss in detail only relationships expressed by straight lines.

The manager of the cost analysis department faces the statistical problem of finding the mathematical equation of a straight line which best fits the data on man-hours worked and total units produced for the 25 job orders given in Table 10.1 and plotted on Fig. 10.1. This line is called a *regression line*. The method of arriving at a suitable regression line is called *regression analysis*. The general notion of a statistical relationship or association between variables is called *correlation*.

EXERCISES

10.1. The relationship between two variables is given by $Y = 4 + 5x$. Make a table of values of Y for several values of x .

10.2. If the relationship between two variables is given by $Y = 5 - 3.5x$, what value of Y is associated with values of $x = 1, 2, 5$?

10.3. What is the estimated total cost of producing 10 units of product according to the data of Fig. 10.3? Of 5 units of product? Of 15 units? What is the estimated selling price of 5 units according to Fig. 10.3? Of 10 units? Of 15 units?

10.2 Cautions about Relationship between Variables

Man's propensity to associate things with each other leads to countless man-hours of work in computing mathematical equations to represent

relationships. No statistical method is more mechanically applied, and some fields of investigation seem to employ only this one analytical tool. Since interpretation is always more difficult than computation, a few words of caution about possible pitfalls and errors in the use of these methods are in order.

Causation and Association. The occurrence of two events together is often interpreted as evidence that one event caused the other. Naming one variable "independent" and the other "dependent" in regression

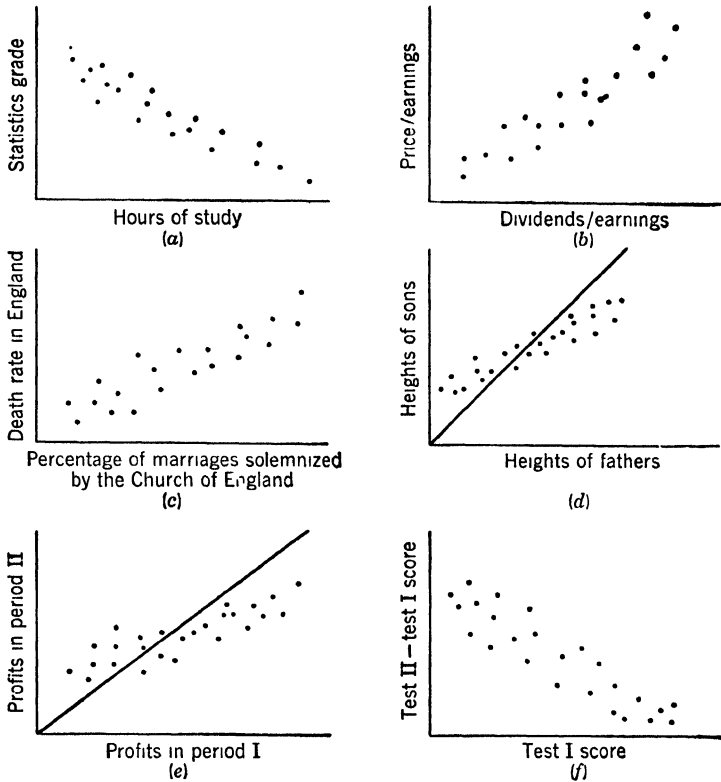


FIG. 10.4. Examples of relationships subject to misinterpretation.

analysis is a carry-over from the mathematics of functional relationships. It is true that values of the dependent variable are determined by substituting values of the independent variable in the functional forms, and that the statistical measure of association between two variables is the regression equation expressed as a particular mathematical function. A statistical association between two variables may be evidence of a real association according to various tests and measurements of the regression analysis, but *statistical association in no way implies causation*. Cau-

sation is based upon other factors. The natural inclination to infer causation from statistical association is a serious pitfall in the application of statistical methods. For example, a correlation between the grades in a course and the time spent in studying might be negative, indicating that the higher grades are received by those who do the least amount of studying (see Fig. 10.4a). This does not imply that the smaller number of study hours *causes* the higher grade. The statistical association is real; the cause and effect interpretation is nonsense. (See succeeding paragraphs in this section for further discussion.)

Correlations with Common Elements. Observed relationships between two variables may be due to the presence of a common element in the variables. This is especially true in the correlation of ratios of which the same variable is a part. The computations in this case are as valid as for any other set of data, but the interpretations are open to serious doubt. In fact, the observation of unwarranted interpretations based upon such relationships prompted Karl Pearson (an early expert in the field of statistics) to label them "spurious correlations."

In practice, the incidence of common elements in variables which are being correlated is not always obvious. For this reason, inexperienced investigators often draw unwarranted conclusions. An outstanding illustration of such a fallacious relationship recently came to the author's attention in a study of the amount of earnings paid out in dividends by a group of companies. The dependent variable was the ratio of dividends to earnings, i.e., the percentage of earnings paid out in dividends. The independent variable was the familiar price/earnings ratio. A resulting positive relationship (see Fig. 10.4b), indicating that a high dividend/earnings ratio was accompanied by a high price/earnings ratio, led to the conclusion that companies which pay out larger shares of their earnings in dividends have stocks which are higher priced relative to their earnings. This phenomenon may be true, but it is an unwarranted conclusion from the regression because *both ratios have the same denominators*. Small earnings tend to make both ratios large, and large earnings tend to make both ratios small.

Correlations Due to Implicit Relationships with other Factors. Correlations between two unrelated variables are often observed when both are related to another factor or perhaps to several factors. A negative statistical relationship between hours of study and grades does not mean that a smaller number of hours causes higher grades, for both variables are related to a third factor, intelligence. The regression equation does not explicitly take this factor into account. High grades are associated with high intelligence; high intelligence is also associated with a smaller number of study hours. The inverse relationship between grades and hours of study is due to the presence of this third factor.

In business and economic problems, the most frequently recurring third factor implicitly responsible for relationships between variables is *time*. A classic example is Professor Yule's observation of a positive correlation between the proportion of marriages solemnized by the Church of England and the death rate in England¹ (see Fig. 10.4c). A low proportion of marriages solemnized by the Church of England coincides with a low death rate and a high proportion of marriages with a high death rate. Here a causal interpretation would imply that suppression of Church marriage ceremonies will reduce the death rate. This is obviously ridiculous. The positive statistical relationship occurs because both variables have been decreasing in England for many years, and *time* is the implicit third factor responsible for the positive relationship.

The Regression Fallacy. The term regression was first applied to the statistical relationship between two variables by Francis Galton in his studies of inheritance. The origin of the expression can best be explained by an example of the relationship between the heights of fathers and the heights of sons. Karl Pearson ascertained the heights of 1078 sons and classified them according to the stature of both son and father.² The relationship reveals that tall fathers tend to have tall sons and short fathers to have short sons, but also that the average height of the sons of a group of tall fathers is less than that of the fathers, and the average height of the sons of a group of short fathers is greater than that of the fathers. In other words, there is a *regression* of the average height of the sons toward the average height of all men. Galton, observing this same phenomenon in many areas, phrased a universal law of regression. The term regression was subsequently employed in statistics to designate a general relationship between two (or more) variables. The occurrence of the regression phenomenon may be misinterpreted. It seems to imply that, after several generations of sons have become fathers, the height of men will tend to a constant average height. This has not happened and probably will not happen, unless the structure of man changes radically.

An example of this regression phenomenon in a business problem is based upon an examination of the profits of a group of firms in two time periods. The more profitable firms made lower average profits in the second period, and the less profitable firms made higher average profits in the second period. The conclusion follows that there is a regression toward mediocrity in business. This is the same phenomenon that Galton observed. The fallacy of the business inference may best be shown by assuming, not unreasonably, that profits have definite upper

¹ G. Undy Yule, Why Do We Sometimes Get Nonsense Correlations? *Journal of the Royal Statistical Society*, vol. 85, p. 95, 1921.

² The data are reproduced in M. G. Kendall, "The Advanced Theory of Statistics," vol. I, p. 327, Charles Griffin & Co., Ltd., London, 1948.

and lower limits. Then, if there is any change at all, the firms with the higher first-period profits will tend to have lower second-period profits, since their profits cannot be larger; similarly, the firms with the lower first-period profits will tend to have higher second-period profits because they cannot have lower ones.

Another example of the operation of this phenomenon emerges from the study of the effect of a stimulus on test scores. Suppose that a group of students is given a test, then a mental or physical stimulus, and finally a second test. The effect of the stimulus is commonly measured by the difference in the two test scores. If a negative relationship is observed in the correlation between this difference and the initial test score for each student, the conclusion often follows that the effect of the stimulus is greater on students with low initial scores than on students with high initial scores. This negative relationship is an example of the regression fallacy. Any random variation in low initial test scores tends to show up in higher final scores, and any random variation in the high initial test scores tends to show up in lower final scores, because the possible scores have lower and upper limits. An initial score near one of the limits has more chance of changing in one direction than in the other.

The regression phenomenon illustrated by these examples is a statistical characteristic of the way in which the data are ordered. Fallacious arguments from these relationships should be avoided.

EXERCISES

10.4. To what factor would you attribute a positive relationship between the salaries of teachers and the consumption of alcohol in the United States over the past 30 years?

10.5. In a recent discussion of the studies of the relationship between cancer and cigarette smoking, someone was heard to say that a positive relationship exists between the incidence of cancer and the wholesale price level in the United States. How could you account for this relationship if it is true? Does a relationship of this kind have any bearing on cancer research?

10.6. Bring to class an example of a relationship between two variables which you believe to lead to a false conclusion or impression. For what reason do you believe that the conclusion is false?

10.3 Universe Regression of Man-hours and Units of Product

The search for a mathematical relationship between two variables is predicated on the idea that they vary together. A *scatter diagram* visually illustrates how they vary together in a sample. The sample observations are plotted on a graph scaled according to the x variable on the abscissa and the y variable on the ordinate. The scatter diagram of the man-hours and units of product is shown in Fig. 10.1. The outstanding

feature of this figure is that high values of y (man-hours) tend to occur with high values of x (units of product), and low values of y with low values of x . In general, the more units of product made, the more man-hours of work required to make them.

Figure 10.1 also shows that the number of man-hours is not always the same for the production of the same number of units. For example, y values for the three values of x equal to 45 are 33, 46, and 48; y values for the two values of x equal to 66 are 63 and 70. In a large sample of more job orders, more values of y will occur at a given value of x . In other words, *a distribution of y values exists for each value of x* . The number of man-hours required to make the same number of units varies from one job to the next, and such a distribution exists for each different size of job order. The distribution of the sample y values for given x values represents the underlying distribution of the population or universe which is sampled.

Here again the basic problem is one of inference. Any relationship in the sample is useful only if it may be projected to the whole universe from which the sample was drawn. The manager of the cost analysis department desires to use the relationship between man-hours and units of product to estimate man-hours for job orders of a given size. This relationship is for all possible job orders from which he has a sample of only 25 observations. The statistical problem is to estimate the parameters of the universe relationship. Before attacking this problem, the statistician must clearly understand the nature of the universe.

In Sec. 1.6, a bowl of 998 chips (Table 1.3) was sampled. Each chip in the bowl was labeled with a value of a variable x . The distribution of the values of x in this bowl was approximately normal with mean $\mu_x = 3.0$ and standard deviation $\sigma_x = 1.0$. The bowl of 998 chips was the universe sampled in that section.

Now imagine a set of several bowls of the same general type. Each bowl is identified by a value of x : $x_1, x_2, \dots, x_i$. Within each bowl is a distribution of values of a variable y . The mean and standard deviation of the distribution of y depend upon the bowl in which the distribution is located, i.e., upon the value of x . The mean and standard deviation are written $\mu(y|x)$ and $\sigma_{y \cdot x}$, which are read: the mean of y given x and the standard deviation of y given x . For example, if the bowl is number x_1 , the values of y within this bowl are distributed with $\mu(y|x_1) = Y_1$ and $\sigma_{y \cdot x_1}$. This distribution is drawn in Fig. 10.5a. If the bowl is number x_2 , the values of y within the bowl are distributed with $\mu(y|x_2) = Y_2$ and $\sigma_{y \cdot x_2}$ (Fig. 10.5b). If the bowl is x_i , the values of y in this bowl are distributed with $\mu(y|x_i) = Y_i$ and $\sigma_{y \cdot x_i}$. In other words, a different distribution of y values exists for each value of x .

The whole collection of bowls or distributions is the universe. The

following assumptions are made about this collection for the universe model of a relationship between x and y :

1. The distribution of y in each bowl is a *normal* distribution.

2. The standard deviation of y is the same for each bowl, that is,

$$\sigma_{y \cdot x_1} = \sigma_{y \cdot x_2} = \dots = \sigma_{y \cdot x_i}.$$

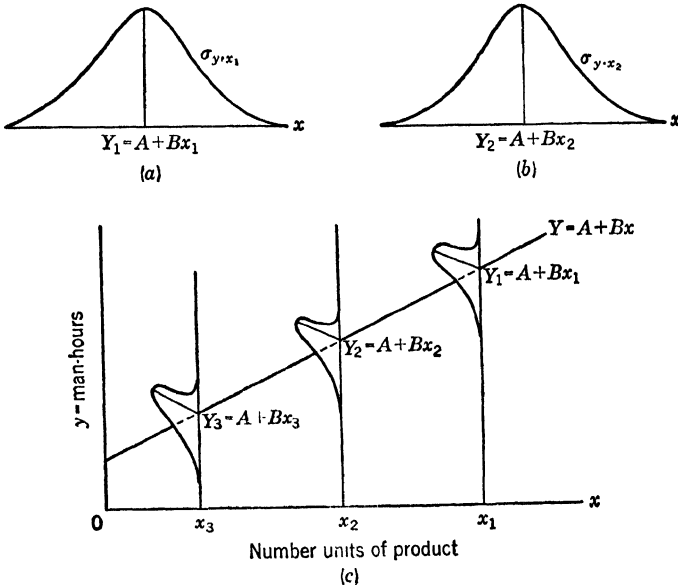


FIG. 10.5. Graphic models of universe regression relationships.

3. The *means* of the bowls are related to each other by a straight-line relationship called the *regression line*,

$$\mu(y|x_i) = Y_i = A + Bx_i$$

Example:

If $x = x_1$,

$$Y_1 = A + Bx_1$$

If $x = x_2$,

$$Y_2 = A + Bx_2$$

The subscript i may be dropped and the model of the universe relationship specified as follows. *values of y are normally distributed with a mean expressed by the relationship $Y = A + Bx$ and a standard deviation $\sigma_{y \cdot x}$.*

In the problem facing the cost analysis department, the values of x are the numbers of units of product; the values of y are the numbers of man-hours required to make these units. The model of the universe regression specifies that:

1. The mean number of man-hours required to make a given number of units of product x is given by the equation $Y = A + Bx$.

2. The number of man-hours required to make x units of product varies about the mean Y in the form of a normal distribution.

3. The standard deviation of this normal distribution is the same for each value of x and is $\sigma_{y \cdot x}$. In other words, the universe relationship is specified as follows: the number of man-hours required to make x units of product is normally distributed with a mean value expressed by the equation $Y = A + Bx$ and with a standard deviation $\sigma_{y \cdot x}$.

A graph of the three-dimensional model is drawn in Fig. 10.5c for three values of x . The distribution function $f(y)$ is a normal distribution; on the graph, it is drawn as if it were in the plane perpendicular to the paper. Figure 10.5c shows the straight line with a positive slope, i.e., both variables increase together. In some instances, one variable increases while the other decreases and the slope of the straight line is negative. The assumptions remain the same; the figure is merely tilted to show the negative relationship.

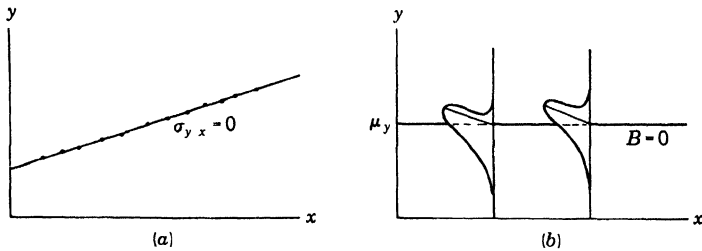


FIG. 10.6. Two limiting models of the universe relationship. (a) Perfect relationship between y and x ; (b) no relationship between y and x .

Two limiting cases of the universe model have interesting implications for the estimation problem. In the first case, $\sigma_{y \cdot x} = 0$. This means that the y values do not vary and hence that they all lie exactly on the regression line. The relationship between x and y is perfect, and the value of y for a particular value of x can be predicted exactly from the regression function (see Fig. 10.6a). In the second case, the regression function is parallel to the x axis with the means of the normal distributions equal to the mean value of y , μ_y . Whatever the value of x , the estimated value of y is μ_y ; no relationship exists between the two variables, and $B = 0$. Accordingly, a knowledge of x is not helpful in predicting values of y (Fig. 10.6b). The data on man-hours and units of product appear to lie somewhere between these two limiting models (see Fig. 10.1).

EXERCISES

10.7. Let a regression universe be specified by $Y = 5 + 10x$ and $\sigma_{y \cdot x} = 1$, with y normally distributed about Y .

- a. Find the mean value of y for $x = 0, 2, 4$, and 6 .
 b. What is the probability that y will lie between 13 and 17 if $x = 1$? Between 63 and 67 if $x = 6$?

- c. Between what values will 90 per cent of all values of y lie for $x = 0$? $x = 4$?
 d. What fraction of all values of y will lie within the band $Y = 5 + 10x \pm 2\sigma_y$?

10.8. Suppose that the population relationship between yearly expenditures on food and yearly income, both expressed in dollars, is defined by the equation $Y = 500 + .10x$, with $\sigma_{xy} = 100$.

- a. What is the average yearly food expenditure for families with incomes of \$1000, \$5000, \$10,000, \$0?
 b. What is the probability that an individual with an income of \$5000 will spend more than \$1300? What assumption do you make to answer this question?
 c. What fraction of individuals with incomes of \$1000 have yearly food expenditures between \$200 and \$800?
 d. Suppose that three individuals are selected at random from this population. What is the probability that all three of them will spend more for food than the average expenditure of persons with their respective incomes?

10.4 Estimating the Relationship between Two Variables

The problem facing the cost analyst is estimation of the values of A and B in the universe relationship $Y = A + Bx$. The straight-line equation computed from the sample to estimate the universe equation will be written $y' = a + bx$. The value of a estimates A ; b estimates B ; and y' is the *mean* value of y for a given value of x which estimates Y . The method of maximum likelihood (used earlier for proportions and means) leads to single-valued or point estimates which are unbiased and efficient.

The probability distribution of the universe is the normal distribution.¹ The maximum-likelihood estimate of A is that value computed from a sample which, if it were the true value of A in the normal distribution, would maximize the probability of observing the sample value which actually is observed. A similar statement holds for the maximum likelihood estimate of B . Again, the derivation of these estimates is a problem for the differential calculus; the estimates are merely stated in the equations

$$b = \frac{\Sigma(x - \bar{X})(y - \bar{Y})}{\Sigma(x - \bar{X})^2} \quad (10.1)$$

$$a = \bar{Y} - b\bar{X} \quad (10.2)$$

where the sample regression line which estimates $Y = A + Bx$ is

$$y' = a + bx \quad (10.3)$$

¹ The normal distribution equation is

$$f(y | x) = \frac{1}{\sqrt{2\pi} \sigma_{y \cdot x}} e^{-\frac{1}{2} \frac{[y - (A + Bx)]^2}{\sigma_{y \cdot x}^2}}$$

and $\bar{X}$ and $\bar{Y}$ are the sample means of x and y , respectively. The computation of b from the sample is easier if the summations in Eq. (10.1) are expanded to involve only the original observations instead of deviations from the mean. Then,

$$b = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2} \quad (10.4)$$

The following quantities are obtained from the data:

Σx : Sum of the x 's

Σy : Sum of the y 's

Σx^2 : Sum of the squares of x

Σxy : Sum of the products of x and y

Partial computations of these sums and the calculation of a and b are

TABLE 10.2. PARTIAL COMPUTATIONS OF THE REGRESSION ANALYSIS*
(a = units of product, y = man-hours)

x	y	x^2	y^2	xy	y'	$(y - y')$	$(y - y')^2$
66	70	4356	4900	4620	64 47	5 53	30 5809
49	46	2401	2116	2254	47 69	-1 69	2 8561
64	65	4096	4225	4160	62 50	2 50	6 2500
45	46	2025	2116	2070	43 75	2 25	5 0625
61	62	3721	3844	3782	59 54	2 46	6 0516
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
61	58	3721	3364	3538	59 54	-1 54	2 3716
34	33	1156	1089	1122	32 89	11	0121
57	49	3249	2401	2793	55 59	-6 59	43 4281
66	63	4356	3969	4158	64 47	-1 47	2 1609
25	24	625	576	600	24 01	- 01	.0001
1331	1297	74,795	71,603	72,935	.	00	482 5400

$$\bar{X} = \frac{\Sigma x}{n} = \frac{1331}{25} = 53.24 \quad b = \frac{25(72,935) - (1331)(1297)}{25(74,795) - (1331)^2} = \frac{97,068}{98,314} = .987$$

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{1297}{25} = 51.88$$

$$y' = -.67 + .99x \quad a = 51.88 - .987(53.24) = -.668$$

* Data are from Table 10.1.

given in Table 10.2. The regression line which is the best estimate of the universe relationship is

$$y' = -.67 + .99x \quad (10.5)$$

This line is drawn in Fig. 10.7.

Interpretation of the Regression Line. The mean number of man-hours required to make a given number of units of product is estimated

by the regression equation

$$y' = -.67 + .99x$$

Each value of y' is the *mean value of y for a given value of x* . In this problem, a value of y' is the mean number of man-hours required to make a given number of units of product x . For example, an estimate of the mean number of man-hours to complete job orders for 50 units is

$$y' = -.67 + (.99)(50) = 48.8 \text{ man-hours}$$

An estimate of the mean number of man-hours for job orders for 65 units is

$$y' = -.67 + (.99)(65) = 63.7 \text{ man-hours}$$

The scatter diagram with the regression line plotted on it shows whether a particular job takes more or less man-hours than the mean (Fig. 10.7). For example, job order 15 with 45 units of product took 33 hours, which is below the mean of 43.7 for 45 units; job order 13 with 45 units of product took 48 hours, which is above the mean.

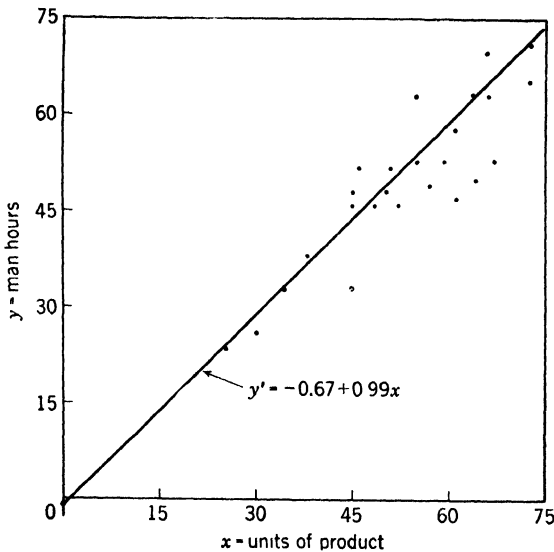


FIG. 10.7. Regression line for man-hours and units of product made.

An interesting question is: How many additional man-hours are required to make one more unit of product? In algebraic terms, b is the slope of the regression line, or the change in y for a unit change in x . Hence, .99 man-hours are required to make one more unit of product. A change in the value of a variable is often represented by the symbol Δ (delta). Then, b has the following meaning: for $\Delta x = 1$ unit of product, $\Delta y = .99$ man-hours.

Finally, what is the interpretation of the value of $a = -.67$ in the regression equation? When $x = 0$, $y' = -.67$. This means that if 0 units of product are made, the mean number of man-hours required to make them is $-.67$ man-hours. This is obviously ridiculous, since a negative number of man-hours is impossible. Even though it has no real meaning here, the value of a is still a legitimate calculation. The occurrence of an impossible value emphasizes a very important caution in the use of regression analysis for estimation or prediction. The values of x range from 25 to 73 units of product. Estimating the mean number of man-hours y' for values of x within this range is called *interpolation*, estimating values of y' outside the range of x (less than 25 or more than 73) is called *extrapolation*. The regression line is computed from a limited range of data, and nothing guarantees that the same linear relationship will hold if values of x less than 25 or more than 73 are observed. The relationship might be entirely different. Using the sample regression line to estimate y' for a value far outside the range of x (like $y' = -.67$ for $x = 0$) is dangerous, as indicated by the occurrence of an impossible value in this problem. Even when the value of a is logically possible, it may not have much meaning if the value of x is outside the range of the data.

Estimating the Standard Deviation. None of the values of y on the scatter diagram (Fig. 10.7) are exactly on the regression line. The observed values differ from their estimated mean value, y' . If the observations are widely scattered about the regression line, the estimates of the mean value of y , y' , for given values of x , are not very precise, i.e., the regression line is not very useful in predicting the mean number of man-hours required to make a given number of units. If the observations lie very close to the regression line, estimates of the mean values of y for given values of x may be quite good. In the population, a measure of the dispersion or scatter of the values of y about the regression equation $Y = A + Bx$ is the standard deviation $\sigma_{y \cdot x}$. The assumption is that $\sigma_{y \cdot x}$ is the same for each normal distribution of y that accompanies a value of x . An estimate of the population standard deviation is the standard deviation of the sample values of y about the sample regression line $y' = a + bx$. Just as the standard deviation of a variable x is measured about its mean $\bar{X}$, so is the standard deviation of y for given values of x measured about the mean value of y for a given value of x which, in this instance, is the regression function $y' = a + bx$. The notation and the equation for estimating this standard deviation is

$$s_{y \cdot x} = \sqrt{\frac{\sum (y - y')^2}{n - 2}} \quad (10.6)$$

Remark. In estimating the standard deviation of a variable x , the formula is $s_x = \sqrt{\sum (x - \bar{X})^2 / n - 1}$, where $n - 1$ is the number of degrees of freedom. One

degree of freedom is lost because one constant $\bar{X}$ is computed from the sample. In estimating $s_{y \cdot x}$, two degrees of freedom are lost: one for the constant a and one for b , both of which are computed from the sample. Thus, the denominator in Eq. (10.6) is $n - 2$ degrees of freedom.

Values of y' are computed for certain values of x in Table 10.2.

Example: (More decimals in a and b are carried for a check.)

$$\begin{aligned} x = 66 & \quad y' = -.668 + (.987)(66) = 64.47 \\ x = 25 & \quad y' = -.668 + (.987)(25) = 24.01 \end{aligned}$$

The deviations of $y - y'$ and their squares are computed to find $s_{y \cdot x}$. From Table 10.2,

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - y')^2}{n - 2}} = \sqrt{\frac{482.54}{23}} = \sqrt{20.96} = 4.58$$

An efficient computational procedure avoids the tedious calculation of deviations from y' by using an expansion of the basic formula.

$$s_{y \cdot x} = \sqrt{\frac{\Sigma y^2 - \frac{a \Sigma y - b \Sigma xy}{n - 2}}{n - 2}} \quad (10.7)$$

Equation (10.7) involves only original values of y and x . From the data of Table 10.2,

$$s_{y \cdot x} = \sqrt{\frac{71,603 - \frac{(-.668)(1297) - (.987)(72935)}{23}}{23}} = \sqrt{20.98} = 4.58$$

The standard deviation $s_{y \cdot x}$ is called the *standard error of estimate*.

EXERCISES

10.9. In a study, The Temporal Stability of Consumption Patterns, *Review of Economic Statistics*, vol. 24, pp. 177-183, November, 1944, W. Allen Wallis reported several linear regressions of various consumption expenditures on income. In 1935, the following relationships were observed for 267 families in Minneapolis and St. Paul, Minnesota.

$$\begin{aligned} \text{Food: } y' &= 191.55 + 0.1994x \\ \text{Clothing: } y' &= -26.59 + 0.1143x \\ \text{Rent: } y' &= 82.14 + 0.1353x \\ \text{Savings: } y' &= -194.33 + 0.1316x \end{aligned}$$

For each of these categories, interpret the values of a and b in the regression equation. Are some values of a logically impossible? Does the knowledge that $x = 0$ is considerably outside the range of the data modify your interpretation?

10.10. The *Survey of Current Business* for April, 1954, carried an article, The Recent Pattern of Consumption. The regression between personal consumption

expenditures in billions of dollars (y) and disposable personal income in billions of dollars (x) based upon data for the years 1922 to 1930, 1937, 1939, and 1940 is $y' = 3.3 + 0.92x$.

- Why were the years 1931 to 1936, 1938, and 1941 and later years omitted?
- What do the values of a and b mean?

c. The range of x is from approximately 40 to 80. Estimate personal consumption expenditures in 1951 when disposable personal income was 225.0 and in 1952 when it was 235.0. How much reliance do you have in these estimates and why? What assumption is made in forecasting by this method? (Note: Actual personal consumption expenditures were 208.1 in 1951 and 218.1 in 1952.)

10.11. The prewar (1929 to 1945) relationship between personal consumption expenditures for food (excluding alcoholic beverages) and disposable personal income is $y' = 0 + .27x$. The postwar (1946 to 1952) relation is $y' = 0 + .24x$. What conclusion do you draw? (The data are from the *Survey of Current Business*, vol. 34, no. 4, p. 11, April, 1954.)

10.5 Testing for the Existence of a Relationship in the Universe

The estimation of the parameters of the universe of man-hours worked and units of product made has proceeded without any test of the existence of a relationship in the population from which the sample of 25 job orders was selected. The sample estimates of a and b and, therefore, the sample regression line will vary from sample to sample because of the chance selection of the sample observations. Even if there is no relationship between man-hours worked (y) and units of product (x) in the universe, the sample could still show some relationship. The model for

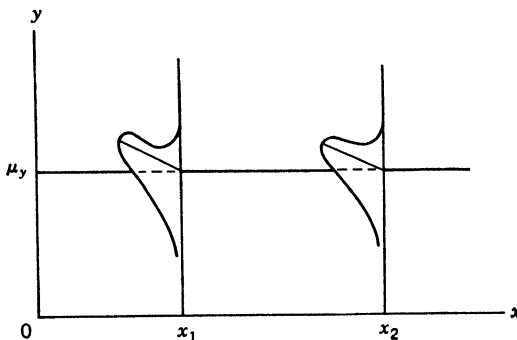


FIG. 10.8. Universe model for no relationship between y and x ($B = 0$).

no relationship between x and y is the limiting case where $B = 0$, discussed in Sec. 10.3 and drawn in Fig. 10.6b. This limiting case is drawn again in Fig. 10.8. The means of the normal distributions of y values (man-hours) for given x values (units of product) are all equal to the average value of y , μ_y , taken without regard to the values of x . The regression line is parallel to the x axis. Whatever the value of x (the

number of units of product), the best estimate of the mean number of hours required to make that number of units is μ_y . One test for the existence of a relationship in the universe is a test for $B = 0$, the model of Fig. 10.8, against the alternative that $B > 0$, the model of Fig. 10.5c.

EXERCISES

10.12. The following questions are a review of the basic concepts of testing hypotheses as they relate to the problem of establishing the relationship between two variables of Table 10.1:

- What is the population sampled?
- What is the sampling unit?
- What are the characteristics of the sampling unit?
- What is an observation of each of these characteristics?
- What is the sample?
- What is the random variable?
- What is the hypothesis to be tested?
- What is the alternative hypothesis?
- What action is to be taken if H_0 is accepted? If H_1 is accepted?
- What is the meaning of error I and error II?
- What are the consequences of making these errors?
- Which of these errors is of the most concern to the cost analyst?
- What is the meaning of alpha? Of beta?
- What does it mean to fix alpha at some preassigned value?
- What is the sample space?
- Where will the region R which leads to the acceptance of H_1 be located in the sample space? Why?

The random selection of sample observations from a universe like Fig. 10.8 ($B = 0$) will result in samples with the b 's positive or negative and not necessarily equal to zero because B is zero. Several scatter diagrams

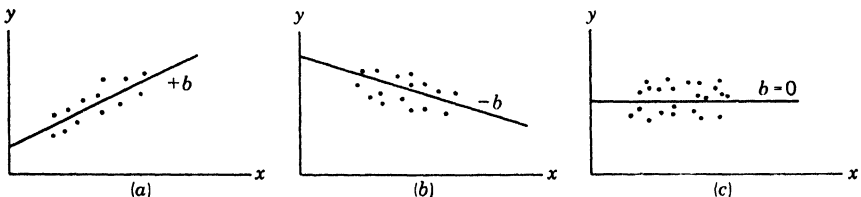


FIG. 10.9. Examples of scatter diagrams with different b 's for samples from the same population.

are shown in Fig. 10.9 for $n = 10$. In Fig. 10.9a, the scatter of randomly selected points shows a positive b , indicating a relationship between x and y in which both variables increase together. In Fig. 10.9b, the random observations indicate a relationship in which one variable decreases as the other increases. Both diagrams are equally probable if $B = 0$ in

the universe. Figure 10.9c shows a scatter in which b is approximately equal to zero. The purpose of these diagrams is to show that the sample value of b varies from sample to sample about the true value of B in the universe, just as the sample percentage varies about the true value of P in the universe and the sample mean about the true value of μ in the universe.

The probability distribution of b is derived mathematically to show how the sample b 's vary from sample to sample. (Sample proportions are distributed according to the binomial distribution; sample means according to the t distribution or the normal distribution, depending upon assumptions about the population.) The probability distribution of b has these characteristics:

1. The mean value of b is equal to the population value B ; in other words, $\mu_b = B$.

2. An estimate of the standard deviation of b is $s_b = s_{y \cdot x} / \sqrt{\Sigma(x - \bar{X})^2}$.

3. The quantity

$$\frac{b - \mu_b}{s_b} = \frac{(b - B) \sqrt{\Sigma(x - \bar{X})^2}}{s_{y \cdot x}} = t \quad (10.8)$$

is distributed according to the t distribution with $n - 2$ degrees of freedom. For a given sample, $s_{y \cdot x}$ and $\sqrt{\Sigma(x - \bar{X})^2}$ are computed from the sample observations. Probability values of the t distribution for several degrees of freedom are given in Table A.4. The formal decision problem is set forth below.

$H_0: B_0 = 0$ (no relationship between man-hours and units of product)

$H_1: B_1 > 0$ (man-hours worked increases with number of units of product)

$\alpha = .05 \quad n = 25 \quad n - 2 = 23 \text{ D.F.} \quad R: \text{upper tail } (b \geq b_r)$

For 23 D.F., the value of t which cuts off $\alpha = .05$ in the upper tail is 1.714. From the sample data of Table 10.2,

$$\sqrt{\Sigma(x - \bar{X})^2} = \sqrt{74,975 - 1331^2/25} = 62.71 \quad s_{y \cdot x} = 4.58 \\ b = .99$$

From Eq. (10.8), since $B_0 = 0$,

$$\frac{(b_r - 0)(62.71)}{4.58} = 1.714 \quad b_r = .125$$

Decision Rule. Accept H_1 if $b \geq .125$; otherwise accept H_0 .

Decision. Accept H_1 and conclude that a positive relationship exists between number of man-hours and units of product because $b = .99$ falls

in the region R . The sample estimate of the linear relationship is $y' = -.67 + .99x$. This regression line can validly be used to estimate or predict the mean number of man-hours required to complete a given number of units of product.

Remark. The decision process described in this section controls the probability of accepting H_1 at .05 when H_0 is true. The probability of accepting H_0 when H_1 is true is taken into account by selecting those values of b which lead to the acceptance of H_1 (that is, the region R) in the upper tail of the distribution. The exact values of β for different values of B_1 under H_1 are not known. They may be computed, however, by selecting various values of B_1 and computing the probability of accepting H_0 for each value. These values of β comprise the OC curve. The numerical details of the OC curve are beyond the scope of this book in practice, but not in principle.

EXERCISES

10.13. In the "Age of Jackson," Arthur M. Schlesinger, Jr., propounds the hypothesis that division of voters into economic classes did not begin with the New Deal but goes back as far as the election of Andrew Jackson. In an article in *Scientific American*, Votes in the Making, November, 1950, pp. 11-13, Prof. Paul Lazarsfeld of Columbia University describes statistical investigations of this hypothesis. The conclusion is that the economic factor played a considerable role in the voting.

These data are taken from the article by Lazarsfeld.

Ward	1836 assessed value per person, hundreds of dollars	1836 Democratic vote, per cent
12	1.7	51
3	2.1	49
1	2.3	53
5	2.4	36
2	3.6	65
11	3.7	35
10	4.7	29
4	6.2	40
6	7.1	34
9	7.4	29
8	8.7	20
7	11.9	23

- Plot the data on a scatter diagram.
- Estimate A , B , and $\sigma_{y \cdot x}$.
- Interpret the values of a , b , and $s_{y \cdot x}$.
- Estimate the democratic vote in a ward with assessed value per person of \$350 and \$1500. In which estimate do you think there is the smaller amount of error and why?

e. Specify exactly the population regression model in this exercise.

f. Test the hypothesis that there is no relationship in the universe, specifying in advance, H_1 and α .

10.14. The elementary units of a population are all stocks traded on the New York Stock Exchange on a given day. Select a random sample of 18 stocks from this population. Record the closing price to the nearest dollar, and measure the change in closing price over the previous day's close in eighth dollars (e.g., $+\frac{3}{8}$ is $+3$). Plot these values on a scatter diagram, estimate the regression equation and standard error of estimate, and test the hypothesis of no relationship between the price of a stock and its closing change in price.

10.15. What is the general shape of the OC curve of the test for the existence of a relationship between man-hours and units of product given in the text? What would a point on this curve mean?

10.6 Prediction Intervals for Regression Estimates

The purpose of the cost analysis department in undertaking a regression analysis of the man-hours of work and the units of product is to determine if they are related so that a prediction can be made about the number of man-hours required for future job orders. These predictions fall into two categories: (1) the *mean* number of man-hours to complete job orders with x units of product; (2) the number of man-hours required to finish a *specific* job order with x units of product. Both prediction problems involve inferences about universe values: the first about a *mean* value $\mu(y|x)$ and the other about an *individual* value of $(y|x)$. The estimates of either $\mu(y|x)$ or $(y|x)$ are subject to sampling variability because of the element of chance in sample selection. Three factors contribute to sampling variability:

1. The variation of the observations about the regression line
2. The variation in the estimated value of b in the regression line
3. The variation in $\bar{X}$, the mean value of x

Prediction of the *mean* y value for a given value of x takes the form of a *prediction interval*, analogous to a confidence interval. For a given value of x , such as $x = x_0$, the prediction interval estimate of $\mu(y|x_0)$ is

$$a + bx_0 \pm t_{c.c.}(n-2) \cdot s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum(x - \bar{X})^2}} \quad (10.9)$$

The values of a , b , $s_{y \cdot x}$, $\bar{X}$, and $\sum(x - \bar{X})^2$ are determined from the sample. The sample size is n ; $n - 2$ is the number of degrees of freedom; C.C. is the confidence coefficient; and $t_{c.c.}(n - 2)$ is the value of t for $n - 2$ degrees of freedom appropriate to the selected confidence coefficient.

For the cost analysis problem, the following quantities are either observed or given:

$$\begin{array}{lll}
 a = -.67 & n = 25 & \Sigma(x - \bar{X})^2 = 3,932.56 \\
 b = .99 & n - 2 = 23 & \text{C.C.} = .95 \\
 s_{y \cdot x} = 4.58 & \bar{X} = 53.24 & t_{.95}(23) = \pm 2.07
 \end{array}$$

$$\text{Prediction interval} = -.67 + (.99)x_0 \pm 2.07(4.58) \sqrt{\frac{1}{25} + \frac{(x_0 - 53.24)^2}{3932.56}}$$

The only variable is the value of x_0 . Table 10.3 gives the 95 per cent prediction intervals for the mean number of man-hours of work required to finish job orders with selected numbers of units of product x_0 . The prediction limits are plotted on Fig. 10.10 and connected by smooth

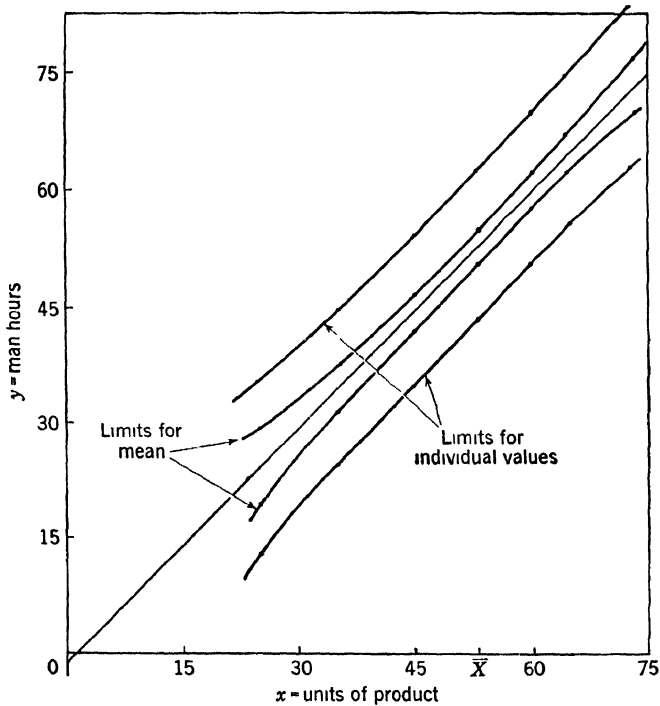


FIG. 10.10. Ninety-five per cent prediction intervals for man-hours of works to make x units of product.

curves to form a prediction band. These prediction intervals are interpreted in the same manner as confidence intervals. The probability is .95 that the interval from 19.4 to 28.7 contains the true mean number of man-hours required to produce job orders with 25 units of product. The probability is .95 that the prediction band contains the true mean number of man-hours required to complete job orders with any number of units of product.

TABLE 10.3. 95 PER CENT PREDICTION INTERVALS FOR MEAN NUMBER OF MAN-HOURS

Units of product x_0	Estimated mean man-hours $y' = -.67 + .99x_0$	Predicted mean man-hours	
		Lower limit	Upper limit
25	24 1	19 4	28 7
35	34 0	30 8	37 3
45	43 9	41 6	46 2
$\bar{X} = 53.24$	52 0	50.1	53 9
60	58 7	56 5	60 9
65	64 7	62 1	67 2
73	72 6	69 1	76 1

Example of calculation: $x_0 = 25$.

$$\begin{aligned}\text{Limits} &= -.67 + (.99)(25) \pm (2.07)(4.58) \sqrt{\frac{1}{25} + \frac{(25 - 53.24)^2}{3932.56}} \\ &= 24.08 \pm 9.181(.49) \\ \text{Limits} &= 19.43-28.73\end{aligned}$$

One characteristic of the interval should be particularly noted. The interval is smallest at $x = \bar{X}$. As the number of units of product becomes either larger or smaller than the sample mean, the prediction interval is wider. The dangers inherent in extrapolating for values of x far away from $\bar{X}$ (often outside the range of the sample data) are clearly indicated by the greater width of the intervals for these values of x . The prediction band is shaped like a hyperbola. It is not, as often implied in elementary texts, bounded by two lines parallel to the regression line.

Prediction of the number of man-hours for a specific job order with x_0 units of product is an interval

$$a + bx_0 \pm s_{y \cdot x} \cdot t_{\alpha/2, n-2} \sqrt{\frac{1}{n} + 1 + \frac{(x_0 - \bar{X})^2}{\sum (x - \bar{X})^2}} \quad (10.10)$$

The only change from Eq. (10.9) for the mean is the addition of +1 under the radical sign. All other data remain the same and are either given or determined from the sample. Table 10.4 gives the 95 per cent prediction interval for the number of man-hours required to complete a specific job order with a given number of units of product x_0 . These intervals are computed from the equation

$$\begin{aligned}\text{Prediction interval} &= -.67 + (.99)x_0 \\ &\quad \pm 2.07(4.58) \sqrt{\frac{1}{25} + 1 + \frac{(x_0 - 53.24)^2}{3932.56}}\end{aligned}$$

The prediction limits for individual values are also plotted on Fig. 10.10. The limits are wider than for the mean number of man-hours. Again

They take the form of a hyperbola and the smallest interval is at the value $x = \bar{X}$. The probability is .95 that the interval from 13.5 to 34.7 traps the true number of man-hours required to produce a specific job order with 25 units of product. The probability is .95 that the prediction band traps the true number of man-hours required to complete specific job orders with any given number of units of product x_0 .

TABLE 10.4. 95 PER CENT PREDICTION INTERVALS FOR SPECIFIC NUMBER OF MAN-HOURS

Units of product x_0	Estimated mean man-hours $y' = -67 + .99x_0$	Predicted individual man-hours	
		Lower limit	Upper limit
25	24 1	13 5	34 7
35	34 0	23 9	44 0
45	43 9	34 1	53 7
$\bar{X} = 53.24$	52 0	42 4	61 7
60	58 7	49 0	68 5
65	61 7	54 8	74 5
73	72 6	62 5	82 7

Example of calculation: $x_0 = 25$.

$$\begin{aligned} \text{Limits} &= -.67 + (.99)(25) \pm (2.07)(4.58) \sqrt{\frac{1}{25} + 1 + \frac{(25 - 53.24)^2}{3932.56}} \\ &= 24.08 \pm 9.481(1.12) \\ \text{Limits} &= 13.46-34.70 \end{aligned}$$

A surprising feature of these prediction intervals is their width. The interval estimate of the number of man-hours required to complete a specific *individual* job order with 60 units of product extends from 58.7 to 68.5 hours. The mean number of man-hours estimated from the regression equation is 58.7. The sampling error added to and subtracted from this mean to form the prediction interval is ± 9.8 , or approximately 10 hours. As a practical matter, is such an estimate accurate enough to be useful in cost analysis? A test of the hypothesis $B = 0$ indicates that a relationship exists in the universe, i.e., that the sample regression line is statistically significant, but statistical significance does not necessarily imply practical significance. It is practical significance that determines the usefulness of the analysis.

The interval estimate of the *mean* number of man-hours required to finish job orders with 60 units of product is from 56.5 to 60.9; the sampling error is ± 2.2 hours. As a practical matter, is an estimate of the mean within ± 2 hours useful to cost analysis? In this problem, the cost analysis department would probably conclude that a relationship between the two variables—man-hours and units of product—does exist, but also

that, on the basis of the present sample with $n = 25$, (1) the sampling variation is too large to permit sufficiently accurate estimates of man-hour requirements for future individual job orders; and (2) the mean number of man-hours for job orders of various sizes can be estimated accurately enough for future use. Convinced that the regression analysis does hold the promise of prediction of man-hours, the cost analysis department would use a larger sample to estimate the relationship and to form prediction intervals.

Remark. For large sample sizes, the value of t for infinity degrees of freedom should be used in Eqs. (10.9) and (10.10).

EXERCISES

10.16. If the decision is to accept H_1 in Exercise 10.13, form the prediction intervals for the *mean* and *individual* values of y . Plot these intervals on the scatter diagram, and interpret them.

10.17. If H_1 is accepted in Exercise 10.14, repeat the directions of Exercise 10.16. Are these intervals of any value in predicting the change in closing price from a knowledge of the price of a stock?

10.7 The Correlation Coefficient

Table 10.5 shows the number of man-hours of work required to produce each of the 25 sample job orders given in Table 10.1. If the manager of the cost analysis department knew nothing about the relationship between man-hours and units of product, he would be compelled to

TABLE 10.5. MAN-HOURS OF WORK REQUIRED FOR 25 SAMPLE JOB ORDERS

70	46	48	26	58
46	68	63	52	33
65	53	48	38	49
46	55	71	63	63
62	52	33	65	21
$\bar{Y} = 51.9$		$s_y = 13.4$		

estimate the man-hour requirements from these data alone. His best estimate of the number of hours of work required for a job order would be the *mean* number of man-hours and a measure of the variation of the individual values about this mean, the standard deviation of the sample, s_y . From the data of Table 10.2,

$$\bar{Y} = 1297/25 = 51.9 \quad s_y^2 = 179.79 \quad s_y = 13.4$$

His estimate would be 51.9 hours for every job order no matter how many units of product were required, with a standard deviation of 13.4 hours.

The analyst's purpose in studying the *relationship* between man-hours and units of product is to make a better estimate of the man-hour requirements than the mean $\bar{Y}$ and to reduce the standard deviation. The relationship between man-hours and units of product is given by the sample regression equation $y' = -.67 + .99x$. A measure of the variation around this regression equation is the standard error of estimate, $s_{y \cdot x} = 4.58$ (see Sec. 10.4). Not only does the regression equation permit a better estimate of the mean number of man-hours for different job orders, but the standard deviation is reduced from $s_y = 13.4$ to $s_{y \cdot x} = 4.58$.

Scatter diagrams for data which show (1) no relationship and (2) a perfect relationship are drawn in Fig. 10.11. With no relationship

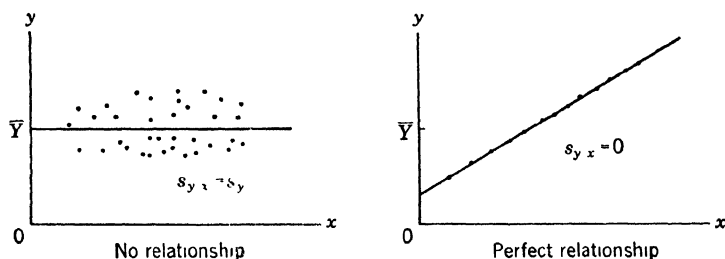


FIG. 10.11. Scatter diagrams showing no relationships and a perfect relationship between y and x .

between y and x , $s_{y \cdot x} = s_y$ because $y' = \bar{Y}$. That is to say, the regression line is horizontal at the mean value of y and the variation around the regression line is the same as the variation around the mean of y . With a perfect relationship between y and x , $s_{y \cdot x} = 0$ because the points do not scatter about the regression line. No relationship between two variables means that knowledge of x does not improve the estimate of y ; a perfect relationship means that knowledge of x reduces $s_{y \cdot x}$ to 0 and improves the estimate of y 100 per cent. A relative measure of the improvement in the estimate of a variable due to the relationship with a second variable is the *correlation coefficient*.

The correlation coefficient is symbolized by r and defined¹ by

$$r = \sqrt{1 - \frac{s_{y \cdot x}^2}{s_y^2}} \quad (10.11)$$

With no relationship between y and x , $r = 0$ because $s_{y \cdot x}^2 = s_y^2$. With

¹ Equation (10.11) is only an approximation. More correctly,

$$r = \sqrt{1 - \frac{n-2}{n-1} \cdot \frac{s_{y \cdot x}^2}{s_y^2}}$$

but if n is reasonably large, the factor $(n-2)/(n-1)$ is approximately unity.

a perfect relationship between y and x , $r = 1$ because $s_{y \cdot x}^2 = 0$. The value of r varies between -1 and $+1$ and takes the sign of b . If b is positive, indicating that both variables increase together, r is positive; if b is negative, indicating that one variable increases as the other decreases, r is negative. When $r = +1$, a perfect positive relationship exists. When $r = -1$, a perfect negative relationship exists. There is no relationship when $r = 0$. The correlation coefficient for the relationship between man-hours of work and units of product is

$$r = \sqrt{1 - \frac{20.98}{179.79}} = .94$$

How is this value of r interpreted?

Equation (10.11) may be solved for $s_{y \cdot x}$:

$$s_{y \cdot x} = s_y \sqrt{1 - r^2} \quad (10.12)$$

A measure of the percentage improvement in the standard deviation may be defined as $(s_y - s_{y \cdot x})/s_y$ —the difference between the standard deviations relative to the standard deviation of y . Substituting the value of $s_{y \cdot x}$ in this measure of the percentage improvement,

$$\frac{s_y - s_{y \cdot x}}{s_y} = 1 - \sqrt{1 - r^2} \quad (10.13)$$

When the relationship between y and x is used instead of the distribution of y alone, the percentage improvement in the standard deviation is directly related to the correlation coefficient. With $r = .94$,

$$1 - \sqrt{1 - r^2} = .66.$$

Thus, the improvement in the standard deviation is approximately 66 per cent when the linear relationship $y' = -.67 + .99x$ is used to estimate the mean number of man-hours instead of the distribution of y by itself.

Table 10.6 shows the percentage improvement in the standard deviation for selected values of r . The reduction in the standard deviation

TABLE 10.6. VALUES OF $100(1 - \sqrt{1 - r^2})$

r	$1 - \sqrt{1 - r^2}$	r	$1 - \sqrt{1 - r^2}$	r	$1 - \sqrt{1 - r^2}$
1 00	100 0	80	40 0	.40	8 4
.99	85 9	.75	33 9	.35	6 3
.98	80 1	70	28 6	.30	4 6
.97	75 7	.65	24 0	.25	3 2
.96	72 0	60	20 0	.20	2 0
.95	68 8	.55	16 5	.15	1 1
.90	56 4	.50	13 4	.10	5
.85	47 3	.45	10 7	.00	.0

remains small and increases very slowly with small values of r . It is only 13 per cent when $r = .5$. For higher values of r , the improvement in the estimate is large and increases rapidly with small changes in the value of r . This table is a valuable device to help interpret the correlation coefficient.

Use of the Correlation Coefficient in Testing for a Universe Relationship. Section 10.5 outlined a test for the existence of a relationship in the universe, involving the hypothesis $B = 0$. An alternative test for the existence of a universe relationship is based upon the correlation coefficient. This test is formally set out below.

H_0 : No relationship between x and y (equivalent to $B = 0$).

H_1 : Positive relationship between x and y (equivalent to $B > 0$).

$\alpha = .05$ $n = 25$ $n - 2 = 23$ D.F.

R : Upper tail (see H_1)

Table A.10 gives values of the correlation coefficient exceeded with probabilities .05 and .01 *if there is no relationship between x and y in the population*.

Examples:

1. For 23 degrees of freedom, the probability is .05 that r will exceed .396 if x and y are not related in the population.
2. For 23 degrees of freedom, the probability is .01 that r will exceed .505 if x and y are not related in the population.
3. Since the distribution of r is symmetrical when there is no relationship between x and y , the probability is .05 that r will be less than $-.396$ and .01 that r will be less than $-.505$ for 23 D.F.

From Table A.10, the value of r in the upper tail corresponding to $\alpha = .05$ and 23 D.F. is .396.

Decision Rule. Accept H_1 if $r \geq .396$; otherwise accept H_0 .

Decision. Accept H_1 because $r = .94$ in the sample falls in the region R . Again the decision is to accept the evidence of a positive linear relationship between man-hours of work and units of product.

EXERCISES

10.18. Compute the correlation coefficient for the data given in Exercise 10.13. Test the hypothesis of no relationship in these data. Interpret the value of r . (Do you agree with Lazarsfeld's conclusion?)

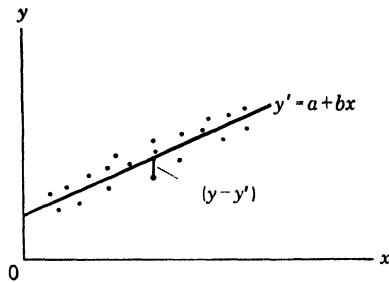
10.19. Repeat Exercise 10.18 for the data of Exercise 10.14.

10.8 The Method of Least Squares

When the plotted values of two variables y and x appear to lie along, although not exactly on, a line on a scatter diagram, a relationship

between the variables may be assumed. Even with no assumption about the normality of the distribution of y for each value of x , or even about the random variation in y , it is often desirable to estimate the mathematical curve expressing the average relationship between the variables. A crude method of estimation is to place a ruler on the scatter diagram and draw a line through the points by eye. Although different people will draw different lines, good preliminary approximations to more elaborate analysis may result from the use of this graphic method. Most widely used in computing a line of "best fit" is the method of least squares.

Suppose that the points on the scatter diagram are as shown below and



that the relationship between y and x is to be represented by a straight line. An indefinite number of straight lines may be drawn through these points. Which one is in some sense best? Whatever line is drawn will be of the form $y' = a + bx$. The deviations $(y - y')$ of the values of y from the line will differ, depending upon the position of the line which is actually drawn. This in turn depends upon the values assigned to a and b . One criterion in determining the values of a and b is to minimize the sum of the squares of the deviations of y from y' : $\sum(y - y')^2$. If this sum is minimized with the calculus, a and b are determined uniquely from the data. This is the method of least squares.

The method of least squares leads to the same values of a and b as the method of maximum likelihood if the values of y are normally distributed about the regression line for given values of x , as in the model of Sec. 10.3. Historically, the method of least squares was developed before that of maximum likelihood, and many statistics books still present it as a method of estimating regression equations without relating it to the maximum-likelihood method of estimation. When the universe models are the same, either rationale of estimation leads to the same results. Without the normality assumption, the method of least squares is widely used in curve fitting, i.e., in passing through a scatter of points a line of best fit in the least-squares sense. The scatter of the observations about the least-squares line may be measured by the standard error of estimate $s_{y \cdot x}$ or by the correlation coefficient r . These, together with the line

$y' = a + bx$, are *descriptive* measures. They describe the line of best fit to a particular set of data. The least squares method is preferable to the more subjective graphic method only because it leads to a unique set of coefficients obtainable by anyone who does the arithmetic correctly.

The literature of statistics contains innumerable least squares estimates of regression equations used for estimating and predicting. The random sampling component and the probability prediction intervals cannot be measured without assuming the regression model specified in Sec. 10.3. Otherwise the worth of the predictions, the usefulness of the least squares line in estimating, and the amount of error in the predictions must be evaluated by judgment. The device is effective in many practical problems only because the analyst has learned to judge correctly from long experience.

10.9 Bivariate Normal Distribution

The bivariate normal distribution is another model of the relationship between two variables x and y . The specification that y and x are *jointly normally distributed* means that the regression of y on x is linear ($Y = A + Bx$) with the y 's normally distributed about the regression line with standard deviation $\sigma_{y \cdot x}$, and that the regression of x on y is linear ($X = A' + B'y$) with the x 's normally distributed about the regression line with standard deviation $\sigma_{x \cdot y}$. Unless the relationship is perfect, the values of A and A' and of B and B' will differ in the two regression equations. A measure of the association between these two variables in the universe is the correlation coefficient, symbolized by ρ (rho) in the population. ρ varies between -1 and $+1$ and has the same meaning for the population relationship as r has for the sample relationship.

Estimates of the regression equations, of ρ , and of the standard deviations may be made by computing the sample values of a , b , $s_{y \cdot x}$, and r , as in the previous sections. The estimates are the same for either model. The bivariate normal distribution has played an important historical role in the theory and application of statistics, though now most, but not all, analyses need only assume a regression model for which the dependent variable is normally distributed about the regression line.

10.10 Some Cautions about the Correlation Coefficient

There is always a difference in statistical work between the use of regression and correlation measures to describe sample observations on the one hand and to estimate universe values on the other. Even when the regression model is applicable, the association between two variables must be verified by statistical test if the sample correlation coefficient is to do anything more than measure the "closeness of fit" of the sample observations to the sample regression line. With no association in the

universe, the sample may still indicate a relationship by a value of r not equal to zero. The sample value of r must be judged in terms of its sampling distribution. The important question is: Could the sample r reasonably occur if there were no relationship between the two variables in the population? In other words, the correlation coefficient must be judged in terms of its probability of occurrence on the hypothesis of no universe relationship. In small samples—and large samples are often impossible or too costly—quite large values of r may occur by chance when, in fact, the two variables are not related in the population.

For example, a random sample of 12 observations drawn from a population with no relationship may yield a sample r as large as .58 with probability .05 and as large as .71 with probability .01. An r of .6 or .7 in a practical problem may appear quite large, but the interpretation may be changed upon consideration of the sample size and the sampling distribution of r . Table A.10 is most useful in inferring a population relationship from the value of a sample correlation coefficient.

Even when the correlation coefficient of a sample is sufficiently large to indicate a relationship in the population, its interpretation may be open to doubt. For example, an r of .5 when $n = 100$ definitely points to a population relationship. But an r of .5 indicates only about 13 per cent improvement in the standard deviation, as interpreted in Sec. 10.7. For estimation purposes, a sample with this value of r is not very useful.

The correlation coefficient for the relation between man-hours and units of product is $r = +.94$. This is large in absolute terms and definitely indicates not only a population relationship between these two variables but also an improvement of 66 per cent in the standard deviation. Even with an r as large as this, the prediction intervals of Sec. 10.6 were so wide as to be practically useless for estimating the man-hours required to finish an individual job order. Although in theory an r of .94 indicates a sample relationship that should be useful for estimating, the sample size and the magnitude of $s_{y \cdot x}$ also affect the practicability of a prediction interval.

If properly interpreted, correlation coefficients are both meaningful and useful in statistics. To achieve this, they must be judged in terms of their sampling distribution and the improvement in the standard deviation. Then, they must be subjected to further doubt in indicating the reliability of estimates by computing the prediction interval and examining its width in practical terms.

10.11 Rank Correlation

The horse-racing problem of Chap. 1 stimulates interest in a further problem: Is there any relationship between the order in which the horses finish a race and the order in which they start the race, commonly called

post position? In other words, does the horse starting on the inside of the track have an advantage over the horse starting on the outside of the track? Data from the Kentucky Derby run at Churchill Downs, Louisville, Kentucky, on May 8, 1952, are used to investigate this problem. The 16 horses in the race are listed by post position and order of finish in Table 10.7. A measure of the relationship between these two variables is the *rank correlation coefficient* denoted by r' (r prime). The rank correlation coefficient is computed from

$$r' = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad (10.14)$$

where d is the difference between ranks and n is the sample size. If $r' = +1$, there is perfect positive correlation of the ranks; the ranks for finish and for post position are 1-1, 2-2, 3-3, . . . , 16-16. In other words, each horse finishes in the same order as the post position. If $r' = -1$, there is perfect negative correlation, or the ranks are paired 1-16, 2-15, 3-14, . . . , the first horse running from the last position, second horse from the second last position, etc. If $r' = 0$, there is no relationship or no correlation between the two ranks. The r' for the Ken-

TABLE 10.7. ILLUSTRATION OF RANK CORRELATION
(Kentucky Derby, Churchill Downs, May 8, 1952)

Horse	Finish	Post position	d	d^2	Position at stretch	d	d^2
Hill Gail	1	1	0	0	1	0	0
Sub Fleet	2	9	-7	49	2	0	0
Blue Man	3	14	-11	121	6	-3	9
Master Fiddle	4	13	-9	81	4	0	0
Count Flame	5	4	1	1	8	-3	9
Arroz	6	16	-10	100	7	-1	1
Happy Go Lucky	7	12	-5	25	3	4	16
Hannibal	8	15	-7	49	5	3	9
Cold Command	9	11	-2	4	11	-2	4
Smoke Screen . . .	10	10	0	0	12	-2	4
Gushing Oil . .	11	7	4	16	9	2	4
Pintor	12	6	6	36	10	2	4
Shag Tails	13	5	8	64	13	0	0
Eternal Moon	14	8	6	36	14	0	0
Brown Rambler	15	3	12	144	15	0	0
Swoop	16	2	14	196	16	0	0
Total			0	922		0	60

$$r' = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6(922)}{16(255)}$$

$$= 1 - \frac{5532}{4080} = 1 - 1.36 = \underline{-.36}$$

$$r' = 1 - \frac{6(60)}{16(255)}$$

$$r = 1 - \frac{360}{4080} = 1 - .09 = \underline{.91}$$

tucky Derby is $-.36$ (from Table 10.7). The minus sign indicates that any existing relationship is negative, i.e., that high values of one variable are associated with low values of the other variable. Does this mean that there is actually a negative relationship? As in other tests, the magnitude of r' is judged by using a *chance model*.

Suppose that there are two rows of 16 boxes each, one row corresponding to order of finish and one to post position. A machine throws a set of 16 balls numbered from 1 to 16 into each row of boxes randomly, one ball to a box. Since the machine knows nothing about horse races, chance alone decides which numbers are paired in the two rows. An r' may be computed for the set of observations generated by this random mechanism. Although r' is expected to equal 0, plus and minus values of r' will actually occur due to chance. The distribution of r' 's resulting from repeating the random pairing until all possible combinations are formed is the *sampling distribution of the rank correlation coefficient*. Critical values of r' are tabled for two one-tailed probability levels in Table A.7. The entries in the table are those values of r' (either positive or negative) which are exceeded with probability .05 and .01 if the population rank correlation ρ' (rho prime) is zero. A ρ' equal to zero means that no population relationship exists between the rank orders. The statistical test takes the following form:

$H_0: \rho' = 0$ (no rank relationship)

$H_1: \rho' \neq 0$ (either + or - rank correlation)

$\alpha = .10 \quad n = 16$

R : two-tailed

Decision Rule. Accept H_1 if $r' < -.425$ or $> +.425$, otherwise accept H_0 (from Table A.7).

Decision. Accept H_0 because $r' = -.36$ does not fall in the region R . The pairings of the ranks observed in the Kentucky Derby of 1952 could have been due to chance, and no relationship between order of finish and post position is established.

Is there any correlation between the order of finish and the order at the beginning of the stretch?¹ Table 10.7 shows an $r' = +.91$ for the Kentucky Derby. The decision is to accept H_1 , that there is a positive rank correlation between the order of finish and the order at the beginning of the stretch. The horses out in front as they enter the stretch tend to be out in front at the finish, as evidenced by this sample.

Uses of Rank Correlation. Rank correlation has wide applicability in the measurement of the relationship between two variables ordered or ranked according to two separate criteria. In many problems, the assign-

¹ The stretch is the straight part of the track just preceding the finish line.

ment of numbers or scores to variables is a natural ordering process. For example, the order of finish and the order of post position in a horse race are numbered from 1 to n . Baseball and football team standings are similarly reported. Very often the standings of students in a school or class are given by their rank: 1, 2, 3, etc.

Variables which are inherently qualitative may also be scored by ordering or ranking methods. Judges in a beauty contest may have no scale for measuring the beauty of the contestants, but instead rank the girls as most beautiful, next most beautiful, etc. Supervisors in a plant must frequently rank their immediate subordinates according to the criteria of efficiency, carefulness on the job, enthusiasm, and the like. In market research, consumers are sometimes asked which brand of a product they like best, which next best, and so on. The point is that a person can frequently tell which of several items he likes best, next best, etc., without actually stating by how much. As soon as he is able to order the items according to two criteria, rank correlation may be used to investigate the existence of a relationship.

In other problems, either the regression analysis of the earlier portion of this chapter or the ranking method may be employed. The choice of technique depends upon the purpose of the analysis. If the problem is one of prediction, rank correlation will not be helpful, since r' does not indicate the nature of the relationship. If the problem is merely to get some idea if two variables are related, the rank correlation coefficient is easier to determine in small samples than the corresponding correlation coefficient r based upon measurements. A high rank correlation coefficient, however, does not necessarily indicate a high correlation between two sets of measurements, because r' does not take into consideration the magnitude of the observations, but only their ranks.

The most important reason for using rank correlation is that *no assumption about the normality of the population distribution is necessary*. Since this assumption of normality, vital to the valid application of the regression model, is often subject to serious doubt in practice, the ranking method is more widely applicable as well as simpler. Rank correlation is another of the distribution-free methods which statisticians are developing in order to lessen the number of assumptions necessary for the application of statistical methods.

Remark. It frequently happens in practice that two or more observations have the same value and consequently their ranks are the same. One method of allocating rank numbers to tied values is to give each of them the average value of the ranks they would possess if they were distinguishable. For example, if the second and third individuals are tied, each gets the rank value $2\frac{1}{2}$; if the fifth, sixth, and seventh individuals are tied, each gets the rank value 6. This method is satisfactory with only a small number of ties. With very many tied values the method breaks down and rank correlation should not be used.

10.12 Multiple Regression

A two-variable regression equation of the form $y' = a + bx$ has been discussed in great detail in this chapter. Relationships among more than two variables are often desirable and statistical methods of handling these problems are available. Although not often included in courses in elementary statistics, the basic ideas of multiple regression and correlation are only an extension of the methods developed for two variables and in principle deserve some attention. The emphasis here is upon the meaning of such regression lines and not upon their calculation or derivation. In fact, no problem is completely worked out, and if faced with a problem, the reader must seek other sources for information on methods.

Data for two variables may be plotted on scatter diagrams with y on the ordinate and x on the abscissa and the regression line drawn as a line which best fits the data. Figure 10.7 shows such a scatter diagram with the regression line drawn on it. Data for three variables must be plotted on a figure with three dimensions. Figure 10.12 shows the points for

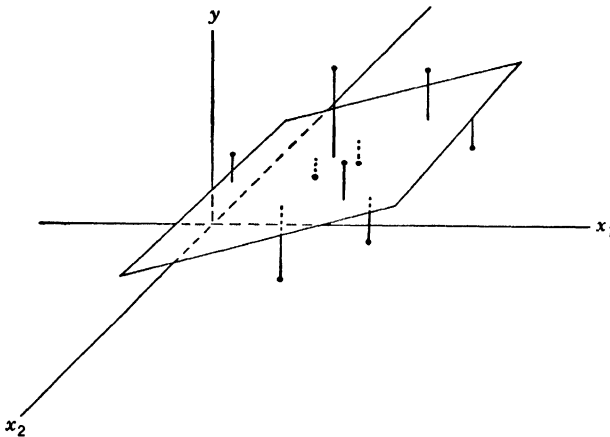


FIG. 10.12. A regression plane in a three-variable regression.

three variables: y , x_1 , and x_2 . The regression equation relating these three variables is of the general linear form

$$y' = a + b_1x_1 + b_2x_2 \quad (10.15)$$

This is the equation of a plane which is drawn on Fig. 10.12. From the sample, estimates are made of a , b_1 , and b_2 . Corresponding to the correlation coefficient r for two variables is a multiple correlation coefficient R , and to the standard error of estimate $s_{y \cdot x}$ is a standard deviation $s_{y \cdot 12}$, where the subscripts 1 and 2 indicate the independent variables x_1 and

x_2 . Confidence-interval estimates of the regression plane and tests of the hypotheses that $B_1 = 0$ and $B_2 = 0$ may be made. All the ideas of simple correlation are readily extended to more than two variables.

As an example of this method, some data are taken from the *Journal of Applied Psychology*, a very fruitful source of regression and correlation applications, especially the latter. In a study concerning predicting success in law school, Berdie and Layton (*Journal of Applied Psychology*, February, 1952) reported the predicting equation for first-quarter grades in law school based upon a sample of 99 students at the University of Minnesota as (the notation of their article differs from that used here)

$$y' = 52.2 + 6.35x_1 + .12x_2$$

where y = fall-quarter grade

x_1 = prelaw honor-point ratio

x_2 = grade on a legal-information test

The multiple correlation coefficient is $R = .49$; the standard error of estimate is not reported. The meaning of the several constants in this equation follow by direct analogy with the two-variable equation. $a = 52.2$ locates the level of the regression plane. $b_1 = 6.35$ means that for each unit change in the prelaw honor-point ratio, the average change in the fall-quarter grade is 6.35 points. $b_2 = .12$ means that for every change in 1 point in the legal-information test, the average change in the fall-quarter grade is .12 points. $R = 0$ is an indication of no multiple correlation and $R = 1$ of perfect multiple correlation, i.e., all the points fall exactly on the regression plane. $R = .49$ is an indication of some relationship and may be interpreted in terms of the improvement in the standard deviation of the estimated y values because of the regression with the two variables x_1 and x_2 .

Multiple relationships are not confined to two independent variables, but may have three or more such variables. A graphic representation in four or more dimensions is not possible and the development is necessarily algebraic. The same article in the *Journal of Applied Psychology* provides a seven-variable equation for predicting first-year law-school grades. The regression equation is

$$y' = 38.2 + .32x_1 + .05x_2 + 3.6x_3 - .38x_4 + .47x_5 + .04x_6 - .22x_7$$

where y = first-year law-school grades

x_1 = first-quarter grades

x_2 = A.C.E. psychological test percentile rank

x_3 = prelaw honor-point ratio

x_4 = score on the mixed relations part of the Iowa Legal Aptitude Test

x_5 = score on the analogies part of the Iowa Test

x_6 = high-school rank

x_7 = score on memory part of Iowa Test

The multiple R using these seven variables to predict the first-year grades is .75.

EXERCISES

10.20. Gather your own data for a rank correlation problem. Compute r' and test its significance from zero.

10.21. The following data are taken from Harold D. Koontz (Domestic Air Line Self Sufficiency, *American Economic Review*, vol. 42, no. 1, p. 112, March, 1952).

RELATIONSHIPS OF SIZE AND OPERATING EXPENSES IN 1949
(Rank order in operating expenses classifications is based upon costs per available ton-mile)

Airline	Available ton-miles	Flying operations	Ground operations	Total operat- ing expenses
American	1	13	10	14
United	2	14	13	11
Eastern	3	15	15	15
Trans World	4	6	14	8
Northwest	5	9	12	9
Capital	6	11	7	13
Delta	7	12	11	12
National	8	16	16	16
Braniff	9	10	8	10
Western	10	7	9	7
Chicago & Southern	11	8	2	5
Mid-Continent	12	4	4	4
Continental	13	5	6	6
Northeast	14	2	3	1
Colonial	15	3	5	3
Inland	16	1	1	2

Compute the rank correlation coefficient between the measure of size, available ton-miles, and each of the three expense classifications. Test the significance of these coefficients. What assumptions are needed for a valid analysis? Are these assumptions met in this problem?

What conclusions would you draw from the rank correlations about economies of size in the airline industry? What other information would you like to have before reaching a definite conclusion about economies of size in the industry?

10.22. The regression of the average annual total expense on gross sales for a given firm for an 8-year period is $y' = 4,000,000 + .78x$ with $r = .95$.

a. What useful information does r provide?

b. A break-even point is defined as the sales which are necessary for a profit margin of zero. What is this firm's break-even point?

c. What other information, if any, would you like to have before using this regression line as a break-even chart?

10.23. The accuracy of the information reported on employment application blanks is an important matter to the personnel director of any firm. A recent study reported in the *Journal of Applied Psychology* investigated the relationship between the claimed and verified weekly wages and duration of employment on a sample of application blank histories. A part of the data of this study is given in the table below.

Number of months job was held prior to application	Male			Female		
	n	r		n	r	
		Wage	Employment		Wage	Employment
0-12	61	.93	.98	65	.94	.98
13-24	21	.98	.99	18	.94	.91
25-36	8	.95	.99	10	.98	.99
37-48	7	.96	.97	10	.89	.99
49-60	4	.97	.99	7	.95	.95
61 and over	8	.90	.98	7	.90	.98

What conclusions do you draw from these data?

10.24. It has recently been argued that the profits of firms in the aircraft industry are outrageously high. Officials of the industry argue that the profits of all manufacturing industries are high and that the aircrafts are not an exception.

The results of three regression and correlation analyses are given below:

1. Net income before taxes and net sales of all manufacturing, 1929 to 1952:
 $y' = -2.4 + .11x$; $r = .97$ (all data in billions of dollars).

2. Profits before taxes and net sales of Lockheed Aircraft, 1935 to 1953:
 $y' = -5.2 + .07x$; $r = .74$ (all data in millions of dollars).

3. Net profits before taxes and net sales of 12 air-frame companies, 1937 to 1953:
 $y' = -8.0 + .07x$; $r = .88$ (all data in millions of dollars).

The data for 1953 are given in the following table:

	Sales	Profits
All manufacturing	265.4	24.4
Lockheed	820.5	48.3
12 aircraft	5120.1	317.1

From all these facts, would you argue for or against the position that aircraft earnings are outrageously high and why?

SUPPLEMENTARY READING

Moroney, M. J.: "Facts from Figures," Chaps. 16 and 18, Penguin Books, Inc., Baltimore, 1953.

REFERENCES

- Burr, Irving W.: "Engineering Statistics and Quality Control," chaps. 5-11, McGraw-Hill Book Company, Inc., New York, 1953.
- Ezekiel, Mordecai: "Methods of Correlation Analysis," 2d ed., John Wiley & Sons, Inc., New York, 1941. This is a complete book on correlation.
- Grant, Eugene L.: "Statistical Quality Control," chaps. 3-11, McGraw-Hill Book Company, Inc., New York, 1952.
- Snedecor, George W.: "Statistical Methods," 4th ed., Iowa State College Press, Ames, Iowa, 1946. Chapters 6, 7, 13, and 14 are good on correlation methods.

CHAPTER 11

THE ANALYSIS OF TIME SERIES: EXAMPLES, POINT OF VIEW, AND TOOLS

A time series, as the name implies, consists of observations of a variable recorded at different points of time. The analysis of such series is particularly challenging to businessmen and economists because many economic data are reported on a monthly, quarterly, or yearly basis. Attempts are made to isolate changes occurring regularly with the passage of time in order to describe what has happened in the past and to predict what may happen in the future.

The tendency of businessmen and economists to recognize and explain such changes is illustrated by their pronouncements in newspapers, magazines, business journals, and speeches. The following are representative quotations:

Total retail trade in February was better than in January on a seasonally adjusted basis.

An analysis of the longer-term trend of earnings of leading manufacturing corporations significantly reveals the extent to which profit margins were squeezed by the pressure of rising labor costs, stiff competition, etc.

Although it is evident that the long-term growth of the economy has been upward, this growth is marked by very serious cyclical fluctuations reaching very high levels in the late 1920's and very low levels in the early 1930's.

11.1 Examples of Time Series

The examples given in Figs. 11.1 to 11.8 below are used only to illustrate very general ideas about time series. It is therefore not necessary to give the original data in tabular form except when the analysis of a particular series is undertaken.

Figure 11.1 shows the population of the United States by decades from 1790 to 1950. The outstanding feature of this graph is the steady and pronounced growth of the population from decade to decade. There are several reasons for this growth: (1) Immigration was a source of population increase, particularly in the early decades. (2) The self-regenera-

tive force of the existing population increases the total number of people. (3) Better food and medical service increase life expectancy. A steady and persistent change in one direction over a long period of time is called a *trend*. The population of the United States has consistently shown a strong upward trend.

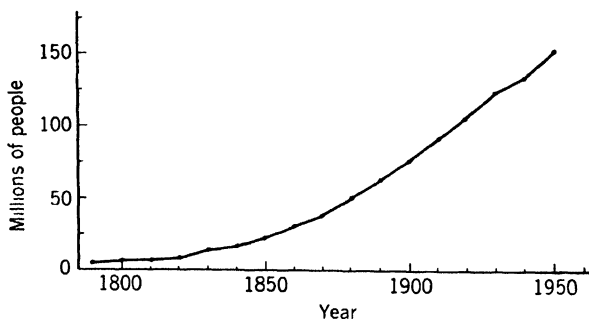


FIG. 11.1. Population of the United States by decades. (SOURCE: "Statistical Abstract," U.S. Department of Commerce.)

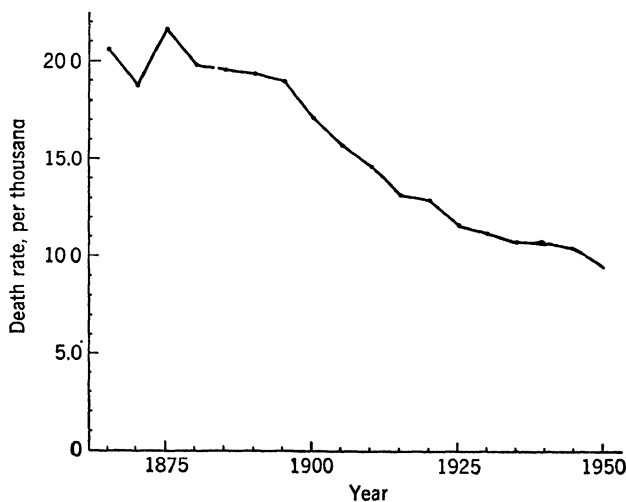


FIG. 11.2 Death registration rates, 1865–1945. (SOURCE: "Statistical Abstract," U.S. Department of Commerce.)

The trend of a series is not always upward but, as illustrated in Fig. 11.2, may be downward. The death rate in the United States has slowly declined between 1865 and 1950, owing to some of the very same factors which caused the population to increase—better food and medical service. With the exception of the irregularity in the initial observations in 1865 and 1870, the death rate declined in every period. The trend of these data is downward.

A trend may be upward in one part of a series and downward in another part. A notable example of this appears in Fig. 11.3, showing the total number of horses on farms in each year from 1900 to 1953. The horse population increased from 1900 to roughly 1917 or 1918, and has been declining ever since. The steady persistent movement is first a gradual increase and then a gradual decline over a long period of time. This

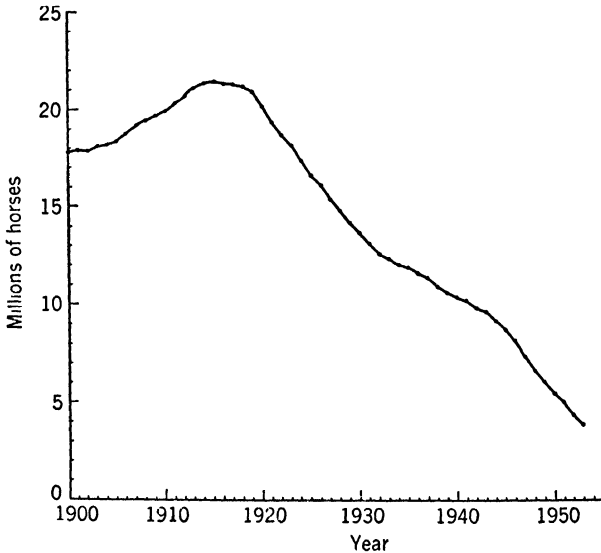


FIG. 11.3. Number of horses on farms (in millions). (SOURCE: "Economic Almanac," National Industrial Conference Board.)

trend is due to the replacement of farm horses by tractors and other mechanized vehicles.

Figure 11.4 shows the revised Consumer Price Index from 1913 to 1953. The pattern of this series is different from that of the series presented heretofore. Over the whole period, the price index rose from a level of approximately 40 to a level of approximately 120, but not by a gradual increase from year to year. On the contrary, some rather violent oscillations marked the pattern of general upward movement. A peak in 1920 was followed by a trough in 1933, and then the upward movement resumed. Such an oscillation, occurring over a period of several years, is commonly called a *cycle*. Whether the cycle which started in 1933 will be completed by a downturn from the recent level remains to be seen. The cyclical movement itself is not a smooth curve, but is subject to variations called *random or irregular* movements. The Consumer Price Index series exhibits several characteristics: an upward trend compounded with a cycle and some random variations.

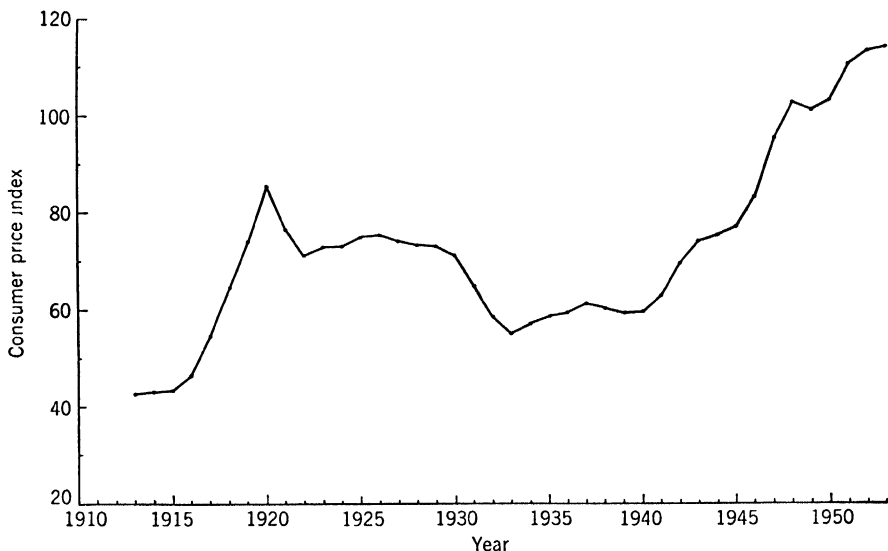


FIG. 11.4. Consumer price index, revised series (1947-1949 = 100). (SOURCE: *Bureau of Labor Statistics*.)



FIG. 11.5. Security price index (1935-1939 = 100). (SOURCE: *Standard & Poor's Corp.*)

The security price series graphed in Fig. 11.5 has a very definite oscillation or cycle. The peak of 1929 was followed by the trough of 1933. Recovery from the depression sent security prices up to another peak in 1936 and 1937, but the outbreak of World War II brought them down to a trough in 1941 and 1942. Still another peak at the end of the war, in 1946, was followed by a decline ending with a low in 1949 and then by a

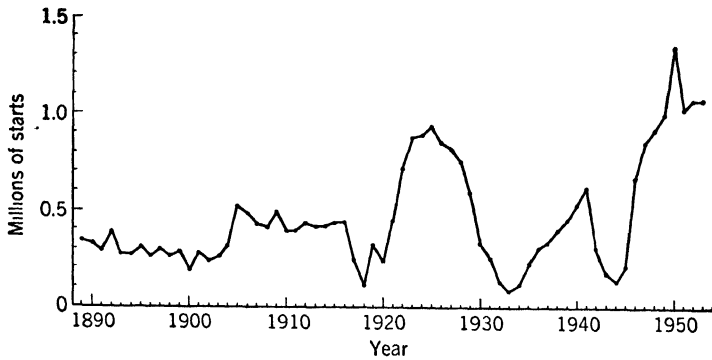


FIG. 11.6. Number of new permanent nonfarm dwelling units started. (SOURCE: David M. Blank, "The Volume of Residential Construction, 1889-1950," p. 67, *National Bureau of Economic Research*, 1954 for the years 1889-1950; 1951-1953 data are from U.S. Department of Commerce.)

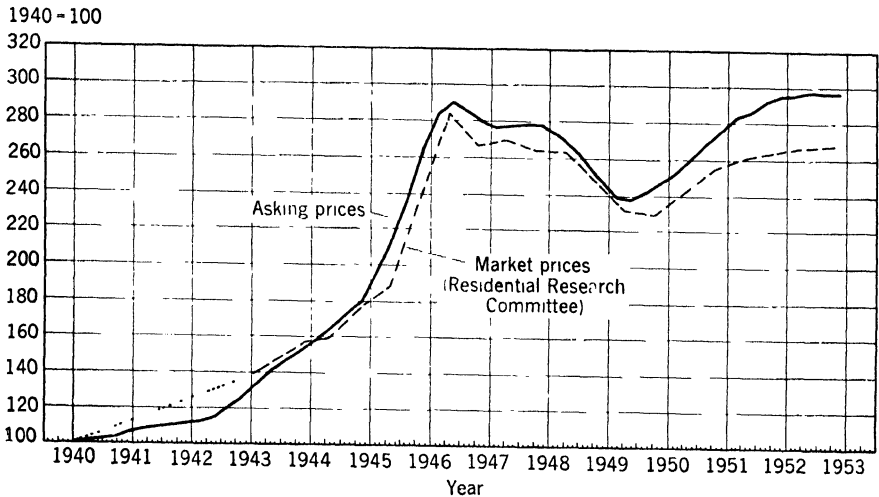


FIG. 11.7. Asking-price and market-price indexes for single-family dwellings, Los Angeles, 1940-1953. (Reprinted from Robert M. Williams, *An Index of Asking Prices for Single Family Dwellings*, *The Appraisal Journal*, vol. 22, no. 1, pp. 33-38, January, 1954, with the permission of the author and the American Institute of Real Estate Appraisers.)

renewal of the upward movement of the cycle. This index does not reveal a trend for stock prices over the entire period, but does show that random or irregular variations marred a smooth cyclical pattern.

The data on residential construction of Fig. 11.6 exhibit a fairly regular cyclical movement. The first trough (in 1900) was separated from the second (in 1918) by 18 years, and the second from the third (in 1933 or 1934) by about 16 years. Students of business cycles point with pride

to the cycle in residential construction, with a period of approximately 15 to 20 years. Irregular variations are apparent in the whole series.

Figure 11.7 shows the sharp increase in asking and market prices of single-family dwellings in the Los Angeles area from 1940 to 1946, followed by a leveling off and an oscillation from 1947 to 1953. The data preceding this figure pertain to the United States as a whole. The study of regional patterns in time series is also an important and interesting phase of economic analysis.

Finally, Fig. 11.8 shows department-store sales from 1919 to 1941. This period between two world wars, frequently regarded as representative of "normal" conditions in the United States, was a happy hunting ground for economists. Department-store sales show a slight long-term

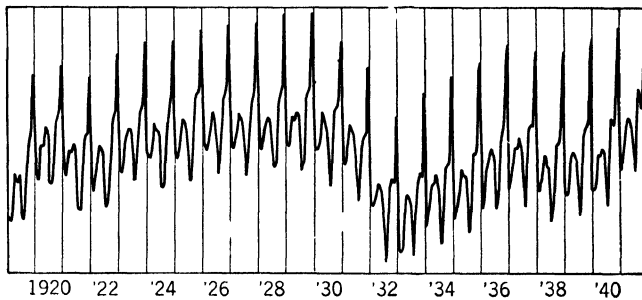


FIG. 11.8. Department-store sales in the United States, 1919 to 1941. (Reprinted from Arthur F. Burns and Wesley C. Mitchell, "Measuring Business Cycles," p. 45, National Bureau of Economic Research, 1946, with the permission of the Bureau.)

upward trend and a cyclical movement around this trend. Another variation present in this series is a fairly regular oscillation within each year which occurs regardless of the general level of trend or the position of the cyclical movement. The department-store sales reach a low near the middle of each year and a peak toward the end of each year. A regularly recurring variation of this kind is called a *seasonal* variation.

A survey of many time series (including those presented in this chapter) suggests that any one series may consist of several components. Although there is disagreement as to the number and terminology of these components, it is usual to assume the existence of four. A series need not exhibit all of them. These four traditional components of a time series are:

1. A *trend* or gradual long-term movement
2. A *cycle* or oscillation which recurs over periods longer than a year
3. A *seasonal* movement or oscillation which recurs regularly each year
4. A *random* or irregular movement

The trend, cycle, and seasonal movement are called *systematic identifiable* forces as contrasted with the *random* forces. The analysis of time series

is the analysis of systematic and random forces like those appearing in the data just examined. The purpose of the analysis here is to present a point of view regarding time series, to gain some insight into the fundamental problems, and to examine several methods of analysis. Other more searching methods that are being developed or are now available are so difficult in computation and theory as to lie beyond the scope of this book.

EXERCISES

11.1. If the cyclical movement in building starts (Fig. 11.6) had continued with its regular period of from 15 to 20 years, when would the next low point have occurred following the trough of 1933? When did the next trough occur? Why did it occur then? What caution can you see in this for the practice of projecting historical data into the future?

11.2. The growth of the population in Fig. 11.1 seems to be fairly regular. What estimate do you make for the population of the United States in 1960 based upon these data? How much reliance would you place in this estimate?

11.2 Semilogarithmic Graph Paper

Suppose that a time series takes on the values shown in Table 11.1. It is interesting to analyze the changes in values of y for changes in

TABLE 11.1. ABSOLUTE CHANGES, RATIOS, AND PERCENTAGE CHANGES

t	y	Δt	Δy	Ratio of each y to the preceding value of y	Percentage change in y relative to initial value of $y(\Delta y/y)$
0	1				
1	2	1	1	2	$1/1 = 100$
2	4	1	2	2	$2/2 = 100$
3	8	1	4	2	$4/4 = 100$
4	16	1	8	2	$8/8 = 100$
5	32	1	16	2	$16/16 = 100$
6	64	1	32	2	$32/32 = 100$

values of t in such a series. The symbol Δ (delta) is used to indicate a change in the value of a variable. Thus, as t changes from 0 to 1 ($\Delta t = 1$), y changes from 1 to 2 ($\Delta y = 1$). As t changes from 1 to 2 ($\Delta t = 1$), y changes from 2 to 4 ($\Delta y = 2$). As t changes from 5 to 6 ($\Delta t = 1$), $\Delta y = 32$. The change in y accompanying a unit change in t is

not constant, but increases with each succeeding Δt . These increases bear a definite relationship to each other: each succeeding Δy is twice the preceding Δy , or the ratio of any Δy to the preceding Δy is always 2. The percentage change in y for unit changes in t is always 100 per cent. Table 11.1 shows these ratios and the equivalent percentage changes.

These data are plotted on ordinary arithmetic graph paper in Fig. 11.9, where changes in y are absolute changes and the increase in y for successive unit changes in t is clear. The graph shows vividly that the change in y is larger for each successive unit change in t . Figure 11.10 shows the same data plotted on a different kind of graph paper, called semi-

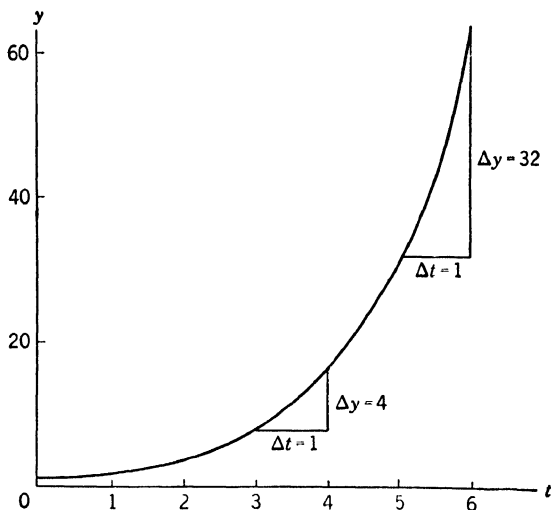


FIG. 11.9 Absolute changes on arithmetic scales.

logarithmic because the vertical scale is ruled according to the logarithms of numbers rather than the numbers themselves. A plotted value of y on this figure is equivalent to the logarithm of y . The data plotted on semilog paper fall on a straight line. The reason for this is apparent from an analysis of changes in y accompanying unit changes in t . A change in y , Δy , is now the difference between the logarithms of two values of y . The difference between the logarithms of two numbers is equivalent to the ratio of the two numbers.¹ From Table 11.1, the ratio of each y value to the preceding y value is always the constant value 2. This constant ratio is the slope of the straight line on semilog paper.

¹ Logarithms do not properly belong in this course of study, but may be reviewed in any algebra book. Here it is sufficient to know that

$$\text{logarithm } \frac{a}{b} = \text{logarithm } a - \text{logarithm } b$$

That is to say, two numbers may be divided by taking the difference between their logarithms.

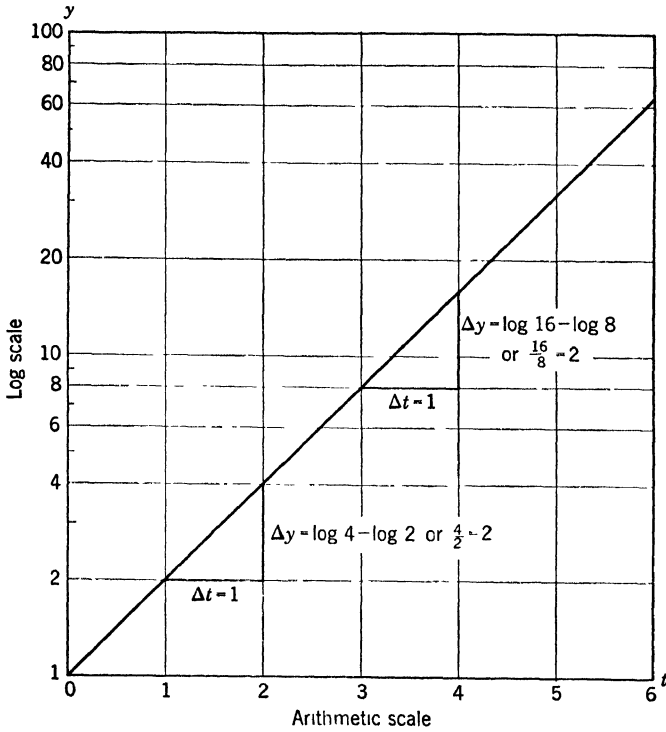


FIG. 11.10. Changes as ratios on semilogarithmic paper.

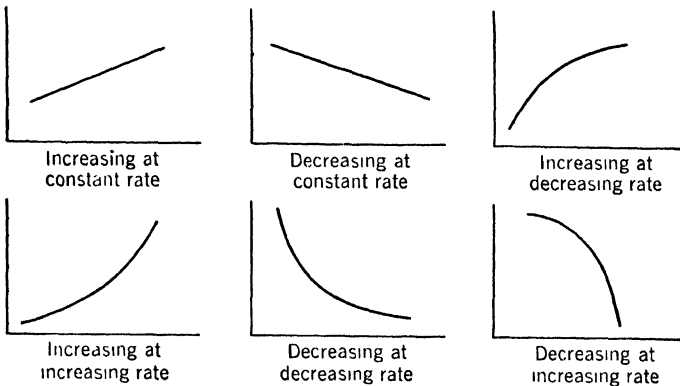


FIG. 11.11. Rates of changes on semilogarithmic paper.

In general, the slope of a line plotted on semilog paper reflects the percentage change in the values of one variable for unit changes in the other variable. These percentage changes are often spoken of as the "rate of change." Since the slope of the straight line of Fig. 11.10 is positive and constant, the series is *increasing at a constant rate*, i.e., a constant per-

centage. The interpretation of changes in other series is given in the diagrams of Fig. 11.11.

Scaling Semilogarithmic Paper. Semilogarithmic graph paper is generally divided into one, two, or three cycles, with each cycle a repetition of the previous one. The horizontal scale is labeled arithmetically; the vertical scale is labeled by starting with any *positive* number. No zero or negative numbers are allowed on the vertical scale, and the number at the top of each cycle must be ten times the number at the bottom. Intermediate scale values follow naturally.

Several different scales are shown in Fig. 11.12. If the paper has two cycles, the second cycle starts at the top of the first cycle and is numbered in the same manner.

Measuring Percentage Changes on Semilog Paper. The ratio or the percentage change of two numbers may be read directly from the original graph with the aid of an auxiliary scale on another small strip of semilog paper. Using the original scale, a value is arbitrarily selected as a reference point (the value is 4 in Fig. 11.13) and labeled 0 on the scale of percentage change and 100 on the percentage scale. The distance from 4 up to 8 on the original scale is equivalent to a 100 per cent increase or a percentage of 200; the distance from 4 down to 2 is equivalent to a 50 per cent decrease or a percentage of 50. The small auxiliary strip of semilog paper is labeled with these percentage changes or ratios. This auxiliary scale may readily be used to measure relative values on the original graph by placing either the 0 or the 100 opposite the first number and reading the percentage change or percentage opposite the second number.

Percentage Changes in Population. The population of the United States by decades is plotted on semilogarithmic paper in Fig. 11.14. The plotted points for values from 1790 to 1860 fall almost on a straight line, indicating an almost constant rate (or percentage) of increase in population. From the scale of percentage change (to the right of the population data), the population increased by about 35 per cent in each decade. The slope of the population graph begins to level off in 1860, indicating that, although the population continued to increase each decade, the

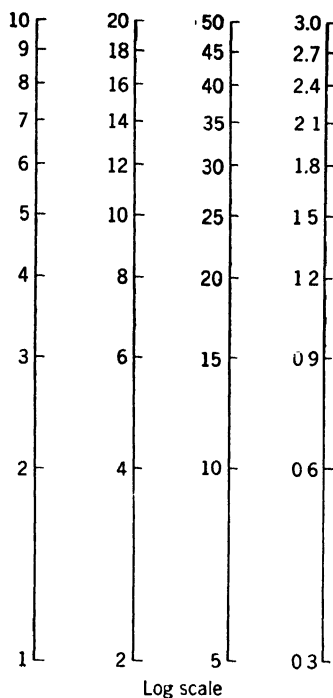


Fig. 11.12. Illustration of different logarithmic scales.

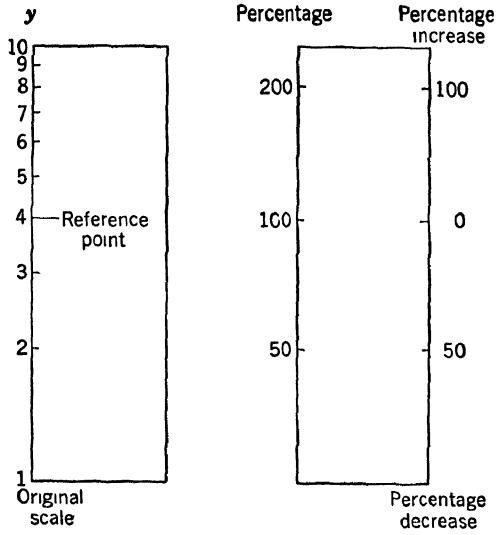


FIG. 11.13. Percentage scales on semilogarithmic paper.

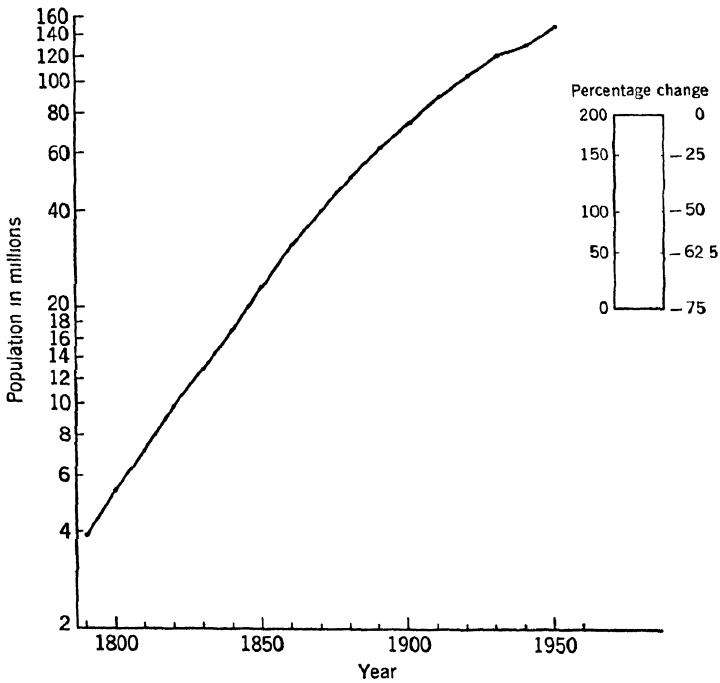


FIG. 11.14. Population of the United States by decades plotted on semilogarithmic paper.

percentage increase was becoming smaller. In the decade from 1900 to 1910, the population increased by about 21 per cent and from 1940 to 1950 by about 14 per cent.

When interest centers directly on percentage changes (either increases or decreases), semilogarithmic charts are most helpful in graphically depicting changes in the rates of increase or decrease and in estimating the actual percentage changes.

EXERCISES

11.3. *a.* Make up a table like Table 11.1 showing the absolute changes, ratios, and percentage changes in the following data:

t	0	1	2	3	4	5	6
y	1	3	9	27	81	243	729

b. Plot these data on three-cycle semilogarithmic paper. Examine changes in y for unit changes in t . Why is the graph of these data a straight line on semilogarithmic paper?

c. Make an auxiliary scale of percentages and percentage changes. Measure some percentages and percentage changes with this scale.

11.4. Data on television-set production are given below. Plot these data on arithmetic graph paper and on semilogarithmic paper. (How many cycles are necessary?) Measure the percentage change in set production from one year to the next with the aid of an auxiliary scale made for that purpose.

Year	Number of sets (thousands)
1946	7
1947	200
1948	975
1949	3029
1950	7464
1951	5385
1952	6096

11.3 The Point of View

The fourth of the traditional time series components listed in Sec. 11.1 is the random or irregular variation. "Random" variations are those resulting from a myriad of unpredictable and partially offsetting forces—the mass of small independent factors that may affect any value in a time series. A systematic component is one that can be identified from this mass of factors subsumed under the heading of random forces. When these systematic identifiable forces display slow changes over a long period, they are called *trend* forces. The identifiable forces may also

show regular periodic oscillations of sufficient magnitude to be distinguished from the random forces. If the period is one year in length, i.e., if the movement recurs regularly each year, the periodic oscillation is called *seasonal*. If the period is longer than one year, economists contend that another set of forces acts with sufficient effectiveness and regularity to be identifiable and predictable, and therefore to be removed from the category of random forces. These are called *business cycle* forces.

In the analysis of a time series, *the first task is to determine whether the forces present in a series are sufficiently distinct and powerful to be identifiable above the mass of minor and offsetting random forces*. If such forces are detected, the next step is to ascertain their nature, origin, and pattern. Any analysis that does not make a careful test for a systematic force will be trying to identify every essentially nonidentifiable force and to explain every fluctuation arising from random forces. At best the result will be a description of chronological events rather than an identification of relevant systematic forces. At worst, it will lead to an enormous waste of time and effort and a fallacious belief that the forces determining the course of a time series can be explained and predicted.

A time series expresses a relationship between two or more variables, one of which is time. The general ideas about relationships and some of the analytical methods discussed in Chap. 10 therefore are applicable in time series analysis. Two limiting models (analogous to the models of perfect correlation and no correlation in Chap. 10) may be hypothesized to describe the mechanism which gives rise to time series data. One model assumes that the economy is completely deterministic and that future values of an economic time series may be predicted with certainty from knowledge of past historical data. (This model is analogous to perfect correlation.) The other model presupposes that the economy is completely random and that only the average value of the series and the variability of the observations about that average can be measured. Knowledge of the order of occurrence of the observations in time is of no assistance in prediction. (This model is analogous to no correlation.)

Economic observations occurring in practice are probably influenced by both systematic identifiable forces and random forces. If the past history is known, it may be possible, under some circumstances, to predict with uncertainty or probability. In other words, the model for an actual series is in between the two limiting models of a completely deterministic economy and a completely random economy. Both systematic identifiable forces and random forces are present. (This model is analogous to a correlation between zero and one.)

In a test for the presence of systematic identifiable forces among random forces, the natural alternative is the random model with no systematic identifiable forces. The two alternative hypotheses will then be:

H_0 : The economy is a random process with no systematic identifiable forces.

H_1 : Systematic identifiable forces are present among the random forces in the economy.

Acceptance of H_0 implies that underlying forces, if present, cannot be estimated *from the data themselves*. The existence of a trend, cycle, or seasonal variation must be based upon other information, for a mechanistic point of view on historical time series is not possible. If H_1 is accepted, the historical data may be examined for systematic components and estimates of the identifiable forces used for inferences about the forces giving rise to the data and for prediction. For further illustration of these alternative hypotheses, a short digression to a discussion of control charts is here interposed.

11.4 A Digression to Control Charts

Control of the quality of mass-produced goods is one of the most fruitful fields for the application of statistical methods. The *control chart*, based upon the ideal of uniform quality, is a simple but extremely valuable statistical tool. Since independent variations in raw materials, machine operators, machine settings, tool settings, etc., cause deviations from the desired uniformity, the actual quality values form a frequency distribution about the mean quality level. If the manufacturing process is operating ideally or, in technical language, is in a *state of control*, the distributions of quality values should be the same at different times in the production process; variations in individual items about the mean quality level are natural to the process and are called *allowable variations*. Any other systematic variation superimposed upon the allowable variation means that a fault exists in the process and can be identified. For example, a tool may wear down and thus cause the production of items whose size differs from acceptable quality standards. The change will show up in the frequency distribution as a change in the average level, in the variability about the average, or in both.

The control chart is used to detect these systematic identifiable forces or *assignable causes*, as they are called in quality-control terminology. If the process is in a state of control, the variability in values in small samples taken at regular time intervals and examined for a quality characteristic such as per cent defective (p) or average value ($\bar{X}$) may be predicted. If such variability differs markedly from the prediction, an assignable cause is assumed to be operating.

The model for a process in a state of control is an unchanging universe of values similar to the bead box or the bowl of chips which were sampled in experiments early in the book. The charts on which were plotted, in

order of appearance, sample numbers and sample averages in repeated samplings of the bead box and the bowl were essentially control charts (see Sec. 1.2 and Sec. 1.7).

The experiment in sampling the bead box is repeated here to illustrate the construction of a control chart. Each red bead now represents a unit of product which is defective according to a selected quality criterion; each white bead represents an acceptable unit under the same criterion. The production process is in a state of statistical control when it turns out 20 per cent defectives. (This particular percentage was selected because the bead box is filled with 20 per cent red beads. So large a percentage of defectives would scarcely be tolerated in actual production.) Each sample of 10 beads from the bead box represents an hourly sample from the production line. The number of red beads in the sample is equivalent to the number of defects. The control chart is made by plotting the number of red beads (defects) on the ordinate and the time interval or the order of the sample selection on the abscissa. The number of red beads will differ from sample to sample; the binomial distribution gives the probability of occurrence of each value if $P = .2$. As long as the variation in the sample values is consistent with $P = .2$, the process is said to be in a state of control. If the variation is such as to induce doubt that the process is in a state of control, the process is halted so that the assignable cause may be detected and remedied. The problem is: How should limits be set upon the variability so as to distinguish between sample values consistent and inconsistent with the hypothesis of a statistical state of control?

Since the control chart is a decision process, it may formally be set out as follows:

H_0 : Process is in a state of control.

H_1 : Process is out of control or assignable causes are operating.

In statistical terms, these hypotheses become:

H_0 : $P_0 = .2$ (for this assumed process)

H_1 : $P_1 > .2$ (assignable causes operate to make more than 20 per cent defectives)

Errors of decision are possible and have the following significance:

I: Accept the process as out of control and look for an assignable cause when, in fact, the process is in control.

II: Accept the process as in control when, in fact, it is out of control.

The standard procedure in determining control limits is to base them upon the probability of making error I, i.e., of looking for an assignable cause when, in fact, none is present. Since the alternative in this problem is one-sided, only an upper control limit will be set. From Table

A.2, the probability of observing 6, 7, 8, 9, or 10 occurrences in a sample of 10 selected randomly from a population with $P = .2$ is about .006. This means that, if a limit is drawn between 5 and 6 and an assignable cause suspected whenever a point lies above the limit, the probability of looking for an assignable cause when the process is actually in control is at most .006. Such a limit is drawn on the graph in Fig. 11.15. The observations in the first 25 samples plotted on the graph are all below the upper control limit, and the decision after each sample is that the process is in control. Beginning with the twenty-sixth sample, 10 red

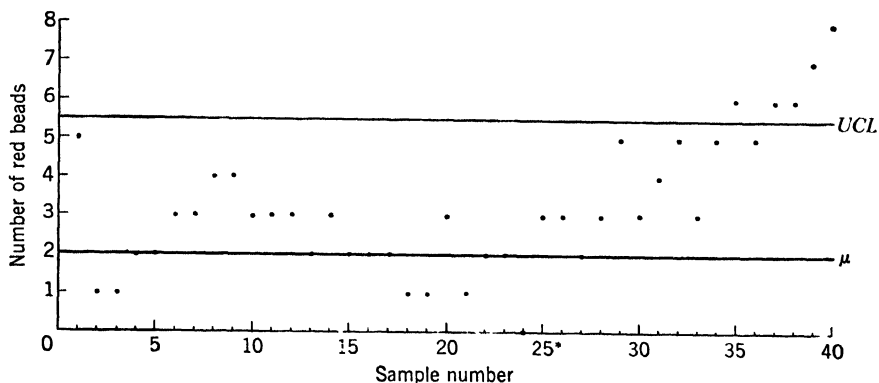


Fig. 11.15. Control chart for bead box containing 20 per cent red beads ($\alpha = .006$) (*Beginning with sample 26, 10 red beads were added before each drawing)

beads are added to the box before each successive sample is drawn. This gradual increase in the percentage of red beads in the box is comparable to the occurrence of an assignable cause in the production process. The control chart clearly shows that the number of red beads in the samples begins to increase until finally an observation is above the control limit—a signal to look for an assignable cause. The last 4 samples dramatically verify the existence of an assignable cause for the process being out of control.

Remark In this problem, only an upper control limit was set. Depending upon its nature or the alternative hypothesis, a control chart may have only a lower limit or both upper and lower limits.

Control charts may be set for averages as well as for the number or percentage of defects. As an illustration, samples from the bowl of 998 chips (Table 1.2) may be plotted on a chart with the average on the ordinate and the order of sample selection on the abscissa. The distribution of values in this bowl is approximately normally distributed with mean equal to 3.0 and standard deviation equal to 1.0. Under these sampling conditions, the sampling distribution of the mean is approximately nor-

mally distributed with a mean equal to the mean of the population and a standard deviation equal to the standard deviation of the population divided by the square root of the sample size. In equation form, these two statements are

$$\mu_{\bar{X}} = \mu_x \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}}$$

This bowl of chips is the model used for the production process in a state of control.

Sample averages vary about the true mean value, and the amount of variability is measured by $\sigma_{\bar{X}}$. In this problem both upper and lower control limits are set because neither too high nor too low an average is desirable. Control limits in $\bar{X}$ charts are customarily set at the 3-sigma level. In other words, the probability of looking for an assignable cause when the process is in control is set at .0026 (comparable to alpha). Other probability levels may be adopted, such as an inner level at the 2-sigma limits and an outer level at 3-sigma limits. The 3-sigma limits have proved successful in much quality-control work.

With the given information about μ_x and σ_x , the upper and lower 3-sigma control limits are

$$UCL = \mu_{\bar{X}} + 3\sigma_{\bar{X}} = 3.0 + 3 \frac{1.0}{\sqrt{10}} = 3.948$$

$$LCL = \mu_{\bar{X}} - 3\sigma_{\bar{X}} = 3.0 - 3 \frac{1.0}{\sqrt{10}} = 2.052$$

When the sample averages fall within these limits, the process is assumed to be in control. A point outside these limits is an indication that the process may be out of control due to an assignable cause (see Fig. 11.16).

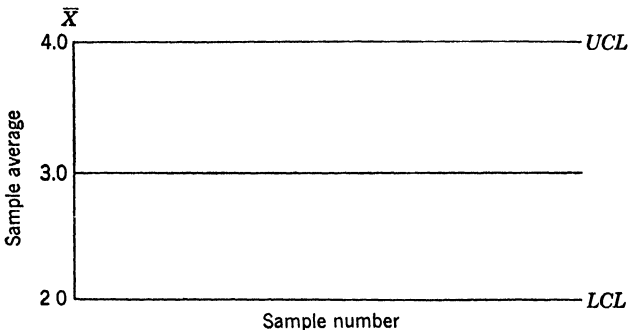


FIG. 11.16. Control chart for $\bar{X}$ with $n = 10$, $\mu_x = 3.0$, $\sigma_x = 1.0$ (3-sigma limits shown).

The most difficult problems in setting up control charts are practical, not theoretical. Bringing the process under control is both a statistical and an engineering problem. Estimates of P or of μ and σ must be

obtained from a large number of observations taken during the experimental stage of control-chart development in order to discover the characteristics of the process when it is in a state of statistical control. Limits are continually revised in the light of recent data on the performance of the process. References cited at the end of the chapter give additional information about control charts.

EXERCISES

11.5. Draw an upper control limit on the graph of the number of red beads plotted in their order of appearance in Exercise 1.4, part (d). Select your own value of α . What determines how large α should be? Did any points indicate lack of control? What is the probability that a point will indicate lack of control when, in fact, the process is in control?

11.6. Draw both upper and lower control limits on the graph of the sample means plotted in their order of appearance in Exercise 1.6, part (d). Did any of the points indicate lack of control? What is the probability that a point will indicate lack of control when, in fact, the process is in control?

11.7. What is the probability that a point will indicate that the process is in control when, in fact, it is out of control in Exercise 11.5? What would you need to know in order to compute this probability? Determine the operating characteristics of this control chart. Are you satisfied with them? How could you change them? What disadvantage is there in setting control limits only on the basis of α ?

11.8. Answer the questions posed in Exercise 11.7 for the control chart in Exercise 11.6.

11.9. Locate in an issue of *Industrial Quality Control* an article which contains a control chart for either percentages or averages. Write a brief summary of the application of the control chart in this article.

11.5 More on the Point of View

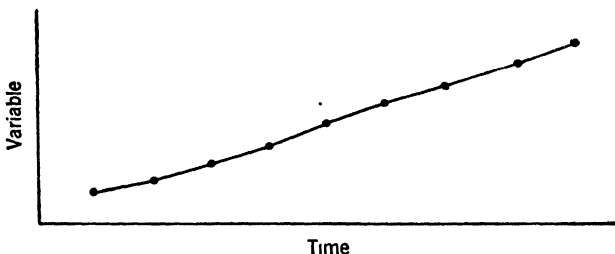
A control chart is essentially a time series because the samples are observed at different points of time and recorded on the chart in their order of appearance. The control chart is introduced here because it is an important statistical tool and because certain ideas connected with it help to elucidate time series analysis.

The unchanging distribution of values in the bowl of 998 numbers serves as a model of a random time series process with no systematic identifiable forces. A value in the series is thought to be composed of many small independent factors and to be a single observation from a distribution of values which has a fixed mean and standard deviation. If the economy is comparable to a production process in a state of control, it produces random series such that values generated in samples are like chance selections from a bowl with fixed mean and standard deviation.

The assignable cause of the control chart is like the systematic identifiable component of a time series. The bowls that are being sampled in time are changeable; i.e., the distributions of values sampled change over time. If the systematic identifiable component is a trend, the change is a slow increase or decrease in the average; if it is a cycle, the averages of the distributions rise and fall in a cyclical pattern. The values of the observations are not completely determined by these systematic components because a distribution of values exists at any instant of time. The myriad of unpredictable and partially offsetting forces—called random forces—are superimposed upon the systematic component to form this universe distribution.

One important difference between control-chart analysis and time series analysis may not be readily apparent. The control-chart analyst—statistician, engineer, machine operator, foreman—tries to keep the machine or production process in a state of statistical control so that it will operate as a random process. Plotting periodic observations from the machine on the chart reveals any departures from a random operation or, in other words, the presence of an assignable cause. The control chart alerts the analyst to detection of assignable causes as soon as they begin to operate. Historical economic data are a matter of record. The economy may be tampered with in numerous ways, but it cannot purposely be put into a state of statistical control so that it will operate as a random process. The objective of time series analysis is the discovery of systematic identifiable forces that may have operated in the past and may continue to operate in the near future.

Why, then, use the random model for the economy? Can the economy and related historical measurements of time series be likened to drawings from a bowl or from an unchanging universe of values? Just as a sample from a population with 50 per cent red beads may by chance contain all red beads, so a finite number of observations from an unchanging bowl when ordered in time may “exhibit,” say, a trend by virtue of random sampling fluctuations and not by virtue of any identifiable systematic component. There is some chance that each successive observation will be higher than the preceding value and that the ordered series will show an upward movement, thus:



The hypothesis that the universe is an unchanging distribution of values with constant mean and standard deviation must be tested and nullified before measurement of a systematic component is attempted. The random model (as defined in terms of the bowl) is a useful starting point because (1) it is simple and assumes no identifiable systematic force; (2) it induces a healthy skepticism as to time series methods; (3) it directs attention to judgments based on other information, which are so essential to time series analysis; and (4) it is easily tested statistically. It may be repeated, therefore, that the random model of a time series with no systematic forces must be nullified before a model which specifies systematic components is accepted and analysis of these components undertaken. This is in keeping with the point of view expressed in Sec. 11.3 that the two alternative models of time series are:

H_0 : The economy is a random process with no systematic identifiable forces.

H_1 : Systematic identifiable forces are present among the random forces in the economy.

11.6 Analysis of a Random Series

Table 11.2 presents 100 observations ordered by the time of drawing. Figure 11.17 gives the general impression that the data have both a

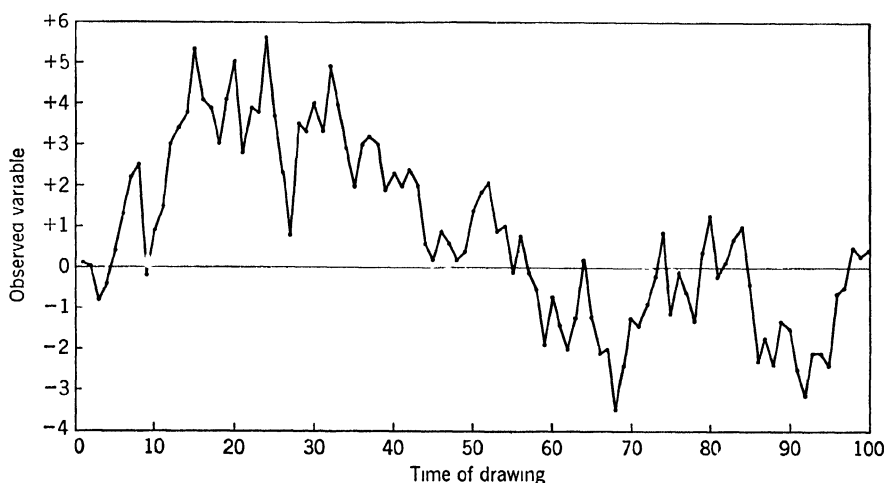


FIG. 11-17. Observations of a random series ordered by the time of drawing (a) Positive trend, (b) negative trend; (c) alternating high and low values; (d) Random series.

slight downtrend and an oscillation. The observations decline from a level of approximately +4 to a level of about -1, and oscillate over their entire range from 0 up to 5, down to -3, and back to 0. Although the

TABLE 11.2. ONE HUNDRED OBSERVATIONS ORDERED IN TIME
(Rounded to Nearest Tenth)

t	First difference	y	t	First difference	y	t	First difference	y	t	First difference	y	t	First difference	y	
1		1	21		2	8	41	2	0	61	-1	4	81	-.2	
2	-		22	+	3	9	42	2	4	62	-2	0	82	.1	
3	-	- 8	23	-	3	8	43	2	0	63	+	1	83	.7	
4	+	- 4	24	+	5	6	44	.6	64		.2	84		1 0	
5	+	4	25	-	3	7	45	.2	65		-1.2	85		- 4	
	+			-											
6		1	3	26	2	3	46	.9	66		-2	1	86	-2 3	
7	+	2	2	27	.8	47		6	67		-2	0	87	-1 7	
8	+	2.5	28	+	3	5	48	.2	68		-3	5	88	-2.4	
9	-	-.2	29	-	3	3	49	.4	69		-2	4	89	-1 3	
10	+	9	30	+	4	0	50	1	4	70	-1	2	90	-1 5	
	+			-											
11		1	5	31	3	3	51	1	9	71	-1	4	91	-2 5	
12	+	3	0	32	4	9	52	2	1	72	-	9	92	-3 1	
13	+	3	4	33	4	0	53	9	73		-	2	93	-2 1	
14	+	3	8	34	2	9	54	1	0	74		9	94	-2 1*	
15	+	5	3	35	2	0	55	-.1	75		-1	1	95	-2 4	
	-			+											
16		4	1	36	3	0	56	.8	76		-	1	96	-.6	
17	-	3	9	37	3	2	57	-	1	77		-	6	97	- 5
18	-	3.0	38	-	3	0	58	-	5	78		-1	3	98	.5
19	+	4	1	39	1	9	59	-1	9	79		.4	99	.3	
20	+	5	0	40	2	3	60	-.7	80		1.3	100		.4	
	-			-											

* This value is -2.08 to two decimal places.

foregoing statements may be adequate as descriptions of the data, they are wholly inadequate as inferences about the mechanism which produced these data. The next sections are devoted to a discussion of such inferences.

First Difference Test for Trend. A simple test for trend in a time series with a large number of observations is based upon the plus and minus signs of the *first differences* of the series. A first difference is the difference between any observation and the preceding observation. The first difference is positive (+) if the succeeding observation is the larger of the two and negative (−) if the succeeding observation is the smaller of the two. If a positive trend is present, successive observations tend to become larger and the series should have a preponderance of + signs. This is shown in Fig. 11.18a. If the trend is negative, successive obser-

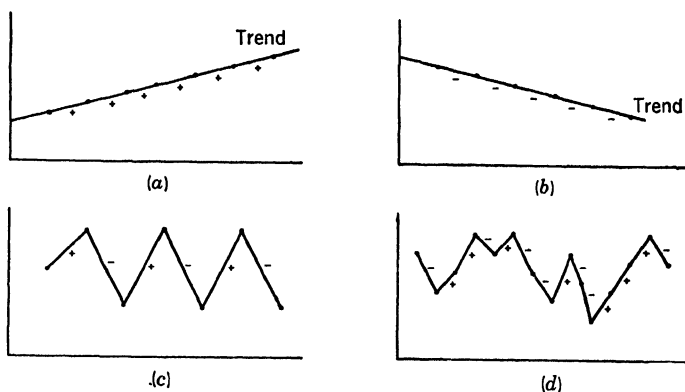


FIG. 11.18. Examples of signs (+ and −) of first differences.

vations tend to become smaller and the series should have a preponderance of − signs, as shown in Fig. 11.18b. If a series is random, approximately half the signs of the first differences should be positive and half negative, in random order. They should not alternate as in Fig. 11.18c, where the series is a succession of alternating plus and minus signs. The pattern should rather be random, as illustrated in Fig. 11.18d.

If the time variable is ordered according to time, $t_1, t_2, \dots, t_n$, and the variable according to $y_1, y_2, \dots, y_n$, there are $n - 1$ first differences:

$$y_2 - y_1, y_3 - y_2, \dots, y_n - y_{n-1}$$

If the observations are a random series, i.e., if they are drawings from a population with constant mean and standard deviation over time, the expected number of positive (or negative) differences is

$$\mu = \frac{n - 1}{2} \quad (11.1)$$

and the standard deviation of the number of positive differences is found to be

$$\sigma = \sqrt{\frac{n+1}{12}} \quad (11.2)$$

When n is large, the distribution of the number of plus (or minus) signs approaches *normality*. A continuity correction is made by subtracting $\frac{1}{2}$ from the difference between the observed and expected frequencies so that

$$Z = \frac{x - \frac{1}{2} - \mu}{\sigma} \quad (11.3)$$

where x is the number of plus signs. This test for trend is applied to the data of Table 11.2.

H_0 : Series is random (sampling is from an unchanging bowl).

H_1 : Series has a trend (sampling is from a set of bowls that is changing on the average).

$\alpha = .05$ $n = 100$ $x = \text{number of } + \text{ signs} = 52$ (Table 11.2)

R : Two-tailed (the trend could be either positive or negative).

With $\alpha = .05$, $\alpha/2 = .025$, and $Z = \pm 1.96$,

$$\mu = \frac{n-1}{2} = \frac{99}{2} = 49.5 \quad \sigma = \sqrt{\frac{n+1}{12}} = \sqrt{\frac{101}{12}} = 2.9$$

From Eq. (11.3),

$$\begin{aligned} +1.96 &= \frac{x_r - .5 - 49.5}{2.9} & -1.96 &= \frac{x_r - .5 - 49.5}{2.9} \\ x_r &= 55.7 & x_r &= 44.3 \end{aligned}$$

Decision Rule. Accept H_1 if $x \geq 55.7$ or $x \leq 44.3$; otherwise accept H_0 .

Decision. Accept H_0 because $x = 52$ falls in the region A . This time series does not give evidence of a trend force. The slight apparent downward trend in Fig. 11.17 is a chance phenomenon.

Remark. In Table 11.2, the observations numbered 93 and 94 both have the same value of y . In the original data, before rounding to one decimal place, the value for observation 94 was -2.08 . This makes the first difference positive.

A Phase Test for a Cycle. A test for the systematic oscillation which makes *no assumption about the distribution of values in the universe* (e.g., whether it is normal) is based upon *sequences of like signs* in the differences between successive observations.¹ A smooth cycle will have long

¹ W. Allen Wallis and Geoffrey Moore, A Significance Test for Time Series Analysis, *Journal of the American Statistical Association*, vol. 36, pp 401-409, September, 1941.

groups of positive signs alternating with long groups of negative signs, as shown in Fig. 11.19. If the series is random, the up-and-down movements leading to groups of positive and negative signs should be of shorter duration, i.e., there should be more alternation between groups of plus and minus signs than if a systematic force is present.

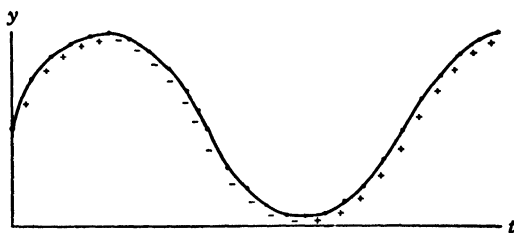


FIG. 11.19. Sequences of like signs in a smooth cycle.

A *turning point* in a series is the point where the series ceases to rise and starts to decline, or ceases to decline and starts to rise. A *phase* is the interval between successive turning points. The length of a phase is the number of intervals between its initial and terminal points. For example:

Ordered series:	1	4	2	1	3	5	6	1	5	1
First difference phases:	+	-	-	+	+	+	-	+	-	
Length of phase:	1	2		3			1	1	1	

The statistical test requires a frequency distribution of the number of phases of lengths 1, 2, and 3 or more. In a series of n independent random observations, the *expected number* (U) of phases of these various lengths is

$$U_1 = \frac{5(n-3)}{12} \quad U_2 = \frac{11(n-4)}{60} \quad U_3 = \frac{4n-21}{60} \quad (11.4)$$

The subscript indicates the length of the phase: 1 for 1; 2 for 2; and 3 for 3 or more. The observed number of phases of these lengths in a series is designated by u_1 , u_2 , and u_3 , respectively. The actual and expected frequencies are compared by computing the following value of the test statistic χ_p^2 :

$$\chi_p^2 = \frac{(u_1 - U_1)^2}{U_1} + \frac{(u_2 - U_2)^2}{U_2} + \frac{(u_3 - U_3)^2}{U_3} \quad (11.5)$$

Selected probability levels of χ_p^2 for n greater than 12 are given in Table A.8.

The original data of Table 11.2 are used to assign plus and minus signs

to the differences between successive observations. Table 11.3 gives the observed and expected number of phases of different lengths. The statistical test follows.

H_0 : No cycle is present in the mechanism which gave rise to the data of Table 11.2 and Fig. 11.17.

H_1 : A cycle is present in the mechanism which gave rise to the data of Table 11.2 and Fig. 11.17.

$$\alpha = .05 \quad \chi_{.05}^2 = 6.898 \text{ (from Table A.8)}$$

TABLE 11.3. PHASE TEST FOR A CYCLE IN THE DATA OF TABLE 11.2

Length of phase	Observed number u	Expected number U	$u - U$	$(u - U)^2$	$(u - U)^2/U$
1	31	40.4	-9.4	88.36	2.18
2	13	17.6	-4.6	21.16	1.20
3 or more	12	6.3	+5.7	32.49	5.16
Total	56	.	.	.	$\chi_p^2 = 8.54$

$$U_1 = \frac{5(n-3)}{12} = \frac{5(97)}{12} = 40.4 \quad U_2 = \frac{11(n-4)}{60} = \frac{11(96)}{60} = 17.6$$

$$U_3 = \frac{4n-21}{60} = 6.3$$

Decision Rule. Accept H_1 if $\chi_p^2 \geq 6.9$; otherwise accept H_0 .

Decision. Accept H_1 because $\chi_p^2 = 8.54$ falls in the region R . The order of the data is consistent with the hypothesis that a cycle is present in the mechanism which gave rise to the data of Table 11.2 and Fig. 11.17.

Remarks on the Construction of the Series in Table 11.2. The statistical tests of the preceding two sections lead to the following inferences about the mechanism which gave rise to the data of Table 11.2 plotted in Fig. 11.17:

1. The data lack evidence of a trend force in the underlying mechanism.
 2. The data give evidence of a cycle force in the underlying mechanism.
- The truth is that the observations were formed by *summing consecutive values of random drawings from a normal distribution* with mean $\mu = 0$ and standard deviation $\sigma = 1$. The drawings were from a population with a fixed mean and standard deviation. No trend force was present in the mechanism which gave rise to these data and, correctly, the statistical test for a trend did not find one. Neither was there a cycle force in the mechanism which gave rise to the data although the statistical test found a cycle force operating. The inference of a cycle force is incorrect, and it is important to examine why it is incorrect.

The decision to accept H_1 , that a cycle force is operating to produce these data, is based upon the arbitrary probability α —the probability of accepting H_1 when, in fact, H_0 is correct. In this problem, α was set at .05 and this sample could be one of those 5 in 100 that will occur on the average just due to chance when H_0 is correct. The fact is that the decision to accept H_1 is not due to an unusual sample. The cycle which is apparent in the series and statistically verified is due to the method of manipulating the data. This cyclical effect is called the Slutsky-Yule effect after two men who have examined it in detail. The sum (or average) of a random variable may be shown to result in a cycle when ordered in time. The meaning and importance of this effect is that *an apparent cycle in a time series may actually be only the result of averaging or summing random data*. This effect, so dramatically illustrated by the data of Table 11.2 and Fig. 11.17, exemplifies one more caution that must be exercised in the treatment of historical time series data, particularly because successive averaging of observations is one of the most popular methods of analyzing data. The discussion of the moving average in Chap. 13 will say more on this point.

Remark. A table of random drawings from a normal distribution with $\mu = 0$ and $\sigma = 1$ is given in Wilfrid J. Dixon and Frank J. Massey, "Introduction to Statistical Analysis," McGraw-Hill Book Company, Inc., Table 2, pp 295-304, New York, 1951. The series of Table 11.2 was formed by starting at random on page 296, column 15, line 2, at the number 0.050 and ending on page 296, column 17, line 2. The beginning of the series is given below.

Observation number t	Random value y	Rounded value y	Sum of the random value (Table 11.2)
1	0 050	0 1	0 1
2	-0 058	-0 1	0 0
3	-0.807	-0 8	-0 8
4	0.420	0 4	-0 4
5	0 791	0 8	0 4
.	.	.	.
.	.	.	.
.	.	.	.

EXERCISES

11.10. The decision rule in the phase test for a cyclical pattern in a set of data was based only upon a selection of α . For what reason that you can suggest is the operating-characteristic curve left undetermined? What is the implication of this practice for β ? For what reason may β be large? What caution does this suggest for practical statistical investigations of cyclical forces with this method?

11.11. With the mean and standard deviation of the lifetimes of incandescent lamps computed in Exercise 8.4, set up a control chart for the mean length of life of samples of 10 bulbs. Plot the averages of the samples reported by Davis on this control chart. (Note that 3 samples are exceptions with only 9 observations.) Do any points fall out of control? What is the probability that a point will fall out of control when, in fact, the process is in control?

SUPPLEMENTARY READING

- Burr, I. W.: "Engineering Statistics and Quality Control," Chaps. 5-7, 9-11, McGraw-Hill Book Company, Inc., New York, 1953.
- Grant, E. L.: "Statistical Quality Control," pts. II and III, McGraw-Hill Book Company, Inc., New York, 1952.
- Moroney, M. J.: "Facts from Figures," Chaps. 11 and 17, Penguin Books, Inc., Baltimore, 1953.

CHAPTER 12

THE ANALYSIS OF TIME SERIES SEASONAL VARIATION

Figure 12.1 is a graph of total egg production (in billions of eggs) in the United States for each month during the five-year period 1948 to 1952, plotted from the data of Table 12.1. Using these historical data,

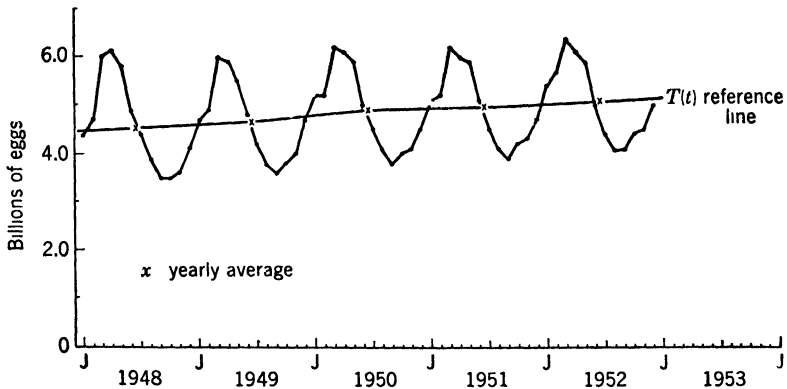


FIG. 12.1. Monthly egg production in the United States, 1948 to 1952. (SOURCE: *U.S. Department of Agriculture.*)

what would be your estimate of total egg production in the United States for January, 1953? This is a matter of vital concern to wholesalers, producers, and retailers of chickens and eggs, since production affects the prices of both. Lacking the specialized knowledge of those directly engaged in the egg business, can you predict future egg production?

12.1 A Test for Seasonal Variation in Egg Production

The immediately obvious and outstanding feature of Fig. 12.1 is the regular fluctuation of egg production within each year. Production reaches a peak in March and April, falls off to a low in September, and then recovers to repeat the oscillation in the next year. Even without any knowledge of chickens, you may realize that the variation depends

to some extent on the weather or the season of the year. Chickens lay fewer eggs in very hot or very cold weather. If the total number of chickens remains the same, the production of eggs will vary with the seasons.

Seasonal variations affect not only egg production, but also such economic series as crop production, industrial employment, residential construction in some cities, department-store sales, and money in circulation. Seasonal variations that are artificially induced by customs or holidays (e.g., the Christmas and Easter effect on department-store sales) are no less real than phenomena directly related to the weather.

TABLE 12.1. MONTHLY EGG PRODUCTION: 1948 TO 1952
(In billions)

Month	1948	1949	1950	1951	1952
January	4 4	4 7	5 2*	5 1	5 4
February	4 7	4 9	5 2	5 2	5 7
March.	6 0	6 0	6 2	6 2	6 4
April . . .	6 1	5 9	6 1	6 0	6 1
May	5 8	5 7	5 9	5 9	5 9
June	4 9	4 8	5 0	5 1	5 0
July . . .	4 4	4 2	4 5	4 5	4 4
August .	3 9	3 8	4 1	4 1	4 1*
September	3 5†	3 6	3 8	3 9	4 1
October	3 5	3 8	4 0	4 2	4 4
November	3 6	4.0	4.1	4 3	4 5
December	4 1	4.7	4 5	4 7	5 0
Total	54 9	56 1	58 6	59 2	61 0
Average	4 57	4 67	4 88	4 93	5 08

SOURCE: U.S. Department of Agriculture, Bureau of Agricultural Economics, *Monthly Egg Production, by Months, 1943-1952*, bulletin released Feb. 27, 1953.

* Higher than the succeeding observation.

† Lower than the succeeding observation.

A feature of the egg-production series less outstanding than seasonal variation is the gradual increase in the average level of production. The average monthly egg production increased for each successive year within the given period (shown in Table 12.1, and on Fig. 12.1 by small crosses), but the effect is so small as to be negligible to the analysis. This leads to the supposition that egg production for this period may consist of two parts:

1. A regular seasonal variation occurring within each year
2. Irregular or random variations

In other words, the egg-production process is presumably affected by one systematic identifiable force—seasonal—and by the random variation.

A mathematical model which specifies these two forces in combination, or with one superimposed upon the other, is

$$y = S(t) \cdot R(t) \quad (12.1)$$

where y is egg production, $S(t)$ is the seasonal component which depends upon time, and $R(t)$ is the random component. The model specifies that these two components are multiplied, as shown in Eq. (12.1). If the seasonal component can be isolated from the random component, it may be used to predict future monthly egg production.

Any attempt to measure the seasonal force is based upon the fundamental premise that an order exists in the process of economic change (egg production in this example) and can be abstracted. The study of statistical regularities in historical time series plus a *criterion* of economic change yields this abstraction. The emphasis upon a criterion is most important. Many economic forecasters will readily admit from hard-won experience that past statistical regularities are not enough. In the egg-production problem the statistical regularity is evident; the criterion of this regularity is the seasonal variation. The underlying economic force in this series is the weather, for, other things being equal, total egg production varies with the season. This criterion is an inherent part of the production cycle of the chicken. With the evidenced statistical regularity of the series and a good criterion, the seasonal variation can validly be measured.

The crucial question is this: Does the mathematical model of Eq. (12.1), used to describe egg production, abstract any of the essential forces of economic change, especially the seasonal variation? The very marked statistical regularity and rational criterion of seasonality are the basis for the assumption that it does. Since all historical time series are not so orderly as the egg-production series, the question is restated: Does the mathematical model $y = S(t) \cdot R(t)$ separate out any of the essential forces which make for economic change, especially $S(t)$?

How is this model to be judged? It is judged relative to another model which specifies that no systematic forces are operating in the mechanism which gave rise to the data. This is in keeping with the point of view expressed in Secs. 11.3 and 11.5 that the two alternative hypotheses for time series analysis are:

H_0 : The economy is a random process with no systematic identifiable forces.

H_1 : Systematic economic forces are present among the random forces in the economy.

Stated in terms of the egg-production process, these hypotheses become:

H_0 : Egg production is a random process.

H_1 : Egg production is affected by a seasonal variation along with a random variation.

In Sec. 11.5, the random process was likened to random drawings from a series of bowls with fixed means and standard deviations, and the process with a systematic component was likened to drawings from a series of bowls with means changing in a regular manner—a gradual increase or decrease for a trend force and a rising and falling pattern for an oscillation. In the egg-production problem, what is the universe that may be likened to a bowl whose contents form a distribution of values with a fixed mean and standard deviation? The universe is the distribution of total egg-production values for a given month. Why is there a distribution of values? Total egg production is so many eggs in a given month. Why can one say that a monthly observation is just one number selected by chance from a distribution of possible egg-production values for that month?

The total egg production in a particular month is the sum of the individual egg productions of about five million farmers in the United States. Approximately 20 per cent of them are large producers who raise only eggs. The rest are small farmers whose flocks of chickens represent only one aspect of their occupation. The number of eggs laid by the individual flocks of these millions of farmers in any one given month is affected by many chance factors, such as the number of chickens, price expectations, temperature, rainfall, amount of daylight, amount of feed, feeding methods, etc. These factors, some of which are beyond the control of the farmer, cause egg production to be higher or lower than it would have been under different circumstances. For example, hens lay eggs only during the daylight hours. Large producers light their chicken coops artificially, but chance clouds and storms reduce egg production for other farmers. Egg production may also vary because dogs and foxes catch hens, hens hide eggs for nesting, eggs break in the dropping or collecting process, or farmers eat eggs which are not counted. Thus, chance whims of the farmer, the fox, or the hen herself make for variation. In summary, actual egg production in any given month depends upon a combination of many seemingly unimportant and chance events on each farm, large or small. Since many different combinations of such events are possible, total monthly egg production could take on many different values, but only one of these is reported in a particular month.

In other words, the egg-production process may be likened to a bowl with a distribution of numbers having a fixed mean and standard deviation. Each month, one number from the bowl represents actual egg production. The hypothesis H_0 , that egg production is a random process

with no systematic identifiable forces, means that the same bowl or distribution of values is sampled each month of the year and that the ups and downs of observed egg production are due only to chance selections of a combination of contributing factors. If the egg-production process is really like this, it is impossible to predict whether production in a particular month will be high or low. At best, the mean level of production can be predicted and the variation around this mean measured.

The alternative hypothesis H_1 , that systematic economic forces are present among the random forces, means that the egg-production process has a regularly recurring seasonal variation. The bowl or distribution of values changes from month to month. The means do not remain the same but vary in a regularly recurring pattern because of seasonal forces. Chance factors still operate to determine that value selected from the distribution each month, but they operate around a different mean value each month. If the egg-production process is like this, and if the seasonal pattern is known, it is possible to predict whether production in a particular month will be high or low. Because of the presence of random forces, the prediction will not be exact, but it will be more accurate than the prediction of the mean egg production in a random process.

Remark. The text has gone to great lengths in describing the factors that affect egg production. This description is not necessary to the technical statistical analysis to follow. A statistician can make the necessary manipulations and measurements without this description, but he should know a good deal about the data and the mechanism which gave rise to them in order to make his analysis meaningful both to himself and to others. Statistical methods in general and time series methods in particular are not machines into which data may be poured and answers ground out. Successful application depends upon a great deal of knowledge about the data and the factors behind them.

A test for the existence of a seasonal force in egg production is the phase test for an oscillation used in Sec. 11.6 for a cycle. Details of the test are given in that section. The test depends upon sequences of like signs in the first differences. The test for the seasonal variation in egg production takes the following form:

H_0 : Egg production is a nonseasonal process.

H_1 : Egg production contains a seasonal movement.

$\alpha = .01 \quad n = 100 \quad \chi_{.01}^2 = 10.25$ (From Table A.8)

Decision Rule. Accept H_1 if $\chi_p^2 = 10.25$; otherwise accept H_0 .

The original data of Table 12.1 are used to assign plus and minus signs to the first differences. (Note that three tied values occur in this table, but that in each case a footnote identifies the larger of the pair. In the original data, values were reported to more than one decimal place.) Table 12.2 gives a summary of the actual and observed phases.

Decision. Accept H_1 because $\chi_p^2 = 44.98$ falls in the region R , and conclude that the criterion of seasonal variation in egg production is a systematic identifiable force of economic change and that it may therefore be measured.

TABLE 12.2. PHASE TEST FOR THE SEASONAL COMPONENT IN EGG PRODUCTION

Length of phase	Observed number u	Expected number U	$u - U$	$(u - U)^2$	$(u - U)^2/U$
1	2	23 75	-21 75	473 06	19.92
2	0	10 26	-10 26	105 27	10.26
3 or more	11	3.65	+ 7 35	54 02	14 80
Total	13	.			$\chi_p^2 = 44 98$

$$U_1 = \frac{5(n-3)}{12} = \frac{5(57)}{12} = 23.75 \quad U_2 = \frac{11(n-4)}{60} = \frac{11(56)}{60} = 10.26$$

$$U_3 = \frac{4n-21}{60} = \frac{240-21}{60} = 3.65$$

EXERCISE

12.1. An index of department-store sales is published monthly in the *Federal Reserve Bulletin*. Data for 1950 to date are given in the table below.

UNADJUSTED DEPARTMENT-STORE SALES, TOTAL UNITED STATES
(1947-1949 = 100)

Month	1950	1951	1952	1953	1954	1955
January. . . .	74	96	83	85	83	
February	77	90	83	88	86	
March	89	98	92	103	89	
April	99	99	103	115	110	
May	99	103	108	108	106	
June .	97	99	105	108	106	
July .	98	84	84	89	88	
August	97	93	98	98	98	
September	114	112	112	112	112	
October	107	112	120	115	118	
November .	122	134	134	136	137	
December .	185	184	196	192	200	

SOURCE: *Federal Reserve Bulletin*.

a. What reasons can you give to support the position that department-store sales could take on many different values in any one month, i.e., that a distribution of sales exists for each month, and that the figure reported is a chance selection from this distribution?

b. Is any systematic component apparent in this series? What is the criterion of this economic change? Specify a model for department-store sales incorporating whatever components are thought to exist.

c. What is the meaning of H_0 : that department-store sales are a nonseasonal series? If H_0 is accepted, what inference may be drawn about these sales?

d. What is the meaning of H_1 : that department-store sales contain a seasonal movement? If H_1 is accepted, what inference may be drawn about these sales?

e. Bring these data up to date by referring to recent issues of the *Federal Reserve Bulletin*. Test for a seasonal force in department-store sales in the period 1950 to date.

12.2 Measuring Seasonal Variation

Several computational methods have been devised to measure seasonal variation. Most of them involve long and complicated arithmetic, and some have additional refinements for adjusting the raw data and averaging the random component. The method described here is a graphic device to measure seasonal variation.¹ Using the data of Table 12.1, the text will explain the procedure and summarize the results of the actual measurement of seasonal variation in the egg-production process.

The model of the egg-production process originally expressed by Eq. (12.1) as $y = S(t) \cdot R(t)$ is changed slightly for the graphic measuring process. Although the slight uptrend in the original data is negligible and is not considered a part of the model in Eq. (12.1), it will be used as a *reference line* from which to measure the seasonal variation. The model is, therefore, changed to

$$y = T(t) \cdot S(t) \cdot R(t) \quad (12.2)$$

where $T(t)$ is the symbol for the *reference line* in the egg production. In the general analysis, each of the original values of y is divided by $T(t)$ in order to isolate the seasonal and irregular components. In equation form, the method is expressed as follows:

$$\frac{y}{T(t)} = \frac{T(t) \cdot S(t) \cdot R(t)}{T(t)} = S(t) \cdot R(t) \quad (12.3)$$

The random fluctuations are then averaged out to leave a measure of the typical seasonal component.

1. Plot the data (of Table 12.1) on semilogarithmic graph paper. Label the abscissa (horizontal scale) with time in months and the ordinate (vertical scale) with egg production. Connect the points with fine straight lines (see Fig. 12.2).

¹ W. A. Spurr, A Graphic Method of Measuring Seasonal Variation, *Journal of the American Statistical Association*, vol. 32, pp. 281-289, June, 1937, W. A. Spurr, A Graphic Short Cut to the Moving Average Method of Measuring Seasonality, *Journal of the American Statistical Association*, vol. 35, pp. 667-670, December, 1940.

2. Measure the reference line freehand, by computing the yearly averages (see Table 12.1) and plotting them at the middle of the year, or halfway between the June and July figures. Plot them as small crosses to distinguish them from the original data. Draw a smooth curve through these five yearly averages to represent the reference line and label it $T(t)$. In the egg-production series, this smooth curve is practically a series of straight lines connecting the averages. (This line is drawn on Fig. 12.2.)

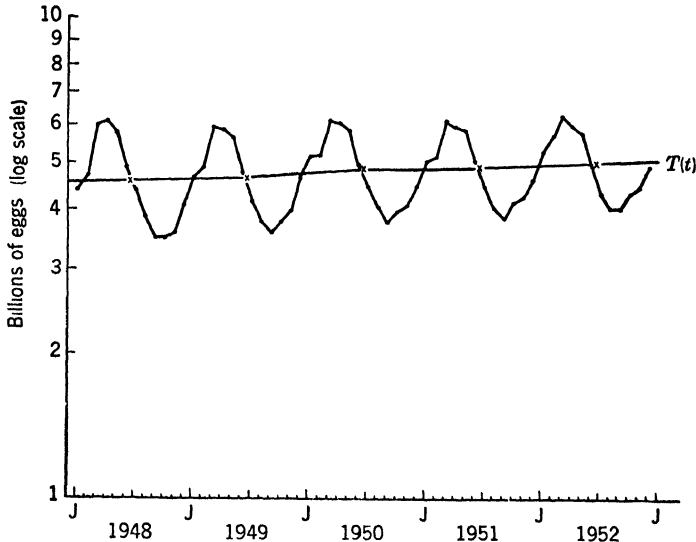


FIG. 12.2. Egg production plotted on semilogarithmic paper.

3. To isolate the seasonal and random components, each monthly value of egg production is divided by the reference value for that month [Eq. (12.3)]. This division is done on the graph paper. Since the values of y and $T(t)$ for each month are points plotted on semilogarithmic paper, the *vertical difference* between them is equivalent to their ratio (see Sec. 11.2). One method of measuring these vertical deviations follows.

a. Take another sheet of semilogarithmic paper [the same kind as used in part (1)], and draw a horizontal line across it near the middle of the paper. Label this line $T(t)$; it is the *reference line* for measuring vertical deviations.

b. Fold this sheet of paper 11 times to make 12 equal vertical sections or strips.

c. Label the sections or strips with the names of the months, January, February, . . . , December. (See Fig. 12.3 for steps *a* through *c*.)

d. Use the first strip to measure a seasonal component for the month of January. Take this strip and place the $T(t)$ line *opposite the reference*

value for January, 1948. Make a mark on the strip opposite the value of y (the January egg production), and label this mark with the figure 8 to identify it as the 1948 value. The January strip will now look like

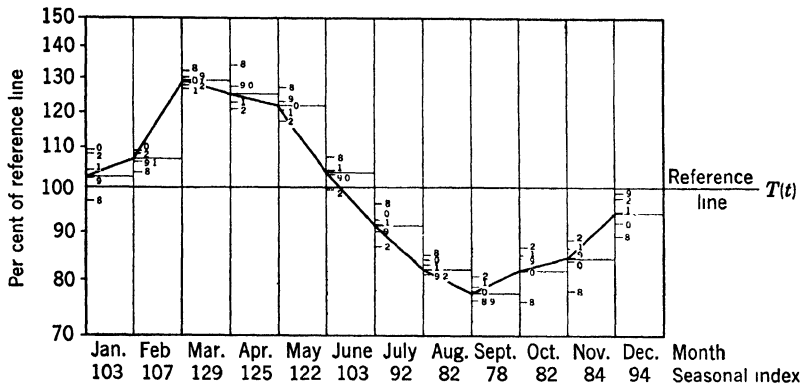
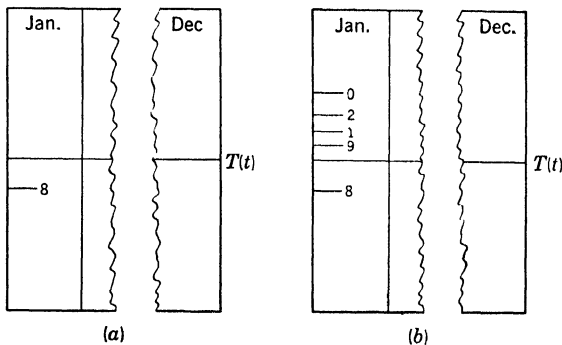


FIG. 12.3. Measurement of seasonal variation on semilogarithmic paper.

(a) below. The deviation of the January, 1948, egg production value from the reference line is thus transferred to the other sheet of graph paper.

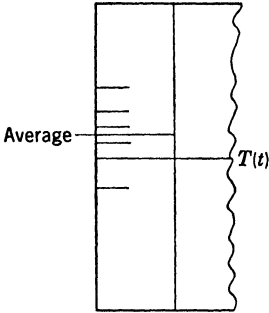
Repeat this process for *each January* of the 5-year period and the strip will look something like (b) below.



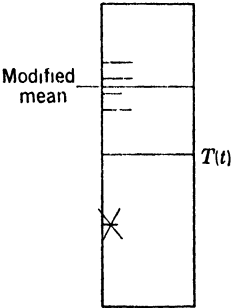
e. Now fold the paper so that the left edge is the strip for February, and repeat the measuring procedure for each of the five Februarys. Continue for each of the twelve months. (See Fig. 12.3 for a graph of the deviations transferred for each month.)

f. Each deviation represents a division of one original observation by its reference value, in other words, a measure of the seasonal and random components: $S(t) \cdot R(t)$. The seasonal-random components for any one month are not the same for all years. The January deviations vary among themselves; the February deviations vary among themselves, and so on. A *typical seasonal index* for a given month is computed by *aver-*

aging the five observations for that month. This average is made on the graph paper by drawing a line by eye in the middle of the five values. For example, the monthly strip for January will look something like the figure below. This average is the *unadjusted seasonal index* for January. The process is repeated for the other months to arrive at their unadjusted indexes. (See Fig. 12.3 for the 12 index averages.)



Remark. In measuring seasonal variation by this averaging method, one month is frequently far out of line with the rest. For example, suppose a strip were to look like this:



Four of the deviations are quite close together, but the fifth is very much lower. In practice, knowledge of the time series in question plus investigation would probably reveal a pertinent reason for such irregularity. A phenomenon *other than a seasonal* influence may have caused the one value to be low. It is customary to omit an irregular extreme from the average and to find a *modified mean*, or a mean of the remaining deviations, for the unadjusted index. Although the egg-production series does not markedly exhibit this behavior, you should be aware of the problem. An extreme observation should be omitted only after careful study to see if it is due to some irregular cause. This is one more instance where the analyst must know a good deal about his data for an intelligent analysis.

4. Scale the second sheet of graph paper in percentages because the seasonal index is measured as a percentage of the reference value (see Sec. 11.2). Let $T(t)$ equal 100 per cent. Label as 200 per cent the point on the ordinate corresponding to a distance on the original egg-production scale in which the series doubles. Label as 50 per cent the point

corresponding to that distance on the original scale in which the series is halved.

5. Read the value of the typical unadjusted seasonal index from this scale and record it as the unadjusted index in a table.

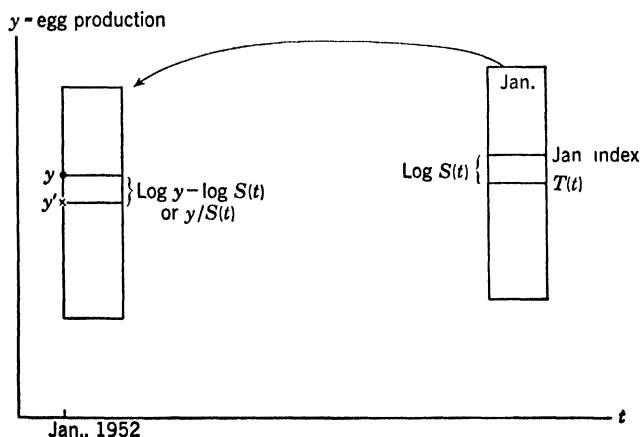
6. Total the 12 unadjusted indexes. If this total is between 1180 and 1220, the unadjusted indexes may be labeled *adjusted indexes*. If the total is outside these limits, adjust the indexes by *multiplying each* of them by the ratio: $1200/(\text{total of the 12 indexes})$. The resulting figures have an average value of 100 per cent and are the *adjusted seasonal indexes*. They should be entered in a table and so labeled. They are the final indexes of seasonal variation.

7. Plot these adjusted indexes on the monthly strips (preferably in a color different from the unadjusted indexes), and connect them to depict graphically the seasonal variation (see Fig. 12.3).

8. Adjust the egg-production series for seasonal variation by dividing the original data by the seasonal index. In equation form, this operation is as follows:

$$y' = \frac{y}{S(t)} = \frac{T(t) \cdot S(t) \cdot R(t)}{S(t)} = T(t) \cdot R(t) \quad (12.4)$$

This step may be done graphically, since the logarithmic division of y by $S(t)$ involves taking the difference between $\log y$ and $\log S(t)$. The $\log y$ is plotted on the original series; the $\log S(t)$ is plotted on the second sheet of paper used to measure the seasonal indexes. The division may be accomplished by subtracting $\log S(t)$ from $\log y$. Place the seasonal *index* for a particular month opposite the appropriate monthly observation (January index for January observation, etc.) of the original series y . Mark a point y' on the original graph opposite $T(t)$ on the seasonal chart. The diagram below illustrates the adjustment for seasonal variation.



9. Connect these adjusted points y' in a different color to show the original egg-production figures adjusted for seasonal variation (see Fig. 12.4).

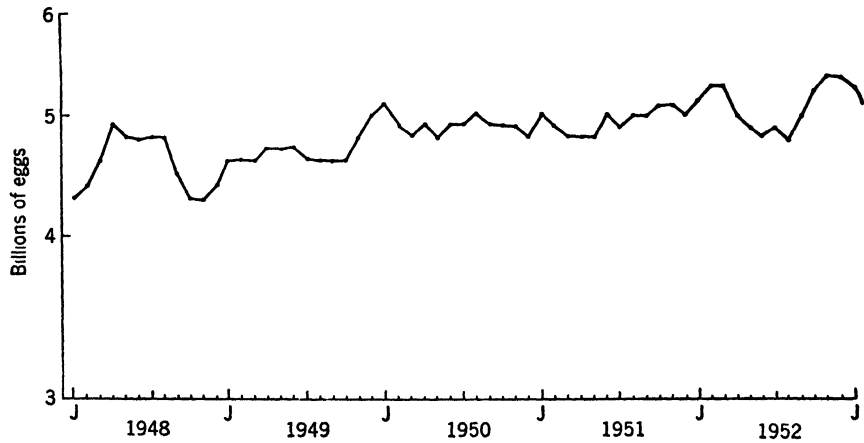


FIG. 12.4. Egg production in the United States adjusted for seasonal variation (plotted on semilogarithmic graph paper).

Table 12.3 shows the seasonal indexes for egg production based upon 1948–1952 data.

TABLE 12.3. SEASONAL INDEXES OF EGG PRODUCTION

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug	Sept.	Oct.	Nov	Dec.
Index (ad- justed)	103	107	129	125	122	103	92	82	78	82	84	94

Remark. It must be understood that $T(t)$ is introduced into the model of egg production merely as a reference point from which to measure the seasonal component $S(t)$ and not because any real trend force is present in the egg-production series for this five year period. In other problems, $T(t)$ may be a bona fide trend force. If so, the appropriate model would be $y = T(t) \cdot S(t) \cdot R(t)$, with the $T(t)$ used as a reference from which to measure $S(t)$, a real trend in the series. In some problems (for example, the department-store sales of Fig. 11.8), a trend and cycle are also present along with the seasonal and random forces. The appropriate model involves all four components and is written $y = T(t) \cdot C(t) \cdot S(t) \cdot R(t)$. The line drawn through the averages then represents the combined trend and cycle, $T(t) \cdot C(t)$; the seasonal index is measured as a percentage of this combined effect.

EXERCISE

12.2. Use the graphic method to measure monthly seasonal indexes of department-store sales given in Exercise 12.1. Adjust the original data with these

seasonal indexes. Compare some of your adjusted figures with those given in the *Federal Reserve Bulletin*.

12.3 Meaning of Seasonally Adjusted Figures

It is tempting to say that a figure adjusted for seasonal variation is the figure that would have obtained had there been no seasonal influence. This interpretation cannot be made. The adjusted egg production for April, 1952, is about 4.9 billion eggs; the unadjusted figure is 6.1. No one could estimate that, in the absence of seasonal forces, egg production in April, 1952, would be 4.9 billion eggs, for no one knows what would have happened without seasonal variation. The seasonal influence upon the chickens which lay the eggs is a very real part of their existence. Who can say that the chicken would be the same bird if there were no seasonal influence? The fallacy of the statement is obvious in the egg-production problem, but not at all obvious in other situations.

The adjusted figure represents an average rate of production. If the actual egg production were equal to the adjusted figure for April, 1952, and if the seasonal pattern of egg production were to begin at that level of production and finish 12 months later, an estimate of the total egg production for the year would be 12 times the adjusted figure. The estimated rate of egg production in 1952 would be 58.8, or 12 times 4.9, based upon the April, 1952, seasonally adjusted production and the experience of the 1948 to 1952 period, compared to actual production of 61.0 billion eggs. Little use is made of seasonally adjusted figures to estimate total production, but the fallacious interpretation of a seasonally adjusted figure is often made.

12.4 Uses of a Seasonal Index to Predict

Seasonal indexes may serve as predictors of future values in a time series. The method is only approximate since the indexes are averages. A prediction of future production may be made by multiplying the actual production in one month by the ratio of the seasonal index of the predicted month to the seasonal index of the predicting month. In November, 1952, for example, actual egg production was 4.5 billion eggs. The seasonal index is 84 for November and 103 for January. A prediction of egg production for January, 1953, based upon the November, 1952, actual egg production, is

$$4.5 \frac{103}{84} = 5.5 \text{ billion eggs}$$

This method of prediction assumes that the general factors affecting the production of eggs in November are still operating to the same extent in January; that the only change in production is due to a seasonal change

measured by the seasonal indexes; and that the magnitude of the seasonal influence has not changed since the time period from which the monthly indexes were computed (1948 to 1952). If any one of these assumptions is untrue, or if all of them depart in some degree from the truth, there will be errors in forecasting. Conditions may not be the same in January, 1953, as they were when the indexes were measured. It is obvious, of course, that these difficulties are always present to plague the economic forecaster.

Errors in Prediction Due to Nonseasonal Factors. Predictions of egg production in the United States for the first 5 months of 1953, made from the actual production of 2 months earlier, are compared to actual production in Table 12.4. The prediction of January production from

TABLE 12.4. PREDICTIONS OF EGG PRODUCTION AND ACTUAL PRODUCTION IN THE UNITED STATES FOR THE FIRST FIVE MONTHS OF 1953

Predicted month	Seasonal index of predicted month	Predicting month	Seasonal index of predicting month	Actual production in predicting month	Prediction	Actual production
Jan	103	Nov.	84	4 5	$4\ 5 \left(\frac{103}{84} \right) = 5\ 5$	5 4
Feb	107	Dec.	94	5 0	$5\ 0 \left(\frac{107}{94} \right) = 5\ 7$	5 3
Mar	129	Jan.	103	5 4	$5\ 4 \left(\frac{129}{103} \right) = 6\ 8$	6 3
Apr.	125	Feb.	107	5 3	$5\ 3 \left(\frac{125}{107} \right) = 6\ 2$	6 1
May.	122	Mar.	129	6 3	$6\ 3 \left(\frac{122}{129} \right) = 5\ 9$	5 9

November's production was good: 5.5 versus actual production of 5.4. The predictions for both February and March overstated actual production. The percentage errors were .4/5.3, or 7.5 per cent, and .5/6.3, or 7.9 per cent, for February and March, respectively. These estimates are fairly good, in view of the fact that they are based upon a mechanical forecast from a historical time series and are not modified by knowledge of the egg industry. More significant, however, is an analysis of the reasons for the overstatement. Was the error made because the magnitude of the seasonal influence changed or because the general factors operating in the predicting month changed in the predicted months? Examination of the number of chickens and of the number of eggs per 100 layers reveals an adequate reason for the decline in production.

There were about 3 per cent fewer chickens on the farms in the predicted months than in the predicting months. The production figure for March would have been even lower and the prediction more in error had not the smaller number of chickens been offset in part by a record rate of laying. The number of eggs laid per 100 chickens reached a record high for the month of March, 1953. Basic nonseasonal conditions changed between the predicting months and the predicted months.

Actual Change versus Expected Seasonal Change. The expected seasonal change from January to March is the ratio of the seasonal indexes for those months: $129/103 = 125.2$ per cent. The percentage change in the seasonal indexes is $(129 - 103)/103 = 25.2$ per cent. The actual percentage change in production was $(6.3 - 5.4)/5.4 = 17$ per cent. The actual increase in egg production from January to March was less than the expected seasonal increase, for reasons given in the preceding section.

Errors in Prediction Due to a Suspected Change in the Seasonal Component. In addition to the change in the basic nonseasonal conditions from late 1952 to early 1953 (a change in the size of flocks), there is evidence that the seasonal component itself is changing. Since valid predictions depend upon a constant typical seasonal variation, any systematic shift in the seasonal will invalidate the predictions.

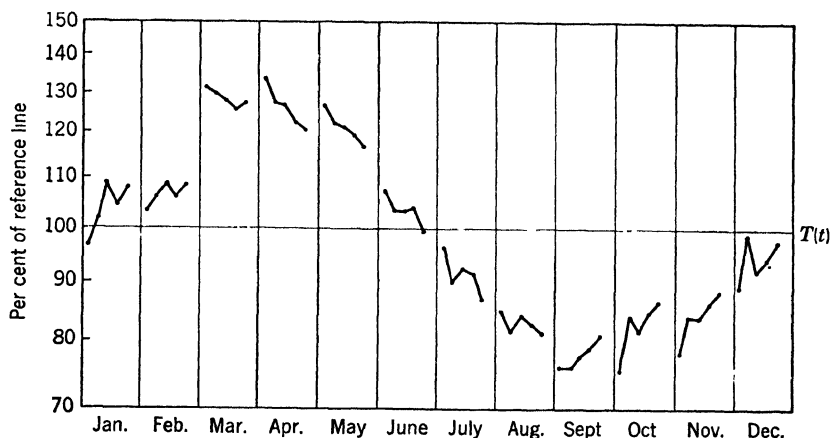


Fig. 12.5. Deviations from reference line plotted in order of occurrence, 1948 to 1952, to show shifting seasonal pattern of egg production.

To detect a shift in the seasonal, the deviations from the reference line are plotted on the monthly strips in order of time, i.e., according to the year in which they were recorded. In effect, a time series of the monthly seasonal deviations from the reference line is made for each month. These deviations are indicated in Fig. 12.5, which dramatically shows a tend-

ency for the percentage deviation from the reference line to decrease from March through August in the period 1948 to 1952. From September through February, the percentage deviation from the reference line tends to increase in the same period of years. Averaging these deviations for any one month is not valid because the variations are not due entirely to the irregular component. The seasonal component itself is shifting. In other words, within the seasonal variations there is an additional systematic component which should be isolated and measured. *Shifting seasonal indexes* are needed to take into account this systematic change in the magnitude of the seasonal variation with time.

The computation of shifting seasonal indexes is not presented here; but the phenomenon is mentioned because of its possible effect on the accuracy of predictions using simple typical seasonal indexes. This phenomenon should be verified for both earlier and later years before acceptance.

EXERCISES

12.3. Use the seasonal indexes computed in Exercise 12.2 to predict department-store sales, and compare your prediction with the actual sales when the data are released by the Federal Reserve Board. Investigate the reason for any serious error in prediction. Is it due to a nonseasonal factor?

12.4. Why would you expect a change in the seasonal pattern of industrial production during a period like 1942 to 1946?

12.5. The sale of life insurance in the United States exhibits a regularly recurring fluctuation within each year. What criterion can you advance to support a hypothesis of a seasonal variation?

The seasonal index of life insurance sales is approximately 90 in September and 116 in December. Life insurance sales were 2589 in September and 3460 in December, 1952. They were 2757 in September and 3735 in December, 1953. What would life insurance sales have been in each of these months had there been no seasonal variation? Were the sales in December above or below the expected seasonal change? (Note that all sales are in millions of dollars.)

12.6. Suppose you are the manager of a food-processing plant that cans tomato, bean, and pea products like soup, strained vegetables, etc. Employment in your plant is very high when these crops are being harvested and very low during the rest of the year. This phenomenon not only creates an employment problem for you but for the whole community where your plant is located. What policies would you consider to alleviate the seriousness of the seasonal employment?

12.7. Egg-production figures for 1953 follow:

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
5.4	5.3	6.3	6.1	5.9	5.1	4.6	4.4	4.2	4.6	4.8	5.3

a. Plot these data on an extension of Fig. 12.2. Project the freehand reference

line through 1953. Does the gradual increase in egg production continue through 1953?

b. Measure the deviations from the reference line, and plot on charts like Fig. 12.3 and 12.5. Is there additional evidence of a shifting seasonal variation?

12.8. As data for subsequent years become available, plot them on semi-logarithmic paper and examine the average egg production and the shifting seasonal. Make predictions of future egg production from time to time and compare the prediction with the actual egg production, as reported monthly in *Crop Production*, Bureau of Agricultural Economics, U.S. Department of Agriculture.

CHAPTER 13

THE ANALYSIS OF TIME SERIES : TRENDS AND CYCLES

Of the three traditional systematic components of a time series—trend, cycle, and seasonal variation—the seasonal force was examined in detail in Chap. 12. This chapter proceeds to a consideration of the trend and cycle. The methods will be analytical where possible and descriptive where necessary.

13.1 Linear Trends

The trend of a series has been represented as a steady persistent movement in the same direction. The examples of time series in Chap. 11 showed trends that were both upward and downward over periods of time. Two general methods of isolating the trend are possible: (1) a procedure analogous to regression methods with a mathematical equation representing the trend; (2) a more general procedure which does not specify a mathematical curve for trend but measures a moving arc of best fit to the data. Both methods are illustrated in the text.

Linear Trend Line. The method of least squares in fitting a regression line to a scatter diagram of two variables may also be used to fit a linear trend line. Data on gross national product¹ (hereafter abbreviated GNP) are shown in Table 13.1 and Fig. 13.1. In this period the GNP is characterized by an apparent gradual growth or trend and an oscillation about this trend. The first step is to measure the persistent uptrend in the data from 1919 to 1941; the second step is to see if this uptrend in the data is a valid trend force in the economy which gave rise to the GNP figures.

¹Gross national product is defined as the total value *at market prices* of all goods produced by the nation's economy, including governmental agencies. Gross national product is equal to national income less subsidies adjusted for current surplus of governmental enterprises plus capital consumption allowances (i.e., depreciation charges, accidental damage to fixed capital, capital outlays charged to current expense), indirect business tax and nontax liability, business transfer payments and statistical discrepancy. (U.S. Department of Commerce.)

The assumption is that any existing trend is linear, that it is of the form

$$T(t) = a + bt \quad (13.1)$$

where $T(t)$ is the trend and t is the time. Estimates of a and b are determined from the sample data according to the method of least squares. The arithmetic is greatly simplified if the time variable is arbitrarily coded by taking a middle year as a reference point and letting $t = 0$ for



FIG. 13.1. GNP, 1919 to 1941. (SOURCE: U S. Department of Commerce.)

that year. For example, if 1930 is the reference year, it is assigned the value $t = 0$. Then every other year is expressed by an integer measured from 1930 as a reference point; e.g., 1931 is $t = 1$; 1925 is $t = -5$. The data and computations are shown in Table 13.2. The formulas for a and b are variations on the linear regression-line formulas with t substituted for x as the independent variable (see Sec. 10.4).

TABLE 13.1. GROSS NATIONAL PRODUCT: 1919 TO 1941
(Billions of Current Dollars)

Year	GNP	Year	GNP	Year	GNP
1919	60.4	1927	75.4	1935	67.6
1920	55.7	1928	76.3	1936	76.4
1921	51.9	1929	81.5	1937	80.9
1922	58.5	1930	73.5	1938	76.4
1923	65.8	1931	67.7	1939	83.7
1924	65.4	1932	57.4	1940	92.1
1925	71.3	1933	56.5	1941	106.2
1926	74.4	1934	62.0		

SOURCE: U.S. Department of Commerce.

TABLE 13.2. LINEAR TREND LINE FITTED TO GNP DATA

Year	GNP y	Time t	t^2	$t \cdot y$	y^2
1919	60 4	-11	121	-664 4	3648 16
1920	55 7	-10	100	-557 0	3102 49
1921	51 9	- 9	81	-467.1	2693 61
1922	58 5	- 8	64	-468 0	3422 25
1923	65 8	- 7	49	-460 6	4329 64
1924	65 4	- 6	36	-392 4	4277 16
1925	71 3	- 5	25	-356 5	5083 69
1926	74 4	- 4	16	-297 6	5535 36
1927	75 4	- 3	9	-226 2	5685 16
1928	76 3	- 2	4	-152 6	5821 69
1929	81 5	- 1	1	- 81 5	6642 25
1930	73 5	0	0	0	5402 25
1931	67.7	1	1	+ 67 7	4583 29
1932	57.4	2	4	114 8	3294 76
1933	56 5	3	9	169 5	3192 25
1934	62 0	4	16	248 0	3844 00
1935	67 6	5	25	338 0	4569 76
1936	76 4	6	36	458 4	5836 96
1937	80 9	7	49	566 3	6541 81
1938	76 4	8	64	611 2	5836 96
1939	83 7	9	81	753 3	7005 69
1940	92 1	10	100	921 0	8482 41
1941	106 2	11	121	1168 2	11278 44
Total	1637 0	0	1012	1292 5	120,113 04

$$\begin{aligned} \bar{t} &= 0 & n &= 23 & s_y t^2 &= \frac{\Sigma y^2 - a \Sigma y - b \Sigma ty}{n - 2} \\ \bar{y} &= \frac{1637.0}{23} = 71.17 & & & &= 93.20 \\ b &= \frac{n \Sigma ty - (\Sigma t)(\Sigma y)}{n \Sigma t^2 - (\Sigma t)^2} & & & s_y t &= \sqrt{93.20} = 9.65 \\ &= \frac{23(1292.5) - 0}{23(1012) - 0} = 1.277 & T(t) &= 71.2 + 1.3t \\ a &= \bar{y} - b \bar{t} = 71.17 \end{aligned}$$

The trend line estimated from the GNP data is

$$T(t) = 71.2 + 1.3t \quad (13.2)$$

Before this equation is used to estimate or predict GNP, the existence of a trend must be tested. In accordance with the point of view of Chap. 11, the two hypotheses are:

H_0 : The economy is a random process with no systematic identifiable forces.

H_1 : Systematic identifiable forces are present among the random forces in the economy.

In the present example, the systematic force is trend. The two hypotheses become, then:

H_0 : GNP is a random process with no trend component.

H_1 : GNP has a positive trend among the random forces.

If no trend component is present in the underlying economic mechanism, the *population* trend line is parallel to the x axis and $B = 0$; if a trend component is present, the trend line has a positive slope and $B > 0$. Therefore, the statistical test for a positive linear trend is a test for $B = 0$ and the two hypotheses become:

$H_0: B_0 = 0$ (no trend)

$H_1: B_1 > 0$ (positive trend)

$\alpha = .05 \quad n = 23 \quad n - 2 = 21$ D.F.

R : Upper tail (see H_1)

From Table 13.2, $\sqrt{\Sigma(t - \bar{t})^2} = \sqrt{1012} = 31.8$; $s_{y \cdot t} = 9.65$. The value of $t_{\alpha}(n - 2) = t_{.05}(21) = 1.72$ (from Table A.4). Replacing x by t in Eq. (10.8),

$$t = \frac{(b - B) \sqrt{\Sigma(t - \bar{t})^2}}{s_{y \cdot t}}$$

$$1.72 = \frac{(b_r - 0)(31.8)}{9.65} \quad b_r = .52$$

Decision Rule. Accept H_1 if $b \geq .52$; otherwise accept H_0 .

Decision. Accept H_1 because $b = 1.28$ in the sample falls in the region R . The data are adequate evidence of a trend force in GNP for the period 1919 to 1941. For these years the basic trend is linear and Eq. (13.2) is an estimate of the linear trend. This trend line is drawn in Fig. 13.2.

In review, the assumptions which are necessary to make this test valid are as follows:

1. The observations on GNP are random, independent drawings.
2. Values of y (GNP) are normally distributed in the universe about a trend line of the form $T(t) = A + Bt$; that is, for each value of t there is a normal distribution of y values with this mean.
3. The residual variance $\sigma^2_{y \cdot t}$ is the same for each of these normal distributions.

With these assumptions, the quantity $t = (b - B)/s_b$ is distributed according to the t distribution with $n - 2$ degrees of freedom.

Remark. The use of t as both the time variable in a time series and the symbol for the t distribution may be unfortunate. Careful reading, however, should decrease the possibility of confusion.

Assumption (1) is sometimes not satisfied in actual time series data because the successive deviations from the trend line are not independent of each other. Independence means here that if one observation is above the trend, there is no reason to suspect that the next observation also will be above the trend line. A *serial correlation* is the correlation

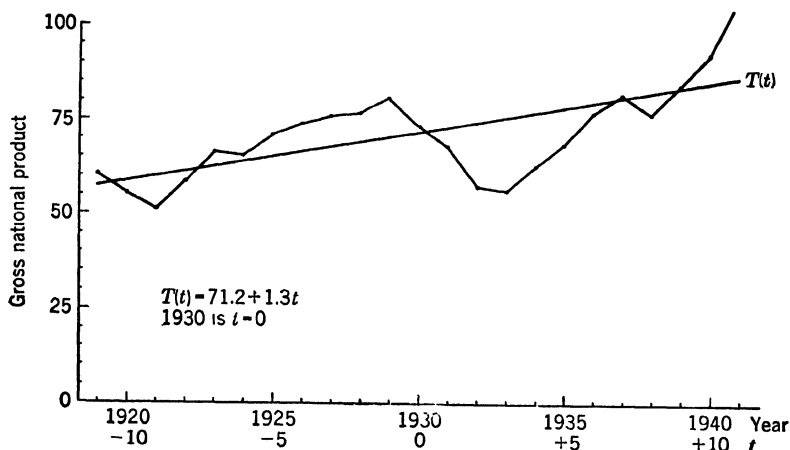


FIG. 13.2. Linear trend line fitted to GNP.

between each deviation from trend and the preceding deviation from trend. A high serial correlation coefficient is evidence of nonindependence, but this test is beyond the scope of this book.

EXERCISES

13.1. What is the meaning of a in the trend line for GNP? What is the meaning of b ? (*Hint:* Make an analogy with a regression equation. Caution: $t = 0$ is the year 1930.)

13.2. What would a value of $s_{y,t} = 0$ mean in the trend relationship between GNP and time? If $s_{y,t} = s_y$, what would be the slope of the trend line?

13.3. Estimate future values of GNP from Eq. (13.2) for the years 1942 to 1952 and enter them in a table. What assumptions are necessary for these estimates to be useful in prediction? From what you know about the economy, would you say that these estimates are high, low, or about right? What reason would you give for your answer to this last question?

13.4. Draw your own graph of GNP adding the following values for 1909 to 1918 and 1942 to 1952. Draw the straight-line trend line, Eq. (13.2), on the graph and extend it back to 1908 and forward to 1952. Does the straight-line trend based upon the 1919 to 1941 data fit the whole series from 1908 to 1952? Does the basic trend in the GNP data for the whole period appear to be linear, although with a different equation, i.e., different values of a and b ? What cautions does this suggest for the practice of projecting empirical trend lines into the future?

Year	GNP	Year	GNP
1909	33 1	1942	161 6
1910	34 6	1943	194 3
1911	35 3	1944	213 7
1912	37 9	1945	215 2
1913	39 0	1946	211 1
1914	38 8	1947	233 3
1915	39 5	1948	259 0
1916	49 5	1949	258 2
1917	63 1	1950	286 8
1918	79 9	1951	329 8
		1952	348 0

A Nonparametric Ranking Test for Trend.¹ Another test for a trend force in a mechanism giving rise to economic data is based upon ranks. It is really a test for the randomness of a series based upon all possible combinations of the observed values. Although the test neglects some information, just as did the rank correlation coefficient r' of Sec. 10.11, its use is more than justified by its lack of need for assumptions. No assumption is necessary either about the mathematical curve of the trend (e.g., whether it is linear) or about the form of the population distribution (e.g., whether it is normal). The test is only to establish that the mechanism generating the data has a gradual growth or decline—a trend—with no specification of what the trend may be.

In a time series, the time variable is already ordered $t_1, t_2, \dots, t_n$. The y variable is ordered $y_1, y_2, \dots, y_n$. If the y variable is now *ranked* from *smallest* to *largest*, a positive trend in a set of data will result in the occurrence of small rank values of y with small values of t , and of high rank values of y with high values of t . A negative trend will result in the occurrence of high rank values of y with small values of t and of small rank values of y with high values of t . For example, consider the two series in the following table. In one series the time order and rank

Time period	Positive trend		Negative trend	
	y values	ranks of y	y values	ranks of y
1	2	1	11	5
4	4	2	9	4
3	7	3	5	3
4	9	4	4	2
5	10	5	2	1

¹ H. B. Mann, Non-parametric Tests against Trend, *Econometrica*, vol. 13, pp. 246ff., 1945; M. G. Kendall, "Rank Correlation Methods," Charles Griffin & Co., Ltd., London, 1948.

order correspond perfectly: 1-1, 2-2, . . . , 5-5. In the other series, the inverse correspondence is perfect: 1-5, 2-4, . . . , 5-1.

A coefficient, which is +1 if the agreement between the time order and rank order is perfect and -1 if the disagreement is perfect, is defined by

$$\tau = \frac{2S}{n(n-1)} \quad (13.3)$$

where $S = 2M - \frac{1}{2}n(n-1)$

n = sample size

M = number of positive scores computed as follows:

The t series is in natural order; the rank series is in rank order. Start at the smallest rank (1), and count the number of ranks which are greater than 1 and which follow it in time order. Call this positive score m_1 .

TABLE 13.3. RANK TEST FOR TREND IN GNP

Year	GNP	Rank of GNP	Positive score (m)
1919	60.4	6	17
1920	55.7	2	20
1921	51.9	1	20
1922	58.5	5	17
1923	65.8	9	14
1924	65.4	8	14
1925	71.3	12	11
1926	74.4	14	9
1927	75.4	15	8
1928	76.3	16	7
1929	81.5	20	3
1930	73.5	13	6
1931	67.7	11	6
1932	57.4	4	8
1933	56.5	3	8
1934	62.0	7	7
1935	67.6	10	6
1936	76.4	17	5
1937	80.9	19	3
1938	76.4	18	3
1939	83.7	21	2
1940	92.1	22	1
1941	106.2	23	0
Total .		..	$M = 195$

Start with the next smallest rank (2), and count the number of ranks which are greater than 2 and which follow it in time order. Call this positive score m_2 . Repeat this process for all rank values. The sum of the m 's is equal to M .

For example, the GNP data of Table 13.1 are given in Table 13.3 and ranked from smallest to largest. The smallest GNP, 51.9 in 1921, is given the rank 1; the next smallest GNP, 55.7 in 1920, is given the rank 2. The largest GNP, 104.7 in 1941, is given the rank 23. The positive score for rank 1 is 20 because all 20 ranks following it in time are greater than 1. The positive score for rank 2 is 20 because of the 21 ranks following it in the table, only one rank (1) is less than 2; all the rest are greater than 2. The positive score for rank 3 is 8, since 8 ranks are greater than 3 and follow it in the table: 7, 10, 17, 19, 18, 21, 22, and 23. The total positive score is $M = 195$. From Eq. (13.4),

$$S = 2(195) - \frac{(23)(22)}{2} = 137$$

From Eq. (13.3),

$$\tau = \frac{2(137)}{(23)(22)} = +.54$$

This value of τ indicates the possibility of a positive trend in GNP, but it must be verified by statistical test.

For $n \leq 10$, probabilities associated with different values of S are shown in Table A.11 of the Appendix. For example, with $n = 10$, the probability that S will exceed 19 is .054; the probability that S will exceed 25 is .014, if there is no relationship between the two variables. The statistical decision rule can be formulated by using this table with the appropriate value of n and α . For n larger than 10, S converges to *normality* with

$$\mu_S = 0 \quad (13.5)$$

$$\sigma_S = \sqrt{\frac{n(n-1)(2n+5)}{18}} \quad (13.6)$$

if the series is random, or if there is no relationship between the two variables. To use the continuous normal probability distribution, a continuity correction is made by adding 1 to S if S is negative and subtracting 1 from S if S is positive. Then

$$Z = \frac{S \pm 1 - \mu_S}{\sigma_S} \quad (13.7)$$

To test for a trend in GNP, the normal distribution is used because $n = 23$. The test takes the following form.

$H_0: \tau = 0$ (no trend in GNP)

$H_1: \tau > 0$ (positive trend in GNP)

$\alpha = .05 \quad Z_\alpha = 1.645$

R : Upper tail (see H_1)

From Eq. (13.5), $\mu_s = 0$; from Eq. (13.6), $\sigma_s = \sqrt{23(22)(51)/18} = 37.9$. Since S is positive, the continuity correction is to subtract 1 from S . Then, from Eq. (13.7),

$$Z_\alpha = \frac{S_r - 1 - \mu_s}{\sigma_s} \quad 1.645 = \frac{S_r - 1 - 0}{37.9} \quad S_r = 62.3$$

Decision Rule. Accept H_1 if $S \geq 62.3$; otherwise accept H_0 .

Decision. Accept H_1 because $S = 137$. These data are evidence of a gradual long-term growth in GNP. The economic system generates larger average GNP figures with the passage of time. A trend force is inferred for this mechanism, but no particular form of trend is estimated.

Moving Averages in Trend Estimating and Data Smoothing. The method of linear trend estimation by least squares presented in Sec. 13.1 is formally equivalent to regression analysis. Other methods of trend fitting are often used because: (1) no mathematical curve is specified to represent trend (e.g., a straight line); or (2) only a smoothing of irregular fluctuations in the series is desired. A simple and widely used method is that of a *moving average*. This method is illustrated by using the fictitious data of Table 13.4 and Fig. 13.3.

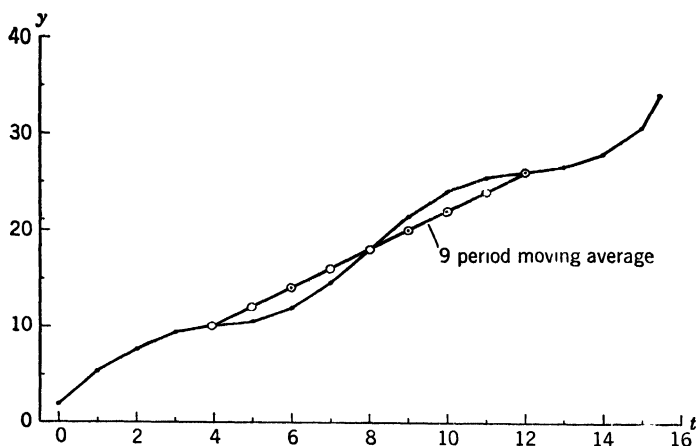


FIG. 13.3. Nine-period moving average fitted to data of Table 13.4.

From Fig. 13.3 it appears that the series has a *linear trend* and a cycle with a *fixed period* and a *constant amplitude*. A moving average will remove the cycle and represent the linear trend exactly if the *period of the moving average* is equated to the period of the oscillation. The period of the series is nine time units. The series starts at $t = 0$ and finishes one cycle at $t = 8$; it starts again at $t = 8$ and finishes another cycle at $t = 16$. A nine-period moving average should result in a complete removal of the cycle, leaving only the linear trend. This moving aver-

age is found by averaging the first nine observations from $t = 0$ to $t = 8$; then averaging $t = 1$ to $t = 9$; then $t = 2$ to $t = 10$, etc. Both the moving totals and moving averages are shown in Table 13.4 with the averages centered at the middle of the period, in this instance opposite the fifth observation. The moving average is plotted on Fig. 13.3.

TABLE 13.4. EXAMPLE OF MOVING AVERAGE FITTED TO TIME SERIES

Time t	Series y	Nine-period moving total	Nine-period moving average	Trend $T(t) = 2 + 2t$
0	2 0	.	.	2
1	5 4	.	.	4
2	8.0	.	..	6
3	9.4	.	.	8
4	10 0	90 0	10 0	10
5	10.6	108 0	12 0	12
6	12 0	126 0	14 0	14
7	14 6	144 0	16 0	16
8	18 0	162.0	18 0	18
9	21.4	180 0	20.0	20
10	24 0	198 0	22 0	22
11	25 4	216 0	24 0	24
12	26 0	234 0	26 0	26
13	26 6			28
14	28 0			30
15	30 6		.	32
16	34 0	...	...	34

The nine-period moving average removes the cycle completely, leaving only the linear trend which coincides with the trend line $T(t) = 2 + 2t$ from which the data were computed. Points are missing at both ends of the series—four points at the beginning and four points at the end. This is a disadvantage, for it means that the moving average cannot be kept up to date, i.e., recorded up to the end of the data without more refined methods. The loss of observations also limits the period of the moving average when used in very short series.

The illustration is hypothetical with perfect specifications (constant period and amplitude of the cycle component) for trend estimation by a moving average. In real economic time series, the existing oscillations will not have fixed amplitudes or periods, although the trend may often appear linear. Hence a moving average will not remove all the oscillation, although it will damp out some of the unwanted periodic fluctuations and thus leave an approximation to trend. When the period of the

cycle varies from one cycle to the next, there is no general rule for selecting a period for the moving average; the judgment, experience, and objective of the investigator must determine what it shall be.

A trend that is not basically linear may be represented as a smooth movement by a moving average. Minor fluctuations are damped out and the long slow changes appear in the moving average. The longer the period of the average selected, the smoother the resulting curve and, also, the more observations that are lost at each end of the series. When used to smooth a curve, a moving average gives an arc of best fit to the data. Refinements in the computing process may give different weights to each of the observations which are averaged, but the simplest case, and the only one given here, assigns equal weights to each observation. The peaks in the original data are lowered by a moving average; the troughs in the original data are raised by a moving average. This behavior, if not known to exist, may lead to incorrect statements about the range of the data if the original data do not accompany the moving average.

The moving average is a most common method of smoothing data or eliminating periodic movements if their period is known. It can be used to eliminate seasonal variations in monthly data because the period of the seasonal variation is, by definition, 12 months. If the trend is basically linear, a 12-month moving average will remove the seasonal variation and approximate the linear trend. If the trend is not linear, a 12-month moving average will represent the trend as a smooth curve. One danger in using a moving average to this end is that it will induce an apparent periodic oscillation into a random series. This is the Slutsky-Yule effect mentioned in Sec. 11.6; more will be said about it in Sec. 13.4.

Example 13.1. The National Bureau of Economic Research has been publishing books and papers on empirical economic research since the 1920s. Any student of time series analysis in business and economics should have some familiarity with its work. The example given below is taken from the Bureau's monumental work on business cycles both because it is a good illustration of how a moving average is used in the study of time series and because it will acquaint the reader with one phase of the Bureau's work.¹

"While we consider it desirable to economize effort in handling secular trends and random movements, experience has taught us not to economize effort by working with annual data. . . . annual data are exceedingly crude materials for comparing the cyclical behavior of different activities in the same period or of the same activity in different periods. They obscure timing relations, they make it impossible to trace cyclical patterns with confidence, often they obscure and sometimes they obliterate cyclical fluctuations. . . .

¹ Reprinted from Arthur F. Burns and Wesley C. Mitchell, "Measuring Business Cycles," pp. 43-46, National Bureau of Economic Research, Inc., New York, 1946, with the permission of the Bureau.

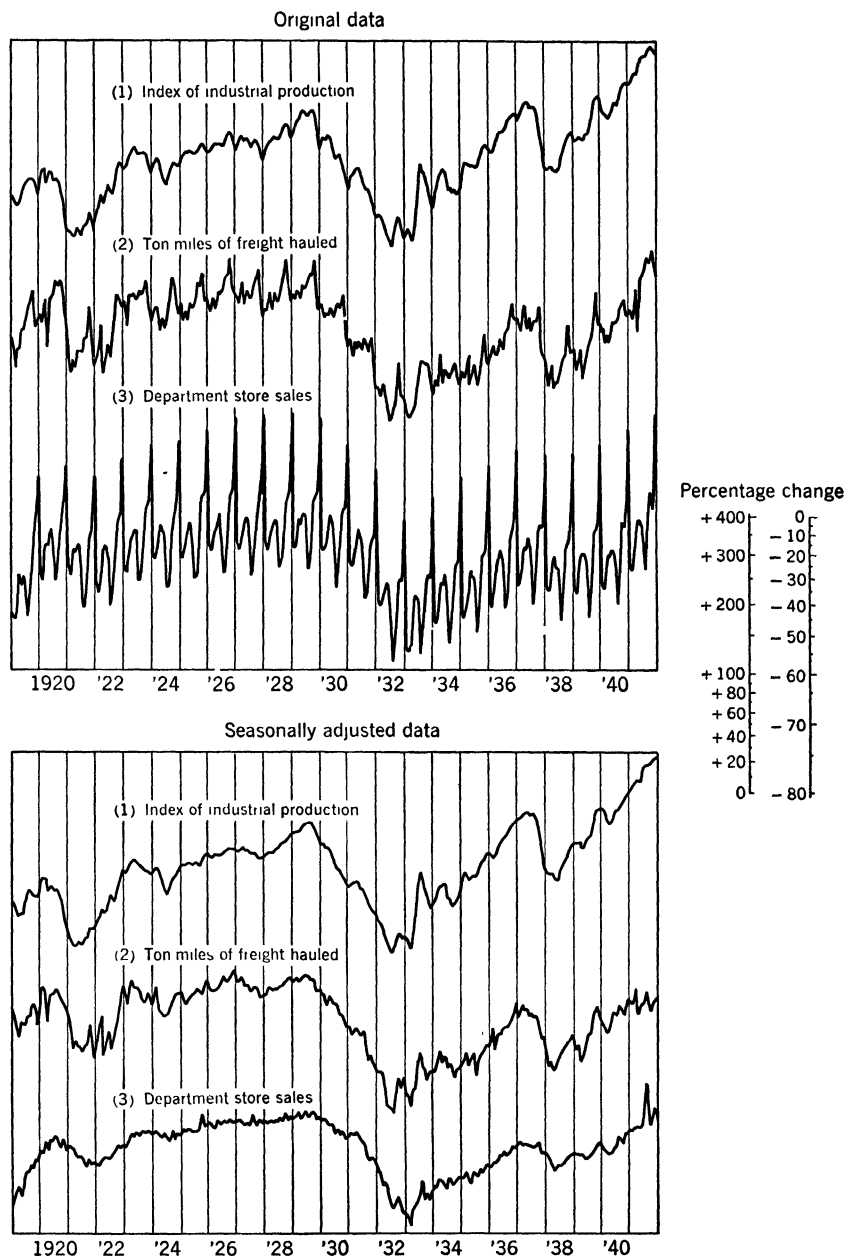


FIG. 13.4. Original and seasonally adjusted data for three American series: index of industrial production, ton-miles of freight hauled, and department-store sales. (SOURCE: This figure is reproduced from Arthur F. Burns and Wesley C. Mitchell, "Measuring Business Cycles," p. 45, National Bureau of Economic Research, New York, 1946, with the permission of the Bureau.)

“Monthly and quarterly records are more difficult to compile and to analyze by our standard technique than annual data; they also impose the burden of removing seasonal fluctuations. . . .

“A statistician who has struggled with seasonal adjustments of numerous time series is not likely to underestimate the part played by ‘hunch’ and ‘judgment’ in his operations. If pressed hard, he may admit that what he does, and not too satisfactorily in many instances at that, is merely to erase the repetitive intra-annual tendency that he observes in the original figures. Seasonally adjusted figures no more show what the behavior of time series would have been in the absence of seasonal forces than indexes of physical production show what production would have been in the absence of changes in relative prices; or than figures on per capita income show what income would have been in the absence of income inequalities. . . . That the removal of seasonal variations can facilitate analysis of the interrelations of cyclical movements in different activities will be plain to any reader who takes the time to study [Fig. 13.4], which depicts the movement in recent years of industrial production, railroad freight traffic, and department store sales in the United States before and after adjustment for seasonality. The seasonal variations follow an individual course in each series, and obscure the protracted fluctuations in freight traffic and department store sales. When the variations associated with the season of the year are removed, the underlying fluctuations in economic fortune come clearly to the surface and can be readily traced in the several series. . . .

“In ascertaining the seasonal variations of monthly data we use two principal methods. . . . Method (2) entails taking a twelve-month moving average of the original figures, placing each average in the seventh month of shifting twelve-month intervals, expressing the original figures as ratios to the ‘centered’ moving averages, and striking averages of these ratios, first for the Januaries, then the Februaries, and so on. . . .”

EXERCISES

13.5. Total factory sales of cars and trucks in the United States from 1924 to 1952 are given in the table below.

TOTAL FACTORY SALES, CARS AND TRUCKS
(In millions)

Year	Total	Year	Total	Year	Total
1924	3 60	1934	2 75	1944	.74
1925	4 27	1935	3 95	1945	.73
1926	4.30	1936	4 45	1946	3 09
1927	3.40	1937	4 81	1947	4 80
1928	4.36	1938	2 49	1948	5 29
1929	5 36	1939	3 58	1949	6 25
1930	3 36	1940	4.47	1950	8 00
1931	2.39	1941	4 84	1951	6.77
1932	1 37	1942	1.04	1952	5.54
1933	1.92	1943	.70		

a. Plot these data on arithmetic graph paper.

b. Fit a linear trend line to these data, letting $t = 0$ for 1938. Plot this trend line on the original graph. Should the years 1942 to 1945 be omitted in fitting a trend line? Should the years 1946 to 1952 be analyzed separately from 1924 to 1941?

c. Fit a linear trend line to either the period 1924 to 1942 or 1924 to 1952, omitting 1942 to 1945, according to your answer to part (b). Plot this trend line on the original graph. Test the hypothesis that factory sales of cars and trucks have no trend. What conclusions do you draw? What assumptions are you making?

d. If you decided to omit 1946 to 1952 in part (b), would you fit a separate linear trend line to these years? Would you fit any trend line to these years?

13.6. Apply the nonparametric ranking test for trend to the wholesale price index of all commodities, 1939 to 1953. Data are given in the table below.

Year	Index	Year	Index	Year	Index
1939	50 1	1944	67.6	1949	99 2
1940	51 1	1945	68 8	1950	103 1
1941	56 8	1946	78 7	1951	114 8
1942	61 2	1947	96 4	1952	111 6
1943	67 0	1948	104 4	1953	

What conclusions do you draw?

13.7. Fit a 15-year moving average to the housing starts plotted in Fig. 11.6. The actual number of starts are in the table on page 320, in thousands. Does this moving average remove most of the cycle from 1890 to 1938? What happens after 1938 and why does it happen? What conclusions do you draw from this analysis?

13.2 Cycles

A model of a particular time series may specify all or only some of the four traditional components of a time series (see Sec. 12.4). No seasonal component is present in the GNP data because they are yearly values and the seasonal variation recurs within each year. The model for GNP specifies only three components as in Eq. (13.8).

$$y = T(t) \cdot C(t) \cdot R(t) \quad (13.8)$$

Algebraically, $C(t)$ may be isolated by dividing through by the trend component, leaving

$$\frac{T(t) \cdot C(t) \cdot R(t)}{T(t)} = C(t) \cdot R(t) \quad (13.9)$$

The cycle-random components are combined in the remainder.

Year	Number of starts	Year	Number of starts	Year	Number of starts
1889	342	1910	387	1930	330
		1911	395	1931	254
1890	328	1912	426	1932	134
1891	298	1913	421	1933	93
1892	381	1914	421	1934	126
1893	267				
1894	265	1915	433	1935	216
		1916	437	1936	304
1895	309	1917	240	1937	332
1896	257	1918	118	1938	399
1897	292	1919	315	1939	458
1898	262				
1899	282	1920	217	1940	530
		1921	449	1941	620
1900	189	1922	716	1942	301
1901	275	1923	871	1943	184
1902	240	1924	893	1944	139
1903	253				
1904	315	1925	937	1945	208
		1926	849	1946	662
1905	507	1927	810	1947	846
1906	487	1928	753	1948	914
1907	432	1929	509	1949	989
1908	416				
1909	492			1950	1,352
				1951	1,020
				1952	1,068
				1953	1,068

The original GNP data and the trend estimates based upon the linear trend, $T(t) = 71.2 + 1.3t$, are given in Table 13.5. Each yearly observation of GNP is divided by its trend value according to Eq. (13.9), leaving the combined cycle and random components measured as a *percentage* of trend. These are also shown in Table 13.5 and plotted on Fig. 13.5. The cycle-random component oscillates about the 100 per cent level. The trough in 1921 is followed by a peak in 1929 and another trough in 1933. These turning points coincide with the recessions of 1921 and 1933 and the boom of 1929. As descriptions of GNP for this inter-war period, statements of this kind are both interesting and valid. Is the apparent cycle also a valid inference for the economic mechanism which generated the data?

A series without trend, either because the trend did not exist in the first place or because it has been removed, is called a *stationary* time series. This stationary series must first be tested as a random series

TABLE 13.5. PERCENTAGE DEVIATIONS FROM TREND IN GNP

Year	t	GNP y	Trend $T(t) = 71.2 + 1.3t$	Cycle and random $C(t) \cdot R(t) = y/T(t)$ per cent	First difference	Phase
1919	-11	60.4	56.9	106.2	-	
1920	-10	55.7	58.2	95.7	-	2
1921	-9	51.9	59.5	87.2	+	
1922	-8	58.5	60.8	96.2	+	2
1923	-7	65.8	62.1	105.9	-	1
1924	-6	65.4	63.4	103.2	+	
1925	-5	71.3	64.7	110.2	+	2
1926	-4	71.1	66.0	112.7	-	
1927	-3	75.4	67.3	112.0	-	2
1928	-2	76.3	68.6	111.2	+	1
1929	-1	81.5	69.9	116.6	-	
1930	0	73.5	71.2	103.2	-	
1931	1	67.7	72.5	93.4	-	4
1932	2	57.4	73.8	77.8	-	
1933	3	56.5	75.1	75.2	+	
1934	4	62.0	76.4	81.2	+	
1935	5	67.6	77.7	87.0	+	4
1936	6	76.4	79.0	96.7	+	
1937	7	80.9	80.3	100.7	-	1
1938	8	76.4	81.6	93.6	+	
1939	9	83.7	82.9	101.0	+	3
1940	10	92.1	84.2	109.4	+	
1941	11	106.2	85.5	124.1		

before the existence of a cycle is established. The same basic question that was asked about the egg-production series must be asked here. Are these drawings from populations with the same mean? Such drawings could show a pattern of fairly regular fluctuations only because of chance selection. This hypothesis must be refuted before the acceptance of the alternative hypothesis that the populations have different means which rise and fall in an oscillatory pattern when ordered in time.

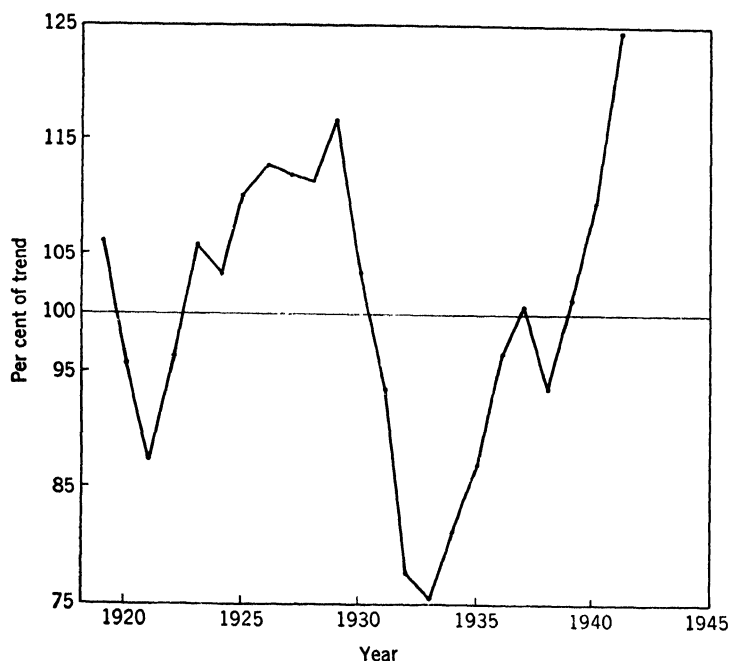


FIG. 13.5. Percentage deviations in GNP from the linear trend $T(t) = 71.2 + 1.3t$

A test for the cycle component in GNP is the phase test used for the random series in Sec. 11.6 and for seasonal variation in Sec. 12.1. The plus and minus signs assigned to the first differences of GNP and the lengths of the phases are recorded in Table 13.5. The expected number of phases is given by Eq. (11.4). In the test for a cycle in GNP, $n = 23$, and Table 13.6 compares the observed and expected number of phases and gives the value of χ_p^2 computed from Eq. (11.5).

The statistical test for the cycle is given below.

H_0 : No cyclical movement in GNP.

H_1 : GNP has a cycle.

$\alpha = .05 \quad n = 23 \quad \chi_{.05}^2 = 6.898$ (from Table A.8)

Decision Rule. Accept H_1 if $\chi_p^2 \geq 6.9$; otherwise accept H_0 .

Decision. Accept H_0 because $\chi_p^2 = 6.3$, and conclude that the economic mechanism which generates GNP does not exhibit a cycle force, as evidenced by the oscillation in GNP from 1919 to 1941.

TABLE 13.6. PHASE TEST FOR A CYCLE IN GNP

Length of phase	Observed number u	Expected number U	$u - U$	$(u - U)^2$	$(u - U)^2/U$
1	3	8 33	-5 33	28 41	3 41
2	4	3 48	52	27	08
3 or more	3	1 18	1 82	3 31	2 81
Total	10				$\chi_p^2 = 6 30$

EXERCISES

13.8. If you accepted H_0 in part (c) of Exercise 13.6, test for a cycle in total factory cars and sales. If you accepted H_1 , express each yearly total as a percentage of its trend according to Eq. (13.9). Test this stationary series for a cycle. What conclusions do you draw?

13.9. Apply the nonparametric test for a cycle to housing starts plotted on Fig. 11.6. Assign the + and - signs to the first differences directly from this figure or from the data given in Exercise 13.7. What conclusions do you draw?

13.3 Some Comments on the Methods of Analysis

The statistical analyses in the preceding sections of this chapter led to the conclusions that a positive trend is present in GNP but that the pattern of the apparent cycle is not evidence of a cycle force in GNP. The long-term trend is substantiated by the data, but the oscillation about this trend is consistent with a random variation and not orderly enough to be a cycle. Since this conclusion about the cycle may appear heretical, a few comments are in order.

As in all statistical decision rules, the acceptance of H_0 or H_1 depends upon the more or less arbitrarily selected value of α , the probability of accepting H_0 when, in fact, H_1 is true. Two different analysts may arrive at entirely opposite conclusions by selecting different values of α . They evaluate differently the seriousness of the two kinds of error and the willingness to commit each. Thus differences arise even among men of good judgment. This is true in any decision process, for the choice of α is a personal matter and a sufficiently large value will make H_1 always acceptable. The GNP analysis points clearly to the necessity for selecting α in advance. Anyone who really wants to detect a cycle in GNP

for the years 1919 to 1941 need only set $\alpha \geq .1$; if he wants to deny the existence of a cycle, he need only set $\alpha \leq .05$. The value of α should be a prior commitment as to the willingness to run the risk of concluding that a cycle exists when, in fact, it does not exist. It should not be used *ex ante* to substantiate a foregone conclusion.

It is entirely possible in this problem to make the error of accepting H_0 when, in fact, H_1 is correct. That is to say, the decision that no cycle is present may be incorrect because a cycle really does characterize GNP. And there may be a large probability of concluding that there is no cycle when, in fact, a cycle is present, although this will never be known. One cause of a large probability β is a small sample, other things being equal. Since GNP data are available for 1909 to 1918 and 1942 to 1952, the sample size may be increased. With this larger sample, the probability of detecting a cycle in GNP if it exists will increase and β will decrease. A new trend line must be fitted to the whole series and a new measurement of the cyclical-random component made. This new trend will not be linear, as indicated in an earlier exercise.

In spite of the evidence of the test for the cycle, analysts with the descriptive approach may still argue that the GNP data do oscillate around the trend line and that this is enough evidence of a cycle. The basic fallacy here is that the order of GNP data measured as percentages of a linear trend could occur with fairly high probability (greater than α) in random samples from a universe in which there is no cycle. A cycle force is a property of the universe of values from which the sample is selected. The position which has constantly been maintained in this book is that any inference about any population characteristic must be based upon a probability statement and that such inferences are not to be confused with descriptions of the sample.

Economists and others who have lived through the periods of the 1920s and 1930s may say that this is nonsense. The deep recession of 1932 to 1933 as well as the sharp but short recession of 1921 is certain evidence of a fluctuating GNP. No statistical test of the randomness of a series will deny for them the seriousness of these movements and the economic chaos created during these low periods. The statistical test, however, does not deny the actual pattern of rises and falls in the data observed. It says only that, for the period 1919 to 1941, the economic system (the statistical population) which generated the GNP data gave high and low values in such an order that they are not inconsistent with drawings of a random series, i.e., drawings from populations with the same mean each year. If an oscillation is to be inferred for GNP, it must be inferred for the economic system from something other than the order of the historical fluctuations. *On the basis of the data themselves, the cycle cannot be isolated from the random component.*

Remark. Random fluctuations may have serious consequences. Earthquakes may occur at random intervals of time, but when they do occur, they often cause much damage and suffering.

In the light of the actual fluctuations in this period, one may even question the use of the hypothesis that the deviations from the GNP trend are a random series as a plausible explanation of these economic data. No one is going to embark on a long and involved analysis of a set of data which he believes to be random. This test is not without its usefulness, however. For if the pattern of the data is found not to be statistically significant on the hypothesis of random variation, it will not be statistically significant on the hypothesis of a systematic component. The nonrandomness of a series must be tested against the natural alternative of randomness before a systematic component is estimated if the estimation is to be anything but an act of pure faith or simply a description of past events.

The real danger is that a cycle may be inferred and projected into the future from mere contemplation of the data. From the trough of 1921 to the trough of 1933 there was a period of 12 years. Should the next trough occur 12 years later in 1945? Is this a valid projection in 1941? It is known now that World War II made this projection erroneous. Even in 1941, anyone sufficiently conversant with the economic system in the United States to be working with GNP data would have been aware of the possibility of war and of its effect on GNP. Mere contemplation and manipulation of the GNP series is not enough. Statistical tests which take into account the variability in a series are helpful restraints on man's propensity to search for a "cause" and then to project his conclusions into the future.

13.4 Explanation of the Cycle

Some economists and statisticians answer no to the question: Are these methods of analyzing the cyclical and random components of a time series really valid? Some feel that any component strong enough to be isolated is obvious and needs no statistical treatment. Others prefer to make models with built-in explanations of the cycle and random components rather than rely upon separate causal factors. The mathematics of a cyclical mechanism for a series is beyond the scope of this book, but two possible explanations of the cyclical mechanism are briefly discussed.

One explanation for an apparent cycle in a series is that the method of handling the data builds an oscillation into the data. The most outstanding example is the Slutsky-Yule effect, where a summing or averaging of successive values of a random series will result in a cyclical behavior. The moving average is a common method of analyzing data, and the series resulting from such an analysis may appear to be oscillatory

not because a real cycle exists in the series but because it was built in by the method of analysis. The periods and amplitudes of random series which have been averaged are not fixed or constant, but very often appear to be like actual time series data. The very act of eliminating a trend or smoothing data with a moving average may induce an artificial cycle into the data.

Another model of a time series having a cycle is called an autoregressive mechanism. The mechanism of an autoregressive system may be likened to the vertical motion of an automobile traveling along a bumpy but flat road. Springs on the automobile permit the car to oscillate up and down over bumps, damping out the movement to the comfort of the passengers. A bump will cause the car to oscillate up and down, and the springs bring the car back to rest in the vertical movement. A continual succession of bumps will cause a vertical oscillation with varying magnitudes or amplitudes and varying periods if the bumps are not equidistant but occur in varying intervals along the road. The result is that the oscillations are continually being damped out by the springs but also continually renewed by the bumps, which may be quite randomly distributed along the road. The behavior of a time series in this manner appears plausible to some people. The regularity of the series is determined by the internal structure of the economic mechanism; the motion itself is generated by external shocks.

13.5 The Separation of Trend and Random Forces

The previous section indicates that the traditional methods of separating the cyclical and random components of time series have been under suspicion for some time. More recently, separation of the trend and random forces of a series has been questioned in an article by M. G. Kendall.¹

Professor Kendall states his purpose and some of his major conclusions as follows:

It has been customary to analyze an economic time-series by extracting from it a long-term movement, or trend, for separate study and then scrutinizing the residual portion for short-term oscillatory movements and random fluctuations. The assumption in this procedure is that the long-term and short-term movements are due to separate causal influences and therefore that the mathematical process of analysis corresponds more or less roughly to a real distinction of type in the generative system. When it began to appear that the classical method of analyzing stationary economic fluctuations into cycles and residual elements broke down, and that better accounts of such movements were given by autoregressive schemes in which the disturbance element played

¹ M. G. Kendall, *The Analysis of Economic Time-series—part I: Prices*, *Journal of the Royal Statistical Society*, vol. 66, series A, part I, pp. 11–34, 1953. Parts of this article are reprinted with the permission of the Honorary Editors of the *Journal*.

an integral part, it was an easy step to proceed one stage further, and to enquire whether the so-called trend in a series was in fact separable from the short-term movements, or whether it should be regarded as generated by a set of forces which also gave rise to the short-term movements. . . .

Broadly speaking, the results are these:

(a) In series of prices which are observed at fairly close intervals the random changes from one term to the next are so large as to swamp any systematic effect which may be present. The data behave almost like a wandering series.

(b) It is therefore difficult to distinguish by statistical methods between a genuine wandering series and one wherein the systematic element is weak.

(c) Until some way has been found of circumventing this difficulty, trend fitting, and perhaps the fitting of any model, is a highly hazardous undertaking. It may be possible for an econometrician to test whether the data agree with a hypothesis suggested by prior analysis, but it may be impossible to discriminate between quite different hypotheses which all fit the data. . . .

(e) An analysis of stock exchange movements revealed little serial correlation within series and little lag correlation between series.¹ Unless individual stocks behave differently from the average of similar stocks, there is no hope of being able to predict movements on the exchange for a week without extraneous information. . . .

Professor Kendall analyzed a price series of wheat extending from 1883 to 1934 with the result that the frequency distribution of 2379 first differences was almost perfectly symmetrical with an appearance of normality.

At first glance the implications of these results are disturbing. If the series is homogeneous, it seems that the change in price from one week to the next is practically independent of the change from that week to the week after. This alone is enough to show that it is impossible to predict the price from week to week from the series itself. And if the series really is wandering, any systematic movements such as trends or cycles which may be "observed" in such series are illusory. The series looks like a "wandering" one, almost as if once a week the Demon of Chance drew a random number from a symmetrical population of fixed dispersion and added it to the current price to determine next week's price. And this, we may recall, is not the behavior in some small backwater market. The data derive from the Chicago wheat market over a period of fifty years during which at least two attempts were made to corner wheat, and one might have expected the wildest irregularities in the figures. To the statistician there is some pleasure in the thought that the symmetrical distribution reared its graceful

¹ Serial and lag correlations will be mentioned frequently in the remainder of this discussion. A serial correlation is a correlation between the observations of a time series at two different time periods, e.g., the price of a stock in one month versus the price of the same stock a month earlier. A lag correlation is a correlation between the observations of one series and the observations of another series at different time periods, for example, the price of steel stocks and the price of automobile stocks 2 months earlier |

head undisturbed amid the uproar of the Chicago wheat-pit. The economist, I suspect, or at any rate the trade cyclist, will look for statistical snags before he is convinced of the absence of systematic movements. And he will be very right to do so.

On the use of yearly versus shorter-term observations, Professor Kendall concludes:

In the past I have often wondered whether annual figures were much use in the study of economic phenomena, except of course in relation to intermittent variables like crop-yields. So much can happen in a year that one feels the underlying causal system to require examination under a finer structure than an annual interval. But if this wheat-series is any guide, it seems that what we gain by observing the phenomenon at short intervals is lost, or at any rate deeply obscured, by the stochastic discontinuity of the process. It may be that the motion is genuinely random and that what looks like a purposive movement over a long period is merely a kind of economic Brownian motion. But economists—and I cannot help sympathizing with them—will doubtless resist any such conclusion very strongly. We can at this point suggest only a few conclusions:

- (a) the interval of observations may be very important;
- (b) it seems a waste of time to try to isolate a trend in data such as these;
- (c) prediction in such a series, from interval behavior alone, is subject to a wide margin of error and the best estimate of the change in price between now and next week is that there is no change.

Several of Professor Kendall's colleagues participated in a general discussion of the paper. Some expressed agreement with his general conclusions. Some disagreed with him, believing that his conclusions followed from the nature of his data on speculative market prices and that prices in nonspeculative markets might not appear random. Others disagreed because, as they expressed it, Kendall had no theoretical framework of the interrelations among variables with which to approach an analysis of his data. Their comments are reproduced in the original article and the interested student is left to his own devices to read them.

13.6 Some Directions of Research

The analysis of time series started out by following the traditional approach: the extraction of long-term movements (or trends) and seasonal movements, and the examination of the residual for oscillatory and random movements. It was shown that the descriptive approach of these methods is not enough, but that the existence of systematic components must be tested before inferences may be drawn about the economic mechanism which generates the data. Finally, the discussion pointed out that it has been known for some time that these methods break down in many problems and that more satisfactory methods are

required. Complete agreement among economists or statisticians on the nature of the more satisfactory methods does not exist. The search has led in many directions. Some of the methods involve advanced mathematics and statistics and a good working knowledge of economic theory, which are not prerequisites to an elementary course in statistics. All the methods require some knowledge of the traditional approach before the basic problems for which the newer methods are developed can be thoroughly understood.

The National Bureau of Economic Research continues to employ a thoroughly empirical approach. (An example of the Bureau's work was given in Sec. 13.1.) The many books, pamphlets, and monographs of this organization present systematically collected measurements of various aspects of the cyclical movements displayed in a large number of different economic time series. The monumental work of this group is "Measuring Business Cycles" by Arthur F. Burns and Wesley C. Mitchell.

In essence the Bureau's approach to the study of cycles is based on a table of reference dates showing the months and years when business cycles reached peaks and troughs. For the United States, these reference dates were first derived from the business annals compiled for the Bureau by Willard Thorp, but they have since been refined, tested, and revised in accordance with the behavior of statistical series. For an individual series the Bureau computes a specific cycle relative and compares

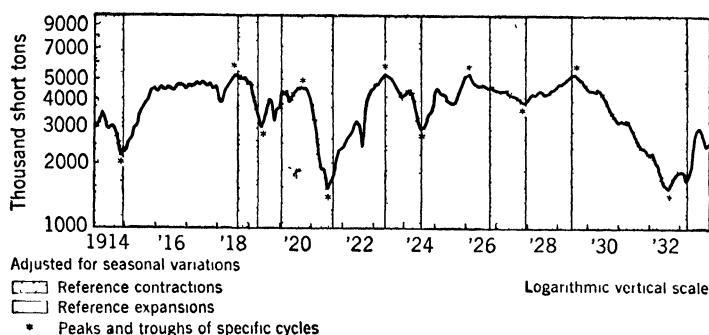


FIG. 13.6. Coke production in the United States, 1914 to 1933, with the turning points of business cycles and the specific cycles in coke production. (SOURCE: *This figure is reproduced from Arthur F. Burns and Wesley C. Mitchell, "Measuring Business Cycles," p. 25, National Bureau of Economic Research, New York, 1946, with the permission of the Bureau.*)

it with the reference cycle, examining the turning points for leads and lags, and finally averaging different phases of the specific cycles. An example of this procedure is given in Fig. 13.6, where the specific cycles in coke production are compared with the shaded reference cycles. As

shown by the asterisks on the graph, coke production tends to follow general business activity very closely.

The Cowles Commission for Research in Economics is the leading exponent in the United States of a sophisticated probability formulation of a mathematical model which attempts to supply a theoretical foundation for the *interrelations* between economic variables and then proceeds to measure the relationships. The Commission's monumental volume is "Statistical Inference in Dynamic Economic Models," edited by T. C. Koopmans (John Wiley & Sons, Inc., New York, 1950). A simple explanation of this point of view is given by E. G. Bennion, The Cowles Commission "Simultaneous Equation Approach": A Simplified Explanation (*The Review of Economics and Statistics*, vol. 34, pp. 49-57, 1952). John R. Meyer and Henry Laurance Miller, Jr., have discussed and amplified Bennion's article in Some Comments on the "Simultaneous-equations Approach" (*Review of Economic Statistics*, vol. 36, no. 1, pp. 88-92, February, 1954).

An example from Bennion illustrates the application of this method to the problem of measuring the marginal propensity to consume, i.e., how much of each additional dollar of income is spent on consumption. Assuming that consumption is a linear function of disposable income and subject to a random disturbance, the equation relating consumption and income is

$$c_t = \beta + \alpha y_t + \mu_t \quad (13.10)$$

where c_t = consumers' expenditures in constant dollars per capita

y_t = disposable income in constant dollars per capita

β = c intercept of the consumption function

α = marginal propensity to consume

μ_t = random element in consumers' behavior

t = time period

The old-fashioned regression method estimates only α and β in this equation. The newer econometric method says that this is not at all satisfactory because the estimates of α and β so derived are biased. Consumption expenditures and disposable income are not independent of each other, but are both related to investment expenditures. The dependence upon investment should be written into the theoretical structure of the consumption relationship. Economic theory says that income should equal consumption plus investment. If z_t = investment expenditures in constant dollars per capita, this relationship is expressed by

$$y_t = c_t + z_t \quad (13.11)$$

The investment expenditure is said to be autonomous or *exogenous* to the system. That is to say, it affects but is not affected by the *endogenous*

variables, consumption and income. Equations (13.10) and (13.11), taken together as

$$\begin{aligned}c_t &= \beta + \alpha y_t + \mu_t \\y_t &= c_t + z_t\end{aligned}$$

are the *economic structure* or the theoretical model which must be studied to estimate α and β of the consumption function. This structure expresses the interrelations among income, investment, and consumption.

A beginning student in statistics can hardly be expected to follow in detail the unsettled controversy over time series methods. He should, however, be aware that the traditional approaches may not be without fault; that they may at best represent merely a description of a set of data and at worst an enormous waste of time and effort and an illusory explanation of systematic movements.

REFERENCES

- Burns, Arthur F., and Wesley C. Mitchell: "Measuring Business Cycles," National Bureau of Economic Research, New York, 1946.
- Davis, Harold T.: "The Analysis of Economic Time Series," Bloomington, Indiana, 1941.
- Kendall, M. G.: "The Advanced Theory of Statistics," 3d ed. vol. II, Chaps. 29 and 30, Charles Griffin & Co, Ltd, London, 1951.
- Klein, Lawrence R.: "A Textbook of Econometrics," Row, Peterson & Company, Evanston, Ill, 1953.
- Koopmans, T. C. (ed.): "Statistical Inference in Dynamic Economic Models," John Wiley & Sons, Inc., New York, 1950.
- Tintner, Gerhard, "Econometrics," John Wiley & Sons, Inc., New York, 1952.

CHAPTER 14

INDEX NUMBERS¹

The cost of living in the United States went up again, reaching the highest peak yet. The Bureau of Labor Statistics reported last week that its index (1947-1949 = 100) increased from 114.7 to 115 with meat, rent, and increased New York subway fares accounting for much of the rise (fresh vegetables, home furnishings, and women's clothing declined a bit). In news accounts of the BLS report, such headline phrases as "new high" tended to obscure the fact that over the past year the retail price level has changed very little.

Industrial production rose slightly since last month, but was still about 9 per cent below a year ago, the Federal Reserve Board reported. The board's index which measures the output of the nation's mines and factories stood at 124 per cent of the 1947-1949 average. This was 1 point higher than last month but 12 points lower than in the same month last year. Durable goods output showed no change this month, mainly because an increase in steel production was offset by a decline in auto assemblies, the Board said.

Each month when the Bureau of Labor Statistics issues its latest index of consumer prices and the Federal Reserve Board its index of industrial production, major news services, newspapers, and magazines carry reports similar to the above. Such reports give rise to several important questions. For example, these questions may be asked about the Consumer Price Index.

1. What is meant by the cost of living or the retail price level?
2. What is the meaning of 1947-1949 = 100?
3. The index rose from 114.7 to 115. Is this an increase in dollar expenditures for an individual or for some group of individuals? Exactly what does it mean?
4. Most consumers feel the rise in meat prices. How does an increase in rent affect home owners, or a rise in New York City subway fares non-residents of New York using other transportation systems?

¹ This chapter may be studied immediately following Chap. 7.

5. What is the relevance of the index and its movements to an individual consumer concerned over his cost of living and desiring an indicator of what is happening to it?

The same kind of questions may be posed for the index of industrial production. These and other questions will be answered after a brief discussion of the elementary principles of index numbers.

14.1 A Simple Price Index Problem

Suppose that an individual has his rent receipts for the 15 years from 1938 to 1953. The rental unit is leased for a period of one year, beginning with July, so that one yearly rental figure describes the rent situation. The dollar rentals given in Table 14.1 show an increasing monthly

TABLE 14.1. RENTAL PAYMENTS, 1938 TO 1953

Year	Monthly rental	Year	Monthly rental
1938	\$35 00	1946	\$52 50
1939	37 50	1947	55 00
1940	40 00	1948	60 00
1941	45 00	1949	70 00
1942	50 00	1950	75 00
1943	50 00	1951	80 00
1944	50 00	1952	77 50
1945	50 00	1953	82 50

rent for this individual from 1938 to 1953, except for the war years 1942 to 1945 when rent was frozen at \$50 a month. Comparisons of rents over the period are often made by putting each yearly rent on a relative basis.

The Notion of a Relative. How much is 1953 rent relative to 1938 rent? Expressed as a percentage, the relative is

$$\frac{\text{Rent}(1953)}{\text{Rent}(1938)} (100) = \frac{82.50}{35.00} (100) = 236$$

In other words, the 1953 rent is 236 per cent of the 1938 rent. The 1953 rent relative to the 1942 rent is

$$\frac{R(1953)}{R(1942)} (100) = \frac{82.50}{50.00} (100) = 165$$

or the 1953 rent is 165 per cent of the 1942 rent. A *relative* is a simple ratio between two numbers expressed as a percentage and is interpreted like any other percentage.

A Time Series of Relatives. A time series of relatives expresses each yearly rental as a relative of the rental in one particular year, which is

called the *base year* or *base period* and given the value 100 as a reference point. For example, if the year 1938 is selected as the base period, each yearly rental is expressed as a relative of the 1938 rental. The 1938 rent relative to itself is obviously $100: (35.00/35.00)(100) = 100$. The 1953 rent relative to 1938 has already been calculated as 236. The time series of relatives for each of the other years with $1938 = 100$ is given in Table 14.2.

TABLE 14.2. RELATIVE RENTS: 1938 TO 1953

Year	Rent index (1938 = 100)	Rent index (1947 = 100)
1938	100	64
1939	107	68
1940	114	73
1941	128	82
1942	143	91
1943	143	91
1944	143	91
1945	143	91
1946	150	95
1947	157	100
1948	171	109
1949	200	127
1950	214	136
1951	229	145
1952	221	141
1953	236	150

The actual prices (rents) cannot be ascertained from Table 14.2, because a series of relatives does not reveal absolute magnitudes but merely prices relative to the base period. The actual dollar rental of any year can be determined only if the dollar rental in the base period, 1938, is known. The relative computed for each year in Table 14.2 is a simple *index number*. Each relative is an index of the yearly rental expressed as a percentage of or relative to the base period. Each index has meaning only in terms of its relation to the base period.

The 1938 base period selected for this index number is a prewar time period. In general, base periods should not be allowed to get too out of date or become atypical of the series studied. A postwar base period for the rent index could be 1947. The index series, recomputed on this base, is also shown in Table 14.2. Again the interpretation is in relative terms. The 1953 rent is now 150 per cent of the 1947 rent; the 1938 rent is only 64 per cent of the 1947 rent. This series, like the other, does not show dollar rentals, but only the rent in any one year relative to the 1947 rent.

Meaning of Index Series. The statistical problem of constructing an index of relatives and interpreting it in percentage or relative terms is simple. What does the series mean? The dollar price of housing has increased from 1938 to 1953. Has the "cost of housing" increased? Is the individual better or worse off in respect to housing? The answers to these economic questions involve value judgments, and are complicated by several factors. Is the contractual monthly rent a payment for the same size apartment, of the same age and location, providing the same services from year to year over the span of years? If not, the higher dollar rent is simply a payment for better housing service. This is equivalent to saying that the basic commodity, housing, must remain the same over the 15 years to permit an adequate comparison of housing costs.

Has the individual's income increased over the same period of time? Even for the same basic house, the higher dollar rents may not be a higher cost of housing because of a rise in income. During the war, when rent controls—a rationing process—froze rents at approximately the same level for several years, alternative housing was scarce. How does this affect the housing cost? Although these questions will not be answered here, they serve to illustrate the practical economic considerations which can arise in a relatively simple statistics problem and to point out pitfalls which become more serious in the interpretation of more complex index number formulations.

EXERCISE

14.1. The total number of civilian radio sets in use at the end of each year from 1922 to 1952 is given below in thousands of sets. Select one year as a base period,

Year	Number of sets	Year	Number of sets	Year	Number of sets
1922	400	1932	18,000	1942	59,300
1923	1,100	1933	22,000	1943	58,000
1924	3,000	1934	26,000	1944	57,000
1925	4,000	1935	30,500	1945	56,000
1926	5,700	1936	33,000	1946	60,000
1927	7,000	1937	37,600	1947	66,000
1928	8,500	1938	40,800	1948	74,000
1929	10,500	1939	45,300	1949	81,000
1930	13,000	1940	51,000	1950	99,000
1931	15,000	1941	56,000	1951	100,000
				1952	114,500

and express each total as a relative of that base year total. Interpret several of these relatives. Why does the relative decline from 1943 to 1945?

14.2 A Price Index of Several Commodities

The next step in the index-number problem involves the purchase of several commodities by one individual. To simplify the problem, assume that only three different foods are bought. Table 14.3 shows the prices paid for potatoes, pepper, and peas at three different time periods. The price of any commodity is represented by the symbol p_{ij} , where i refers to the commodity ($i = 1, 2, 3$), j refers to the time period ($j = 0, 1, 2$) and $j = 0$ is the base period. For example, p_{10} is the price of potatoes in the base period; p_{21} is the price of pepper in time period 1; and p_{32} is the price of peas in time period 2. The subscript consists of two serial numbers which identify the commodity and the time period for which the price is quoted. The individual buying this budget of three commodities is concerned with the price changes in his budget. He would like a measure of how the prices change with time; that is to say, he would like to have an index of the changes in the prices of the commodities which he regularly purchases as part of his budget.

TABLE 14.3. FOOD PRICES FOR AN INDIVIDUAL BUDGET

Commodity $i =$	Price in cents per pound		
	Base period $j = 0$	Period 1 $j = 1$	Period 2 $j = 2$
Potatoes 1	9	18	18
Pepper 2	41	41	82
Peas 3	10	11	10
Total	60	70	110

A Simple Price Index for One Individual. A simple price index can be made by merely adding up the prices of the three commodities in each of the three time periods and comparing them to one of the time periods used as a base period or a reference point. These computations are shown below. The indexes for the periods are 100, 117, and 183, respectively.

<i>Time period</i>	<i>Total of prices</i>	<i>Index (0 = base)</i>
0	60	$(60/60)(100) = 100$
1	70	$(70/60)(100) = 117$
2	110	$(110/60)(100) = 183$

The total of the prices per pound of the three commodities in period 1 is 117 per cent of the base total, and in period 2 is 183 per cent of the base total.

The index computed from a simple sum of prices is often not very satisfactory because it may not accurately represent a typical budget. Pepper, for example, is a relatively minor item, usually bought in quantities considerably less than a pound. Yet the index changed sharply from period 1 to period 2 (117 to 183) because the price of pepper doubled, rising from 41 cents to 82 cents a pound. On the other hand, the more staple item, potatoes, doubled in price between period 0 and 1 (from 9 to 18) with relatively little effect on the index. High-priced items tend to dominate this index. Since prices may be quoted in various units of measure, the index compiler arbitrarily determines which commodity dominates. If potatoes were quoted in cents per 100-pound sack, they would dominate the index. For this reason prices in an index are customarily *weighted* by a measure of the importance of each commodity, usually the quantity purchased, sold, or produced, depending upon the purpose of the index.

Weighted Price Indexes. The data are changed to illustrate weighted price indexes. A more representative budget item is substituted for pepper, and both prices and *quantities consumed* are given for each product in Table 14.4. The symbols p_{ij} and q_{ij} refer, respectively, to the price and quantity of the i th commodity in the j th time period, and $j = 0$ is the base period. There are various possible ways of making a price index using the quantity weights, depending upon the kind of question that is asked.

TABLE 14.4

Commodity $i =$	Time period					
	$j = 0$		$j = 1$		$j = 2$	
	p	q	p	q	p	q
Potatoes 1	10	2	15	3	15	1
Squash 2	15	3	20	4	10	3
Pens 3	20	4	12	5	15	4

Suppose the individual were to add up the products of the prices and quantities consumed in each time period and then make an index by expressing the totals for periods 1 and 2 as relatives of the total for the base period 0. The total amounts spent in each period are

$$\text{Period 0: } 10(2) + 15(3) + 20(4) = 145$$

$$\text{Period 1: } 15(3) + 20(4) + 12(5) = 185$$

$$\text{Period 2: } 15(1) + 10(3) + 15(4) = 105$$

Using the total amount spent in period 0 as the base of the relatives, the indexes are

Period 0: 100

Period 1: $185/145 = 128$

Period 2: $105/145 = 72$

The total amount spent in time period 1 is 128 per cent and, in time period 2, 72 per cent of the total amount spent in the base period. How much money is spent in each period is not known from the index itself; only the amount relative to or expressed as a percentage of the amount spent in the base period is known. In general, this index is expressed by

$$\frac{\sum p_{1j} \cdot q_{1j}}{\sum p_{10} \cdot q_{10}} \quad (14.1)$$

If the individual wishes to have an index of how much he spends on his budget of goods each month, this index is a good one. Throughout this discussion, the purpose has been to measure changes in *prices*. The index computed above is not a good one to measure price changes because it may change either with a change in the prices of the commodities or the quantities consumed. What part of a change in the index is induced by a change in price and what part by a change in the quantity consumed is not readily apparent. For this reason, the individual should redefine the method of computing an index number consistent with the purpose of measuring only price change.

Suppose that the individual decides to use the quantities consumed in the base period as representing a typical budget for him --a sort of standard market basket of goods which he regularly buys. He decides to price this fixed basket of goods in each time period and to make an index number by expressing the total cost of this same basket in each period as a percentage of the base period cost. The total costs of the period 0 market basket in each time period are

Period 0: $10(2) + 15(3) + 20(4) = 145$

Period 1: $15(2) + 20(3) + 12(4) = 138$

Period 2: $15(2) + 10(3) + 15(4) = 120$

In percentages, the indexes for the three periods are

Period 0: 100

Period 1: $138/145 = 95$

Period 2: $120/145 = 85$

The total cost of the fixed market basket of goods consumed in period 0 is 95 per cent in period 1 and 85 per cent in period 2 of the total cost of that basket in the base period. Each index is less than 100, indicating that the price of these commodities, weighted by the quantities con-

sumed in the base period, has gone down. The change in the index is due solely to a change in price because the quantity weights remain the same. Only the prices change. In general, this index is expressed by

$$\frac{\sum p_{i1}q_{i0}}{\sum p_{i0}q_{i0}} \quad (14.2)$$

This index is based on the assumption that the fixed basket of goods of the base period is purchased in each succeeding period. Any shift in the actual consumption of the individual makes this method of weighting inaccurate, relative to what the individual actually buys. The index does not purport to measure changes in the cost of what he buys. It is strictly a price index and a peculiar kind of price index—one which compares only the outlays for the same typical market basket of goods or budget of commodities in each of several periods. If the base period becomes atypical of the actual consumption pattern of the individual, the index may be way out of date and relatively useless. This is a severe limitation on the index, and it must be revised from time to time as important shifts in the typical market basket take place if it is to be kept current and relevant.

An alternative to this index number is to use the quantities consumed in the various periods as weights. The quantities actually consumed represent what is in fact purchased and they are not out of date as may be a typical market basket of some base period. The individual decides to compare the cost of *today's* market basket with *its* cost at base period prices. The cost of the given year basket at both the given year and base periods prices are

Period 1 basket at period 1 prices: $15(3) + 20(4) + 12(5) = 185$
 Period 1 basket at period 0 prices: $10(3) + 15(4) + 20(5) = 190$
 Period 2 basket at period 2 prices: $15(1) + 10(3) + 15(4) = 105$
 Period 2 basket at period 0 prices: $10(1) + 15(3) + 20(4) = 135$

The index numbers written as percentages of the base period 0 costs are

Period 0: 100
 Period 1: $185/190 = 97$
 Period 2: $105/135 = 78$

Thus, the cost of those commodities purchased in period 1 is 97 per cent of what the cost of the same commodities would have been at period 0 prices. The cost of those commodities purchased in period 2 is 78 per cent of their cost at period 0 prices. In general, this index is expressed by

$$\frac{\sum p_{ij}q_{ij}}{\sum p_{i0}q_{ij}} \quad (14.3)$$

Between the base period and any one given time period, the index may change only with a change in price because the given year quantities are fixed. Thus, prices went down in period 1 relative to the base period, and in period 2 relative to the base period. Did prices also go down between period 1 and 2? A change between these two periods may be because of a change in price, a change in quantity, or both. The market basket continues to change each time period and makes an ambiguous measure for comparisons of price changes other than with the base period. The index does have the advantage of keeping the market basket used as the weighting system up to date with actual consumer buying habits. It has a practical disadvantage in that the consumer must keep a record of what quantities he actually buys in each time period. If an index number is to be computed for a large group of consumers, the physical aspects of making continuing budget studies is prohibitive both with respect to the time element involved and the cost of making the study. For this reason, if an index is desired which will measure *only price changes*, the index with the *fixed market basket* of some base period is customarily used. Only one budget study need be made and the quantities determined may be used for several years until the market basket becomes atypical, when a new budget study is made and the whole index revised to a new base period.

Regardless of which index is chosen, however, real problems are connected with its use. This is true even in the simple situation of one consumer's buying a limited number of commodities. What prices and what quantities shall be used in the index? Peas, for example, have no one price in the market. The price depends upon the brand, size of pea, size of can, quality of vegetable, the retail outlet, and special sales. To be comparable over time, the commodity must have definite specifications with only one price so that it will represent a typical purchase in the budget. Also, the quantity weights of the index must represent typical purchases for either the base period (q_{i0}) or the given period (q_{i1}). The quantities to be used are determined by a budget study of the individual, a process which is complicated by large orders, special sales, and purchases for inventory stocking. Changes and substitutions in the individual consumption pattern blur the comparability of the data. Thus, great care must be exercised in defining the commodity, its price, and the quantity purchased.

EXERCISE

14.2. Suppose that the data in the table below represent your food purchases in two stores at three different time periods. Compute an index of changes in the price of each food group and for all foods combined. You may use any one of the three indexes defined by Eqs. (14.1), (14.2), or (14.3). You must state why you

used this particular index, what it measures, and the conclusions you draw from changes in the index.

INDIVIDUAL FOOD PURCHASES IN TWO STORES DURING THREE TIME PERIODS*

Food	Period 1				Period 2				Period 3†			
	Store											
	A		B		A		B		A		B	
	p	q	p	q	p	q	p	q	p	q	p	q
Meat:												
Steak	90	2	95	1	95	3	100	1	90	1	110	4
Roast	80	3	85	3	85	4	90	2	80	1	110	5
Pork chops	75	3	80	4	80	5	85	2	75	2	100	5
Bread spread:												
Butter	60	3	60	1	65	1	65	1	60	1	60	5
Oleomargarine	20	2	20	1	25	1	25	1	20	1	20	3
Jam and jelly	10	2	10	1	15	1	15	1	10	1	10	4
Vegetables:												
Peas	15	2	15	1	20	2	20	1	15	1	25	3
Beets	10	1	10	1	15	1	15	1	10	1	20	3
Corn	10	2	10	1	15	2	15	1	10	1	20	2
Beans	15	1	15	1	20	1	20	1	15	1	25	4
Miscellaneous:												
Bread	20	3	20	1	20	3	20	1	20	3	20	1
Milk	21	6	21	1	21	6	21	1	21	6	21	1

* All prices are in cents, all quantities in pounds.

† Price controls go into effect: store A adheres to price controls and has limited quantities; store B violates controls and has almost unlimited quantities.

14.3 A Simple Production Index

A manufacturer of bicycles produces 25, 15, and 50 bicycles in three successive time periods. A simple relative of these units of production, using the first period as a base, gives indexes of 100, 60, and 200. These data are given in the table below.

Production period	0	1	2
Bicycles	25	15	50
Relative	100	60	200

The production of bicycles by this manufacturer is 60 per cent in period 1 and 200 per cent in period 2 of the production in the base period. The actual number of bicycles produced in each period is not known unless the base period units are known.

An extension of this problem is to several manufacturers producing several products. For example, suppose that three manufacturers produce bicycles, refrigerators, and automobiles in the following quantities.

Production period	0	1	2
Bicycles	25	100	10
Refrigerators	18	18	15
Automobiles	3	1	10

An index of the quantity of production cannot be made by simply adding up the total units because they represent different products. What does the total of 25 bicycles, 18 refrigerators, and 3 automobiles—46 physical units—mean? These production data could be expressed in a common unit of measure, say, total weight in tons. Thus, the 25 bicycles might weigh 750 pounds, the 18 refrigerators 3600 pounds, and the 3 automobiles 6000 pounds, a total of 10,350 pounds, or a little more than 5 tons. Such a summation of common physical units is not too satisfactory, except in special cases, where the ton (or some other measure of weight) is a common mode of expression, e.g., in steel production or shipping volume. The basic problem is one of combining diverse measures of production in an index number. In combining the measures of changing product, some gauge of the relative importance of the various products must be chosen. Although it is the avowed purpose of an index of production to measure changes in physical quantities of production, the gauge used in most indexes is in value terms, e.g., the price of the product or the value added to the raw materials in its manufacture.

TABLE 14.5. PRODUCTION DATA FOR THREE MANUFACTURERS

Product $i =$	Time period					
	$j = 0$		$j = 1$		$j = 2$	
	q	p	q	p	q	p
Bicycles 1	25	\$ 30	100	\$ 10	1	\$ 60
Refrigerators 2	18	200	18	200	1	400
Automobiles 3	3	2000	1	2000	10	2000

Table 14.5 gives the prices and quantities produced of each of three commodities in three time periods. Thus, p_{ij} and q_{ij} are the prices and quantities of the i th commodity in the j th time period. The base period

is $j = 0$. For these three consumer durable goods, an index of production is to be defined.

One method of constructing an index is to weight each quantity produced by its price, compute the total value of the three products in each time period, and express the total value in periods 1 and 2 as relatives of the total value in period 0. The total values for the periods are

$$\text{Period 0: } 25(30) + 18(200) + 3(2000) = 10,350$$

$$\text{Period 1: } 100(10) + 18(200) + 1(2000) = 6,600$$

$$\text{Period 2: } 1(60) + 1(400) + 10(2000) = 20,460$$

With period 0 as the base period, the indexes are

Period	0	1	2
Index	100	47	198

The total value of products made is 47 per cent in period 1 and 198 per cent in period 2 of the total value in the base period. Is the change in the index from 100 to 47 an indication that total physical production has declined between the two periods? Part of the drop in the index is occasioned by the drop in the price of bicycles; to this extent, then, the index is ambiguous as a measure of quantity change. Even without the price change, does the index indicate a drop in total physical production? The refrigerator manufacturer maintains his level of production, and the bicycle manufacturer quadruples his production. The dominant product in the index is the automobile producer whose production is cut by two-thirds.

From period 0 to period 2, the index increases 98 per cent. Is this an indication of an increase in total physical production? Both the refrigerator and bicycle manufacturers are practically out of business although they are selling the one unit they produce for twice the base period price. The automobile manufacturer is enjoying a great boom in production with 10 cars. The index increases partly because of the price increase, but mostly because of the large change in automobile production. Does this index measure the total physical output? Has production increased almost twofold since the base period?

In general terms, this index number is expressed by the relationship

$$\frac{\sum q_{ij}p_{i0}}{\sum q_{i0}p_{i0}} \quad (14.4)$$

Its greatest disadvantage is that a change in the index is brought on by a change in price as well as a change in quantity, a behavior not desired in an index of production.

Another method of constructing an index number is to use only the base period prices as the weights, following the formula

$$\frac{\sum q_{ij}p_{i0}}{\sum q_{i0}p_{i0}} \quad (14.5)$$

The total values of the production in each period at base period prices are

$$\text{Period 0: } 25(30) + 18(200) + 3(2000) = 10,350$$

$$\text{Period 1: } 100(30) + 18(200) + 1(2000) = 8,600$$

$$\text{Period 2: } 1(30) + 1(200) + 10(2000) = 20,230$$

The indexes with period 0 as a base period are

Period	0	1	2
Index	100	83	195

The total value of the production at base period prices is 83 per cent in period 1 and 195 per cent in period 2 of the total value of production in the base period. The relative valuations of the individual products in the market place—their prices—remain fixed in this index number. As a result the index may be inaccurate if substantial price changes take place. Again, the question is asked: Does a change in this index reflect a change in total physical production? Is total production almost doubled between the base period and period 2, or does it decline approximately 17 per cent (index = 83) in period 1?

An index can be defined with given period prices as the weighting system according to the formula

$$\frac{\sum q_{ij}p_{ij}}{\sum q_{i0}p_{ij}} \quad (14.6)$$

This index compares the total value of the production in a given period with the value of the base period production at the given year prices. In practice, this method is seldom used because of the very large costs involved in obtaining the information on prices each period. The index with fixed weights of the base period is the most common in measuring changes in production (as it is in measuring price changes).

When an index of production is extended to many different products, some of which may be the raw materials of others, the use of the price of the product as a weight for the quantity of production is not a good one because the same physical goods may be counted several times. For example, when a sink is made in a foundry a piece of iron is taken and a sink is cast from the molten metal. The iron has already been counted in the production of the iron works; it cannot be counted again in the

foundry. All that may be counted as the production of the foundry, to avoid an overlap with the raw materials, is the *value added* to the product by virtue of making it into a sink from a chunk of iron. In a large-scale production index, the weights are a measure of the value added to the product by the manufacturer—a measure of net output—rather than the price of the product. The methods of using value added weights are the same as with the prices. The index numbers have the same limitations, except that now they are indexes of value added to production. If the index of production employs the common fixed weights of some base period, the index compares the value of the net production at constant prices in a given year and the base year.

The data of Table 14.6 are changed to the quantities (q_j) produced and the value added (p'_j) to each product in its manufacture. An index

TABLE 14.6

Product $i =$	Time period					
	$j = 0$		$j = 1$		$j = 2$	
	q	p'	q	p'	q	p'
Bicycles 1	25	5	100	1	1	10
Refrigerators 2	18	25	18	25	1	50
Automobiles 3	3	500	1	500	10	500

with the fixed weights of value added in the base period ($j = 0$) is constructed according to Eq. (14.5).

$$\frac{\sum q_{i1} p'_{i0}}{\sum q_{i0} p'_{i0}}$$

The indexes for the three time periods are

Production period.	0	1	2
Index	100	70	242

Each index is a relative of the base period total output weighted by the value added, a total dollar amount. A change in the index is induced only by a change in output, because the p'' 's remain constant. Somehow the feeling is that if this index goes up, total production has increased in the segment of the economy for which the index is computed. Since an index number is only an average, individual manufacturers may have decreased total production although the index shows an increase. Changes in the index reflect some kind of change in production and prob-

ably bear some relation to changes in total physical production, although basically the indexes are relatives of values, and not physical units of production.

EXERCISE

14.3. Data on quantities produced (q), prices (p), and value added to production (p') are given in the table below for four products manufactured in a given locality. Construct an index of production, using any method you wish, for this locality. Defend your choice of index number by stating what it measures, why you used it, what are its limitations, and how changes in it are interpreted.

Product* $i =$	Time period								
	$j = 0$			$j = 1$			$j = 2$		
	q	p	p'	q	p	p'	q	p	p'
Cast iron 1	10	\$20	\$5	20	\$20	\$5	30	\$20	\$6
Canned peas 2	100	3	1	50	3	1	25	3	1
Davenport 3	11	100	50	10	100	50	15	75	40
Sinks 4	50	50	30	100	50	30	150	50	35

* Units of measure: cast iron in tons and canned peas in cases. All prices in dollars per unit.

14.4 A Description of Two Index Numbers

These introductory sections have attempted to explain with simple examples the basic elements of index numbers: the notion of a relative, the problem of weighting, and the practical problems of definition and interpretation which arise in even the simplest index-number formulations. When the price index is extended to a group of individuals buying several commodities or the production index to many firms making several products, these problems are magnified and new ones arise. For example, different consumers have different budgets and make different purchases. They live in different areas and shop in different stores. An item important in the budget of one may be unimportant in the budget of another. Of all the possibilities for a price and quantity situation, the one selected must represent in some sense an average or typical budget for a certain broad class of individuals.

A host of indexes of prices and production exists both in the United States and other countries of the world; there are so many, in fact, that it is practically impossible even to list them. As one quip says, almost every respectable agency and certainly every respectable country own at least one index number. The phenomenon is so widespread that every individual, as part of his general education, ought to be more aware of it,

although it is considered a moot point by some whether he would be any the worse off if most of the indexes were suddenly discarded. Since the plethora of index numbers makes a survey of many of them impractical, only the two most widely known indexes in the United States will be discussed in any detail: the Consumer Price Index and the Federal Reserve Board Index of Industrial Production.

14.5 The Consumer Price Index

The full name of this index is Index of Change in Prices of Goods and Services Purchased by City Wage-earners and Clerical-worker Families to Maintain Their Level of Living, although it is commonly and correctly called the Consumer Price Index and often incorrectly the Cost of Living Index. The index is compiled monthly by the U.S. Bureau of Labor Statistics and is perhaps the best known index in the country. The discussion of the index is taken from several government publications.¹

The Consumer Price Index is a measure of changes in prices of goods and services bought by families of city wage earners and clerical workers. It measures changes in prices, not changes in the kinds of goods families buy, the amounts that they buy, or the total amount that they spend. Neither does it compare differences in living costs in different places.

The numbers and kinds of commodities included in the index must be in keeping with the purpose of the index. The commodities must be goods and services bought by city wage earners and clerical workers. These include food, clothing, household goods, doctor fees, rents, transportation charges, and so on. An urban family is defined as consisting of two or more persons living in the 3000 towns, cities, and suburbs of the United States, ranging in population from small cities of 2500 to New York City. The heads of the families are wage earners or salaried clerical workers. The average size of the families included in the present index is about 3.3 persons and the average family income is about \$4160 after taxes. These families represent about 64 per cent of all people living in urban places and about 40 per cent of the total population of the United States. The goods and services represented are not for the families with the lowest and highest extremes of the income scale. They are not for professional workers or farmers or single people.

The 18 million urban wage earners and clerical workers in the 3000 cities and towns in the United States buy thousands of different kinds and qualities of goods and services in many thousands of stores and service establishments. It is impossible to find out what every family buys and, even if this were possible, to include in an index every good or service bought. Consequently, a sample of these goods and services is

¹ See *The Consumer Price Index, a Short Description of the Index as Revised, 1953*, Bureau of Labor Statistics, Washington, D C., 1953.

selected for the index. Since the prices of many groups change at about the same rate, only one of a few commodities in the group need be included in the sample to represent the whole group. The principle of stratification by price movements is used. For example, the price movement of fresh milk can be used to typify the change in the prices of buttermilk and various other milk products. The price changes in round steak, chuck roast, rib roast, and hamburger may be used to represent all beef products. Based upon an extensive study of price changes in hundreds of items, the Bureau selected about 300 items which taken together represent all the thousands of goods and services bought by families. These items typify the price movements of larger groups of commodities, are important in the family budget, and are carefully specified so that the same quality of goods will be priced each time. Changes in quality upset the index because part of a price change may represent a change in the basic good being purchased.

Every index number is weighted, either deliberately or haphazardly. The weights used depend upon the purpose of the index number. The purpose of the Consumer Price Index is to measure changes in prices alone. Immediately, then, a value index with a mixture of changing prices and quantities is eliminated. The choice lies between an index with changing quantities each period or with fixed quantities of a base period. A fixed weighting system allows the weights to get out of date and become unrealistic as new commodities come into existence or as consumers change their buying habits. A changing system of weights reflects a change in both the standard of living and in prices. Both systems are wrong to a certain extent and a choice must always be made between inaccuracy and ambiguity.

The Consumer Price Index is generally of the type with a system of fixed weights. The Bureau made this choice in the beginning not because it desired to have an inaccurate index, but because it wished to measure only price changes and because the cost of the physical aspects of a system of changing weights is staggering. The weights are not absolutely fixed, because substitutions are made from time to time as some commodities drop out of use and others take their place. The possible inaccuracy is further reduced by changing the whole weighting system every few years.

The quantity data which are used as weights in the Consumer Price Index represent the relative importance of the various items in the consumer's budget; i.e., in his market basket of goods and services. The market basket is determined by a survey of consumer expenditures in a sample of cities. For example, in 1950, this sample included all 12 of the largest urban areas with populations of over 1,000,000, plus a sample of large, medium-sized, and small cities. The cities for the sample were

selected in such a way as to represent different characteristics of size, climate, density of population, and level of income. In all, 97 cities were selected to represent all types of cities in the United States.

In each city, a sample of families was interviewed to obtain a complete record of the kinds, qualities, and amounts of food, clothing, furniture, and all other goods and services purchased, and to ascertain how much was spent for each item. The records for families in the sample were averaged for each city. These data provide the fixed market basket quantities.

The prices collected by the Bureau are retail prices—prices to the consumer, including sales and excise taxes. From the 97 cities covered by the expenditures survey to determine the quantity weights, the Bureau samples prices in 46 cities: the 12 largest, 9 other large cities, 9 medium-sized cities, and 16 small cities. In each city, the Bureau collects prices from a sample of stores representative of chain stores, independents, department stores, and specialty stores where families of wage and salary workers buy their goods and services. Prices reported by these stores are averaged together for each city to determine the average price changes. Periodic checks are made to determine whether the sample of stores is still typical for that city.

The prices are collected at intervals ranging from 1 month to once every 4 months. Food prices change rapidly and are priced in each city each month. Prices of some goods and services are obtained every month in the five largest cities. In other large cities and medium-sized cities, prices of items other than foods, fuels, and rent are obtained every 3 months. In small cities, they are collected every 4 months. Food prices are collected around the middle of the month by local part-time agents trained to do this work. Other prices are obtained by full-time Bureau agents. Rent information is collected by mail every month from each city. Streetcar fares, public utility rates, newspaper prices, and a few others are also obtained by mail. Most important, the price obtained is for a well specified and identifiable commodity.

The Bureau of Labor Statistics computed its first Consumer Price Index on a survey of consumer expenditures in 1918. Because goods and services available and people's buying habits changed, a new study and a new market basket of goods was made for the years 1934 to 1936. During World War II, when goods were scarce and many were rationed, interim changes were made in the weighting system until a complete post-war survey of consumer expenditures could be made. Then, in 1950, the Bureau estimated consumer purchases again and revised its index to a 1952 base with the first index coming out for January, 1953. The 1952 market basket brings the index up to date and includes such things as TV sets and frozen foods which were not important in earlier family

market baskets. Although the market basket was determined by a survey in 1950 and the data about consumer expenditures were adjusted to 1952, the statistical base (100) of the present index is the average for the years 1947 to 1949. This period is used to facilitate comparison with other series also based upon $1947-1949 = 100$. Each monthly index is a relative expressed as a percentage of 1947-1949. In equation form,

$$\text{C.P.I.} = \frac{\sum p_{x7-49}}{\sum p_{47-49} / 47-49} \quad (14.7)$$

An index is made for each individual group of commodities (for example, food and housing) and subgroup of commodities (e.g., cereals, meats, dairy products). Expenditures weights for groups of items are obtained for each city and a combined index made for some of the larger cities. Finally, with the population of individual cities used as weights, an index is made to represent price changes for all urban wage-earner and clerical-worker families in the United States. The population weights are adjusted with new Census data as they become available.

Limitations of the Consumer Price Index. Several limitations of the Consumer Price Index are implicit in the preceding discussion.

1. Sampling imposes one kind of limitation on the accuracy of the index. The expenditures in the base period market basket were based upon 8000 interviews. The index itself is calculated from the price changes of only 300 items priced in only 46 cities and only at certain times. Because reports are not available for every person in every city at all times, sampling errors are present in the final index number. The Bureau makes a definite effort to keep these sampling errors to a minimum, but it must be remembered that the final index is only a sample estimate, not an exact measurement.

2. Another kind of error occurs because people cannot report information accurately. In the survey to determine consumer expenditures for various goods and services, the Bureau found that most families can estimate quite accurately what they spend on rent or on automobiles. They have more difficulty remembering all their expenses for small and more frequently purchased items. These errors may tend to cancel out as some people report more than they spend and some report less. Good survey methods and highly trained interviewers tend to keep these errors to a minimum.

3. Errors in pricing can also occur. The agent must know the specifications of the item to be priced and be able to recognize it. Although they sometimes miss special discounts or premium prices, well-trained agents will in general report correct prices.

4. The index is calculated to measure the average change in prices of a select group of commodities for a select group of people. It represents

all wage earners or clerical workers, but not necessarily any one family. It does not apply to very low or very high income families, to single people, or to farmers. The index does not measure total family spending or reflect the experience of an individual housewife shopping for sales or lower-quality items. Since the index has a fixed system of weights, it does not measure any changes in the kinds of goods or services purchased, and it may get out of date and become inaccurate with the passage of time if substantial changes in consumer spending habits take place.

The index represents about the best that can be done in measuring the price changes of consumer goods and services.

Industrial Production Index. One of the most widely used indicators of industrial production in the United States is the Index of Industrial Production issued by the Federal Reserve Board since the 1920s. The following discussion of the index summarizes the more detailed accounts found in the *Federal Reserve Bulletin*.¹

The Index of Industrial Production measures the physical volume of output in the manufacturing and mining industries which account for about three-fourths of the national income in the United States. The index measures output changes at all stages within manufacturing and mining operations—intermediate as well as final product. The index does not concern itself with production in the farm, construction, utility, transportation, trade, or service industries. Changes in the index reflect broad changes in the total output of the economy and are useful in a general analysis of business conditions and business fluctuations.

Thousands upon thousands of different commodities are produced in the United States each year. Obviously, not all of these can be included in an index of total production. In keeping with the purpose of the Index, only manufacturing and mining commodities are included. All others are eliminated by definition. All manufacturing and mining commodities are grouped together into rather broad classifications in order that an index may be computed for each group and production measured for a few products as representative of the whole group. The classification into groups could follow various schemes: major materials used, markets for products, types of processes or raw materials, stages of fabrication, or some combination of a few or all of these. For industrial production, the only feasible classification principle depends upon the data which are available. A purist position on any classification principle is impossible.

The Federal Reserve Board has adopted industry classifications according to the Standard Industrial Classification Manuals issued by the

¹ For a discussion of the latest revision, see Industrial Production Index Revised, *Federal Reserve Bulletin*, vol. 39, no. 12, pp. 1239–1328, December, 1953.

Bureau of the Budget. The use of this scheme provides a very definite classification principle with industry groups which are comparable to other series, thus facilitating comparison of the index with other series. The SIC lists 26 major industry groups—21 in manufacturing and 5 in mining. The total manufacturing group is broken into 2 major subgroups: durable and nondurable manufactures. Within durable manufactures are subgroups like primary metals; metal fabricating; clay, glass, and lumber products; furniture and miscellaneous manufactures. Within nondurables are subgroups like textiles and apparel; rubber and leather products; paper and printing; chemical and petroleum products; foods, beverages, and tobacco. The mining industry is broken into groups of mineral fuels, including coal (anthracite and bituminous) and crude oil and natural gas, and metal, stone, and earth minerals, including metal mining and stone and earth minerals. The index is based upon a total of 175 monthly series which represent all these various manufacturing and mining industries.

The Index of Industrial Production uses the *value added* by each manufacturing process as weights for the physical output. Value added avoids a duplication of production by successive fabricators and measures the relative importance of the production based upon the value added to the purchased raw materials in the process of fabrication. It reflects the relative valuations placed upon the various production activities in the market place. In the manufacturing industries, the value added is determined by subtracting the cost of materials, supplies, containers, fuels, etc., from each industry's gross value of products.

The value added data used as weights in the production index are obtained from the 1947 Census of Manufactures for manufacturing industries and from 1947 estimates prepared by the Bureau of Labor Statistics for the mining industry. In the present index, total manufacturing is given a weight of 90.02 and total mining a weight of 9.98. Within the total manufacturing group, the durables have a weight of 45.71 and nondurables of 44.85. The weights assigned to other classifications on the basis of the 1947 estimates are readily obtainable from the previous reference.

The production figures used in the index are obtained from regularly published data of other government agencies and private organizations. These data are usually gathered for another purpose on a monthly or yearly basis and are taken over by the FRB for use in the index. Government agencies supply about 83 per cent of the annual index data. Problems in gathering the data arise from the fact that they are frequently fragmentary or often wholly lacking. The compilers of the index have the task of evaluating the available data, selecting a method of representation, and adjusting the data to increase their adequacy. The pre-

ferred data are physical units of product—tons, yards, etc. In practice, physical data are lacking and substitute methods of measurement are used. These take several forms:

1. Physical quantity shipped less stocks of goods
2. Quantities of materials consumed in production
3. Value of goods shipped adjusted for price changes
4. Production-worker man-hours adjusted for output per man-hour
5. Indirect representation by other data with similar movements

Indirect representation accounts for from 3 to 7 per cent of the index. The man-hour series accounts for about 4 per cent in the annual data and 45 per cent in the monthly data. Thus, about half of the monthly index to measure physical output is not based upon physical output but upon man-hours of work and output per man-hour.

All data are adjusted for calendar variations. The data available cover different reporting periods and are adjusted to a daily average basis. Some industries normally work 7 days and some 6; holidays may reduce some work weeks to 4 or 5 days. The working-day adjustment is only for calendar variations; it does not include days lost due to work stoppage for other reasons.

The period for determining the weights—value added by manufacturing or mining—was 1947. Estimates of the relative importance of each production classification were made in that year. The base period index equal to 100 is the average for the three years 1947 to 1949. The use of this base period facilitates comparisons with other indexes which are calculated on this base. The previous base period was the average index for 1935–1939, which was adjusted to the new 1947–1949 base after World War II for obvious reasons. Each index number, monthly or yearly, is a relative of the average 1947–1949 total output weighted by value added, a total dollar amount.

The practical aspects of taking a census of manufactures or of mining to estimate the value added weights dictate an index with fixed weights of some base period. For computational convenience, the index is expressed by the formula:

$$\text{I.I.P.} = \sum \left(\frac{q_x}{q_{47-49}} \cdot \frac{q_{47-49}p_{47}}{\sum q_{47-49}p_{47}} \right) \quad (14.8)$$

That is to say, the output (q_x) in a given time period x is expressed as a relative of the average output of the 1947–1949 period (q_{47-49}). This relative is weighted by the fraction expressing the ratio of the total dollar figure for that product ($q_{47-49}p_{47}$) to the total dollar figure for all commodities ($\sum q_{47-49}p_{47}$). This symbolic statement is equivalent to the index with fixed base period weights because " q_{47-49} " appears in both the numerator and denominator of the weighted average of relatives. The

choice of the weighted relative expression is merely a convenience for calculation.

The Federal Reserve Board publishes a monthly index of industrial production both adjusted and unadjusted for seasonal variation. The production data are marked by more or less regularly recurring patterns of fluctuation during the course of one year which tend to be repeated in successive years. The technical method of eliminating the seasonal variation is described in the Federal Reserve *Bulletin*, June, 1941, pp. 518-528. The seasonal adjustment for the postwar period has been complicated by many factors, but the 1952 revision of the index included new estimates of the seasonal variation. These, together with the seasonal variation estimates for the old index, are given in the table below.

SEASONAL VARIATION IN THE TOTAL INDEX, 1952

Month	New index	Old index
January	98 9	98 0
February	101 0	98 2
March ..	101 8	99 0
April	100 1	99 5
May .	99 1	100 2
June	100 4	101.0
July	94.2	100 1
August	100 0	102 1
September	101 7	101 7
October	103 0	101 9
November	101 0	100 0
December	98 5	98 8

Limitations on the Index. The Index of Industrial Production is subject to the same general limitations as the Consumer Price Index with respect to sampling errors and errors in reporting information. In addition, no satisfactory method of taking quality changes into account has as yet been found. Since the Index is of the type with fixed period weights, it may be inaccurate if the actual relative importance of the different manufacturing and mining groups changes drastically from the base period weighting system. The Index attempts to measure only the average change as defined by the technical index formula. It does not tell what has happened to the production of one plant or of one product. It is not concerned with changes in farm, construction, utility, or various other productions. With the refinements made in the new Index, the Federal Reserve Board believes that it has a much improved tool for analysis and a more representative summary of industry's output for those interested in general business developments.

14.6 Economic Theory of Index Numbers

All the material on index numbers in this chapter is descriptive. Price and quantity indexes are described with emphasis placed upon methods of constructing actual index numbers and not upon a theoretical discussion. An economic theory of index numbers has been discussed by some writers, both with respect to price and production indexes. The interested reader may consult the references at the end of the chapter for accounts of the discussion.

14.7 Sources of Index Numbers

So many index numbers are in use that a listing of them is impossible. A good reference which locates and briefly describes 449 different series is Arthur H. Cole, "Measures of Business Change," Richard D. Irwin, Inc., Homewood, Ill., 1952. Other references to actual indexes are given at the end of the chapter.

EXERCISES

14.4. Answer each of the questions posed at the beginning of this chapter in the light of the general discussion of index numbers and the fairly detailed discussion of the Consumer Price Index.

14.5. How does the BLS Index of Consumer Prices apply to you as an individual? Do movements in the index from month to month indicate what is happening to your own cost of living?

14.6. In May, 1948, the United Automobile Workers and the General Motors Corporation negotiated a contract in which the BLS Consumer Price Index was used as a basis of wage-rate escalation. This agreement provided for a quarterly adjustment in wage rates of 1 cent per hour for every change of 1.14 points in the Index on the 1935-1939 = 100 base. With the revision of the Index to a 1947-1949 base, the original agreement was revised in 1953 to allow for a quarterly adjustment in wage rates of 1 cent per hour for every 0.6 points in the revised index.

a. If you were an employee of General Motors, would you like to have your wage rate adjusted according to this procedure?

b. What do you think of the principle that wage rates should be adjusted for cost of living by tying them to changes in the Consumer Price Index from the statistical point of view?

14.7. The Consumer Price Index for *total foods* stood at 113.1 for Los Angeles, 111.0 for New York, and 110.4 for Chicago in April, 1954. Does this mean that food is cheaper in Chicago than in either of the other two cities? Exactly what does it mean?

14.8. The food index stood at 113.1 and the transportation index at 127.3 for Los Angeles in April, 1954. Does this mean that the average family covered by the index paid more for transportation than for food in April in Los Angeles?

14.9. The Dow-Jones Industrial Average is computed by dividing the total prices of 30 industrial stocks by a constant. This constant changes from time to time to allow for stock substitutions and stock split-ups which change the importance of an individual issue in the index. In August, 1954, the divisor was 5.89.

If each of the 30 stocks used in the index increases \$1 in price, what is the increase in the index? If each stock decreases \$1, what is the decrease in the index? What are the implications of the way in which the index is computed for analyzing stock market price fluctuations?

14.10. In connection with the Little Steel Formula for granting wage increases in 1943, the AFL and CIO attacked the Cost of Living Index because they believed it understated the rise in the cost of living from 1941 to 1943 by from 20 to 40 points. The reasons given for the understatement were:

1. Quality deterioration of goods purchased
2. Underreporting of prices of goods and services actually bought because of ceiling prices and the OPA
3. Forced uptrading by the lack of cheaper goods which were included in the index but were not available for purchase.
4. Elimination of special sales at which large amounts of goods were previously purchased, especially groceries.

Comment on this attack on the index by evaluating each of the four reasons given in terms of the construction of the BLS Consumer Price Index.

SUPPLEMENTARY READING

Moroney, M. J.: "Facts from Figures," pp. 48-54, Penguin Books, Inc., Baltimore, 1953.

REFERENCES

- Allen, R. G. D.: The Economic Theory of Index Numbers, *Economica*, August, 1949, pp. 197-203.
- Brambrick, James J., Jr., and Harold Stieglitz: "Cost of Living Provisions in Union Contracts," National Industrial Conference Board, Studies in Personnel Policy, No. 113, New York, 1951.
- Joint Committee on the Economic Report: "Historical and Descriptive Supplement to Economic Indicators," U.S. Government Printing Office, Washington, 1953.
- : "Economic Indicators," U.S. Government Printing Office, Washington. (Monthly.)
- Mitchell, Wesley C.: "The Making and Use of Index Numbers," U.S. Bureau of Labor Statistics, *Bulletin* 656, Washington, 1938.
- Mudgett, Bruce D.: "Index Numbers," John Wiley & Sons, Inc., New York, 1951.
- U.S. Bureau of Labor Statistics, Techniques of Preparing Major BLS Statistical Series, *Bulletin* 993, U.S. Government Printing Office, Washington, 1950.

CHAPTER 15

MISCELLANEOUS TOPICS

This book has not covered the whole field of statistics. Many problems exist for which statistics can provide solutions beyond what has already been taught in this course, and this chapter will present typical examples of several of them. The solution to each is given with little discussion of either the principles behind the methods or their widespread application. The intention is only to give the reader enough familiarity with more advanced statistical ideas so that if he meets situations for which answers have not been provided in this book, he may think to inquire whether such answers do exist and to seek them out either from other books or from a statistician.

15.1 Two Sample Proportions

A newspaper recently studied the effectiveness of newspaper advertising in selling real estate. As a part of this study it found that out of a random sample of 500 homes that were sold by brokers, 425 or 85 per cent of the buyers were first brought in contact with the broker by an advertisement in a newspaper. At the same time, it found that out of a random sample of 500 homes which were sold by owners without the services of a broker, 450, or 90 per cent, of the purchasers established connections with the seller directly through a newspaper advertisement. Is it reasonable to believe that the percentage of owners who established a contact through newspaper advertisements is higher than the percentage of brokers who made a contact through newspapers? The sample percentages are higher, but the percentages of two samples from the same population will differ. The question to be answered is: Did these two random samples come from populations with the same or different values of P , the percentage of contacts made through newspapers?

Let P_1 be the population proportion of owners who established a contact through newspaper advertisements. Let P_2 be the population proportion of brokers who made a contact through newspaper advertisements. The basic hypothesis to be tested is: Does $P_1 = P_2$? The alternative to this hypothesis is: Is $P_1 > P_2$? That is to say, is the

percentage of owners who established a contact through a newspaper advertisement greater than the percentage of brokers who established a contact in this manner? The data are set out in a fourfold table.

OBSERVED NUMBER OF SALES

Seller	Newspaper contact		Total
	Yes	No	
Broker	425	75	500
Owner	450	50	500
Total	875	125	1000

This table is read as follows: 425 purchases were made from brokers with whom the contact was made through a newspaper; 75 purchases were made from brokers with whom contact was not made through a newspaper; 450 purchases were made directly from owners with whom a contact was established through a newspaper; 50 were made from owners with whom a contact was not made through a newspaper.

The true percentage of contacts made through a newspaper is unknown, and no theory furnishes such a percentage. Among the 1000 home purchases, 875, or 87.5 per cent, were made by contacts through a newspaper. If the basic hypothesis to be tested, that $P_1 = P_2$, is true, the percentage of contacts made through newspapers should be the same whether the seller is a broker or an individual. The fraction of sellers who will establish contacts through a newspaper is therefore taken as .875, on the general hypothesis that $P_1 = P_2$. The expected number of sales through brokers with contacts through a newspaper is $Ex = .875(500)$, or 437.5. The expected number of sales through brokers with whom a contact was not made through a newspaper is $500 - 437.5$, or 62.5. The expected number of sales in each of the four categories is given in a table similar to the observed numbers.

EXPECTED NUMBER OF SALES

Seller	Newspaper contact		Total
	Yes	No	
Broker	437.5	62.5	500
Owner	437.5	62.5	500
Total	875	125	1000

A measure of the discrepancy between the observed and expected numbers is χ^2 discussed previously in Chap. 8. Using the former notation,

let O be the observed number and E be the expected number in any category. Then,

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (15.1)$$

From the data in the two tables,

$$\chi^2 = \frac{(450 - 437.5)^2}{437.5} + \frac{(425 - 437.5)^2}{437.5} + \frac{(75 - 62.5)^2}{62.5} + \frac{(50 - 62.5)^2}{62.5} = 5.7$$

The χ^2 distribution has one parameter: the number of degrees of freedom. This fourfold table has only *one* degree of freedom, because only one of the four entries in the table may be filled arbitrarily keeping the same totals. The other three follow automatically from this one value by subtracting from the totals. The magnitude of χ^2 is judged according to the χ^2 probability distribution with one degree of freedom.

All these ideas may be set out in the form which has been used throughout this book.

$$\begin{array}{ll} H_0: P_1 = P_2 & \text{D.F.} = 1 \\ H_1: P_1 > P_2 & \alpha = .05 \end{array}$$

For 1 D.F., the value of χ^2 exceeded with probability .05 is 3.841 in Table A.9.

Decision Rule. Accept H_1 if $\chi^2 \geq 3.841$; otherwise accept H_0 .

Decision. Accept H_1 because $\chi^2 = 5.7$ falls in the region R . The conclusion is that the populations which gave rise to the samples have different values of P . The proportion of contacts made from a newspaper advertisement with owners to sell their homes is greater than the proportion of such contacts with brokers.

15.2 Several Sample Proportions

A survey to determine the characteristics of the buyers of homes financed with a Veterans' Administration Guaranteed Loan (commonly called a G.I. Loan) gave the following information on the ages of the buyers, classified according to veteran or nonveteran status.

Buyer	Age group				Total
	Under 25	25-34	35-44	Over 45	
Veteran	10	97	37	32	176
Nonveteran	14	37	97	64	212
Total	24	134	134	96	388

The data are set out in a 2-by-4 contingency table. The question to be answered is: Are the age distributions the same for the veteran and non-veteran buyers? Is the age distribution independent of the veteran classification or does it depend upon whether the buyer is a veteran? In statistical terms, this question is stated: Is the percentage of veteran and nonveteran buyers the same within each age bracket? The percentage may differ from one age bracket to the next, but within an age bracket is it the same for each category of buyer?

The percentage in each of the age groups is unknown but may be estimated from the sample. The percentage of buyers under 25 is $24/388$. Since 176 buyers are veterans, the expected number of veteran buyers under 25 is $(24/388)(176) = 10.9$. It follows that the expected number of nonveteran buyers under 25 is $24 - 10.9 = 13.1$. The percentage of buyers over 45 is $96/388$, and the expected number of veteran buyers over 45 is $(96/388)(176) = 43.5$. It follows that the expected number of nonveteran buyers over 45 is $96 - 43.5 = 52.5$. The other expected numbers are computed in similar manner. From the observed and expected numbers,

$$\chi^2 = \frac{(10 - 10.9)^2}{10.9} + \frac{(14 - 13.1)^2}{13.1} + \dots + \frac{(32 - 43.5)^2}{43.5} + \frac{(64 - 52.5)^2}{52.5} = 61.9$$

How many degrees of freedom does this sample have? Although 8 expected numbers were computed, the requirement that they add to the same totals as the observed numbers for the same columns and rows permits only 3 of the expected numbers to be filled in arbitrarily. The other 5 follow from the marginal totals. This sample value of χ^2 has 3 degrees of freedom and must be judged according to the χ^2 distribution with 3 D.F.

The formal presentation of the statistical decision follows.

H_0 : Percentages within each age group are the same for veterans and nonveterans.

H_1 : Percentages within each age group differ for veterans and nonveterans.

D.F. = 3 $\alpha = .05$ $\chi^2_{.05}(3 \text{ D.F.}) = 7.8$ (from Table A.9)

Decision Rule. Accept H_1 if $\chi^2 \geq 7.8$.

Decision. Accept H_1 and conclude that the age distributions differ from each other. They are not independent of the veteran or nonveteran classification. In general, the nonveteran purchasers tend to be older than the veterans.

15.3 Two Sample Means

A buyer of a product for which the breaking strength is a critical factor is supplied with a sample of 10 items by each of two competitors. The product of manufacturer *A* is more costly than that of *B*, but it is advertised as being the best product of its kind on the market. Each of the samples is tested with an appropriate breaking-strength test resulting in the following:

Sample <i>A</i> :	66	82	71	78	69	73	75	61	90	75	$\Sigma x = 740$	$\bar{X} = 74$
Sample <i>B</i> :	55	68	43	15	52	58	56	61	54	48	$\Sigma x = 540$	$\bar{X} = 54$

The average breaking strength of sample *A* is 74 and of sample *B* is 54. Is this an indication that product *A* should be purchased?

Just because the sample means are different is no reason to conclude immediately that product *A* has a greater breaking strength. Two samples from the same population will have different means because of the sampling variation inherent in the process of random selection. Only when the difference in sample means is so large as to cast doubt on the hypothesis that the two samples came from the same population is it valid to infer that the populations from which the samples came have different means. Different methods of analysis are possible for this problem. The one given here is used because it leads directly to the more general problem of several sample means.

If the two samples really come from the same population, the variation between the sample means must reflect the same population variance as the variation within samples. If it should happen that the "between sample mean" variation is greater than the "within sample variation," it is reasonable to suspect that the two samples do not come from populations with the same mean but from populations with different means, so that to the within population variation is added a between population variation. The basic method of analysis, called *analysis of variance*, is devised to break up the total variation in the 20 different sample products into two sources of variation: a variation within samples about the individual sample means, and a variation between sample means. The data given are simple enough so that the complete structure of computations may be explicitly set out, although in practice substitute computing procedures are used to reduce the amount of labor involved.

Throughout this entire procedure it must be kept in mind that the variance is a sum of squared deviations from a mean divided by the number of degrees of freedom. For one sample, the variance is defined by $s_x^2 = \Sigma(x - \bar{X})^2 / (n - 1)$, a basic formula given before. The computations here are variations on this formula.

Disregarding the source of the products—manufacturer *A* or manufacturer *B*—the grand mean, symbolized by $\bar{X}$, is 64. The total of all 20 observations is 1280, and $\bar{X} = 1280/20 = 64$. The total sum of squared deviations of all 20 observations from this grand mean is 3114, as derived from the table below which gives each observation as a deviation from the grand mean.

DEVIATIONS FROM THE GRAND MEAN: $\bar{X} = 64$									
-9	4	21	-19	-12	-6	-8	-3	-10	-16
2	18	7	14	5	9	11	-3	26	11
Σ (deviations) = 0					Σ (deviations) ² = 3114				

The mean of sample *A* is 74 and the mean of sample *B* is 54. Expressing each original observation as a deviation from its respective mean gives the deviations in the following table.

Sample <i>A</i> (deviations from 74)	-8	8	-3	4	-5	-1	1	-13	16	1
Sample <i>B</i> (deviations from 54)	1	14	-11	-9	-2	4	2	7	0	-6

The sums and sums of squares of these deviations are

$$\begin{aligned}\Sigma (\text{deviations of sample } A) &= 0 & \Sigma (\text{deviations})^2 &= 606 \\ \Sigma (\text{deviations of sample } B) &= 0 & \Sigma (\text{deviations})^2 &= 508\end{aligned}$$

The total sum of squared deviations *within* samples is $508 + 606 = 1114$.

The sum of squares *between* samples is found by substituting the sample mean for each of the original observations—74 for sample *A* and 54 for sample *B*. The deviations of these substituted sample means from the grand mean of 64 are given in the table.

Sample <i>A</i>	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10
Sample <i>B</i>	10	10	10	10	10	10	10	10	10	10
Σ (deviations) = 0					Σ (deviations) ² = 2000					

With the appropriate number of degrees of freedom, each of these three sums of squared deviations from a mean may be converted to an estimate of a variance. For the total sum of squares, there are $20 - 1 = 19$ D.F. For the within sample sum of squares, each sample, with 10 observations, contributes 9 D.F., and the total sum of squares within samples has 18 D.F. Finally, the between sample variation is based upon 2 sample means, or 1 D.F. These results may be summarized in a more or less standard table for the analysis of variance.

ANALYSIS OF VARIANCE TABLE

Source of variation	D.F.	Sum of squares	Variance estimate
Between samples	1	2000	2000
Within samples	18	1114	61.9
Total	19	3114	

If the two samples came from the same population, the two estimates of variance should be the same except for random sampling variation. The random sampling variation is taken into account by the probability distribution of a measure of discrepancy between the two independent estimates of variance. It happens that the measure of discrepancy for which the probability distribution is well established is the *ratio* of the two variance estimates: *between samples to within samples*. This ratio is given the symbol F , and if the hypothesis that the two samples come from the same population is true, this ratio is distributed according to the F distribution with the appropriate number of degrees of freedom. The F distribution is a probability distribution which depends upon two parameters: the number of degrees of freedom associated with each variance estimate. In this example D.F. = 1 for the numerator and D.F. = 18 for the denominator of F . Values of F at different probability levels for 1 and 18 degrees of freedom are tabled in many places. A complete table of F is not given here, but the references at the end of the chapter provide readily accessible tables.

The complete statistical decision process may be set out formally as follows.

$H_0: \mu_A = \mu_B$ (the two population means are equal)

$H_1: \mu_A > \mu_B$ (the mean of population A is the larger)

D.F. = 1 and 18 $\alpha = .05$ (for 1 and 18 degrees of freedom, the value of F which is exceeded with probability .05 is $F = 4.41$)

Statistical-decision Rule. Accept H_1 if $F \geq 4.41$; otherwise accept H_0 .

Decision. Accept H_1 because the observed value of $F = 32.3$ falls in the region which accepts H_1 and rejects H_0 . The conclusion is that product A has the higher mean breaking strength.

15.4 Several Sample Means

The basic method of testing for a difference between two sample means by the F test is readily extended to several samples. The arithmetic is a bit more involved but short-cut methods of computing the basic sums of squares diminish this aspect of the problem.

A study of TV viewing and viewer characteristics gave information on the average number of hours of TV viewing per person per day for adults

classified into six age groups. The individual observations are not given, but the totals that are important are reproduced.

Age	18-25	26-35	36-45	46-55	56-65	Over 65
n	23	50	62	35	18	12
Σx	73 6	125 0	167 4	105 0	43 2	52 8
$\bar{X}$	3 2	2 5	2 7	3 0	2 4	4 4
Σx^2	5,416 96	15,615 0	28,022 76	11,025 0	1,866 24	2,787 84

The mean number of hours of TV viewing is not the same for each age group in the sample. The question is: Do these data give evidence that the samples come from populations with different means, i.e., is the average TV viewing per person per day different for different age groups in the universe? Samples from the same population will give different sample means. Is the variation among the sample means large enough to conclude that the population means are not the same but differ with age?

The basic data and method of analysis are the same as for two samples. The arithmetic is simplified by adopting some short-cut methods. The sum of squares of deviations from the mean of each age group may be computed from a basic expansion learned earlier in Chap. 7. Equation (7.3) is reproduced here as a refresher.

$$s_x^2 = \frac{\Sigma(x - \bar{X})^2}{n - 1} = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1} \quad (7.3)$$

The numerator is computed from Σx and Σx^2 of the original data, in the form $\Sigma x^2 - (\Sigma x)^2/n$. From the data of the previous table, the sum of squares of the deviations from the sample means, i.e., the sum of squares within samples, is computed below.

Age	$\Sigma x^2 - \frac{(\Sigma x)^2}{n}$	Age	$\Sigma x^2 - \frac{(\Sigma x)^2}{n}$
18-25	39 5	46-55	25
26-35	74 5	56-65	15 3
36-45	123 0	Over 65	22 7

Example: For age group 46-55, $n = 35$, $\Sigma x = 105.0$ and $\Sigma x^2 = 340$. $\Sigma x^2 - (\Sigma x)^2/n = 340 - (105)^2/35 = 25$. The total sum of squares *within samples* is 300.0, the total of the six individual sample sums of squared deviations from their respective means.

Disregarding the age classifications, the 200 observations have $\Sigma x = 567.0$, $\bar{X} = 2.8$, and $\Sigma x^2 = 1951$. The total sum of squared deviations from the grand mean is $1951 - (567)^2/200 = 343.6$.

The sum of squares *between* samples may be found by subtraction in the analysis of variance table in which the data are summarized below.

ANALYSIS OF VARIANCE TABLE

Source of variation	D.F.	Sum of squares	Variance estimate
Between samples	5	43 6	8 72
Within samples	194	300 0	1 55
Total	199	343 6	

The sample $F = 8.72/1.55 = 5.6$. The formal statistical decision process is set out below.

$H_0: \mu_1 = \mu_2 = \dots = \mu_6$ (mean viewing time per person per day is the same for all six age groups)

$H_1: \mu_1 \neq \mu_2 \neq \dots \neq \mu_6$ (mean viewing times differ with age)

D.F. = 5 and 194 $\alpha = .05$

The table of F does not include values of D.F. = 5 and 194. The closest is D.F. = 5 and 200 which is satisfactory for this problem. The value of F for 5 and 200 D.F. which is exceeded with probability .05 is 2.26.

Statistical Decision Rule. Accept H_1 if the observed $F \geq 2.26$; otherwise accept H_0 .

Decision. Accept H_1 because the observed $F = 5.6$ falls in the region R . The decision is that the mean viewing time per person per day differs from one age group to the next. From the original data it appears that the oldest (over 65) and the youngest (18 to 25) adults view the most.

15.5 Design of Experiments

The analysis of variance technique, illustrated with two simple problems in Secs. 15.3 and 15.4, is a powerful and important technique in advanced statistics, for it enables experiments to be designed in such a way that the influence of different factors on the experimental outcome can be measured. Examples of some rather elaborate controlled experiments in marketing research are reported by Max E. Brunk and Walter T. Federer, *Experimental Designs and Probability Sampling in Marketing Research*, *Journal of the American Statistical Association*, vol. 48, no. 263, pp. 440-452, September, 1953. One of these investigations concerns the effects of different pricing, displaying, and packaging practices on the sale of apples.

For example, should apples be sold in 4-, 5-, 6-, or 8-pound Polythene bags? An experiment to determine which method of packaging results in the largest sales must take other factors into account as well. If the 4-pound bag is used on Wednesday and the 8-pound bag on Saturday, an increase in sales of apples is not necessarily attributable to the bigger bag alone, because grocery sales are normally higher on Saturday and the increased sale of apples may only reflect the increased sale of all groceries. Day-to-day variations in sales must be accounted for in order to isolate the change in sales ascribed to the difference in bags. Another source of variation in the sales of apples is the store in which the experiment is to be conducted. Apple sales are higher in some stores than in others. If the experiment is to be performed in several stores, the different levels of sales in these stores must be accounted for in order to ascribe increases in sales of apples to the size of the bag. The complete experimental design is given in the table below where the stores are numbered 1, 2, 3, and 4, and the basic treatments, the sizes of the bags, are lettered *A*, *B*, *C*, and *D*.

First part of the week					Second part of the week				
Day	Store				Day	Store			
	1	2	3	4		1	2	3	4
Monday	<i>B</i>	<i>C</i>	<i>D</i>	<i>A</i>	Friday A.M.	<i>B</i>	<i>A</i>	<i>C</i>	<i>D</i>
Tuesday	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	Friday P.M.	<i>C</i>	<i>D</i>	<i>B</i>	<i>A</i>
Wednesday	<i>D</i>	<i>A</i>	<i>B</i>	<i>C</i>	Saturday A.M.	<i>A</i>	<i>B</i>	<i>D</i>	<i>C</i>
Thursday	<i>C</i>	<i>D</i>	<i>A</i>	<i>B</i>	Saturday P.M.	<i>D</i>	<i>C</i>	<i>A</i>	<i>B</i>

Each treatment is given once in each store and once in each time period. On any one day, all four treatments are used, and over the whole time period, each store uses all four treatments. The arrangement of the treatments among the stores and the time intervals is random. With this experimental arrangement, the effect of the stores and the day of the week on apple sales may be accounted for, and the change in apple sales attributed directly to the different size bags measured.

This experimental design measures the effects of three factors. Other designs are available for two factors or for more than three factors. The traditional method of experimenting is to hold all factors but one constant and then to vary that one to measure its effects. The statistical design of experiments permits and, in fact, encourages several factors to vary together and then, by the method of analysis, measures the effects of each of the factors separately and also the joint effects of their acting

together. With the proper arrangement, these designs get the maximum amount of information from the fewest number of observations.

15.6 Final Comment

This chapter has only hinted at the vast number of problems for which statistics can provide solutions. Statistical methods will not solve all the problems of the world, even those dealing directly with numerical data. The application of the scientific method is becoming more widespread in all fields, including business and industry, and since this application depends largely upon making measurements, it rests heavily upon statistical methods.

If this book has taught you some basic principles of these methods or has prompted you to think about the possibility of application in your own field or to be more critical than you were in your interpretation of numerical data, it has accomplished its purpose. If it has inspired you to learn more about statistics, it has more than accomplished its purpose.

REFERENCES

Almost every publishing house has one or more books in elementary statistics which cover the same topics as this book. Most of them emphasize the descriptive rather than the decision-making aspects of statistics. Intermediate and advanced books are written to cover both the field of statistics in general and some very specialized topic. The short bibliography to follow lists some of the recent general books at both the intermediate and advanced level of study.

Intermediate

- Dixon, Wilfrid J., and Frank J. Massey, Jr.: "Introduction to Statistical Analysis," McGraw-Hill Book Company, Inc., New York, 1951.
- Fisher, R. A.: "Statistical Methods for Research Workers," 11th ed., Oliver & Boyd, Ltd., Edinburgh and London, 1950.
- Neyman, Jerzy: "First Course in Probability and Statistics," Henry Holt and Company, Inc., New York, 1950.
- Snedecor, George W., "Statistical Methods," 4th ed., Iowa State College Press, Ames, Iowa, 1946.
- Tippett, L. H. C.: "The Methods of Statistics," 4th ed., John Wiley & Sons, Inc., New York, 1952.
- Walker, Helen M., and Joseph Lev: "Statistical Inference," Henry Holt and Company, Inc., New York, 1953.

Advanced

- Hoel, Paul G.: "Introduction to Mathematical Statistics," 2d ed., John Wiley & Sons, Inc., New York, 1954.
- Kendall, M. G.: "The Advanced Theory of Statistics," 4th ed., vol. I, 1948 and 3d ed., vol. II, 1951, Charles Griffin & Co., Ltd., London.
- Mood, Alexander: "Introduction to the Theory of Statistics," McGraw-Hill Book Company, Inc., New York, 1950.

APPENDIX A: STATISTICAL TABLES

- A.1 Squares and Square Roots of Numbers
- A.2 Individual Terms of the Binomial Distribution for $n = 10$
- A.3 Areas of the Normal Distribution
- A.4 Probabilities of the t Distribution
- A.5 Confidence Interval for the Median
- A.6 Tolerance Factors for Normal Distributions
- A.7 Critical Values of the Rank Correlation Coefficient
- A.8 Probabilities of χ_p^2 for the Phase Test in Time Series ($n > 12$)
- A.9 Probabilities of the χ^2 Distribution
- A.10 Critical Values of the Correlation Coefficient
- A.11 Probability Levels of S for τ in Rank Correlation

TABLE A 1. SQUARES AND SQUARE ROOTS OF NUMBERS

<i>N</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	1 000	1 020	1 040	1 061	1 082	1 103	1 124	1 145	1 166	1 188
1.1	1 210	1 232	1 254	1 277	1 300	1 323	1 346	1 369	1 392	1 416
1.2	1 440	1 464	1 488	1 513	1 538	1 563	1 588	1 613	1 638	1 664
1.3	1 690	1 716	1 742	1 769	1 796	1 823	1 850	1 877	1 904	1 932
1.4	1 960	1 988	2 016	2 045	2 074	2 103	2 132	2 161	2 190	2 220
1.5	2 250	2 280	2 310	2 341	2 372	2 403	2 434	2 465	2 496	2 528
1.6	2 560	2 592	2 624	2 657	2 690	2 723	2 756	2 789	2 822	2 856
1.7	2 890	2 924	2 958	2 993	3 028	3 063	3 098	3 133	3 168	3 204
1.8	3 240	3 276	3 312	3 349	3 386	3 423	3 460	3 497	3 534	3 572
1.9	3 610	3 648	3 686	3 725	3 764	3 803	3 842	3 881	3 920	3 960
2.0	4 000	4 040	4 080	4 121	4 162	4 203	4 244	4 285	4 326	4 368
2.1	4 410	4 452	4 494	4 537	4 580	4 623	4 666	4 709	4 752	4 796
2.2	4 840	4 884	4 928	4 973	5 018	5 063	5 108	5 153	5 198	5 244
2.3	5 290	5 336	5 382	5 429	5 476	5 523	5 570	5 617	5 664	5 712
2.4	5 760	5 808	5 856	5 905	5 954	6 003	6 052	6 101	6 150	6 200
2.5	6 250	6 300	6 350	6 401	6 452	6 503	6 554	6 605	6 656	6 708
2.6	6 760	6 812	6 864	6 917	6 970	7 023	7 076	7 129	7 182	7 236
2.7	7 290	7 344	7 398	7 453	7 508	7 563	7 618	7 673	7 728	7 784
2.8	7 840	7 896	7 952	8 009	8 066	8 123	8 180	8 237	8 294	8 352
2.9	8 410	8 468	8 526	8 585	8 644	8 703	8 762	8 821	8 880	8 940
3.0	9 000	9 060	9 120	9 181	9 242	9 303	9 364	9 425	9 486	9 548
3.1	9 610	9 672	9 734	9 797	9 860	9 923	9 986	10 05	10 11	10 18
3.2	10 24	10 30	10 37	10 43	10 50	10 56	10 63	10 69	10 76	10 82
3.3	10 89	10 96	11 02	11 09	11 16	11 22	11 29	11 36	11 42	11 49
3.4	11 56	11 63	11 70	11 76	11 83	11 90	11 97	12 04	12 11	12 18
3.5	12 25	12 32	12 39	12 46	12 53	12 60	12 67	12 74	12 82	12 89
3.6	12 96	13 03	13 10	13 18	13 25	13 32	13 40	13 47	13 54	13 62
3.7	13 69	13 76	13 84	13 91	13 99	14 06	14 14	14 21	14 29	14 36
3.8	14 44	14 52	14 59	14 67	14 75	14 82	14 90	14 98	15 05	15 13
3.9	15 21	15 29	15 37	15 44	15 52	15 60	15 68	15 76	15 84	15 92
4.0	16 00	16 08	16 16	16 24	16 32	16 40	16 48	16 56	16 65	16 73
4.1	16 81	16 89	16 97	17 06	17 14	17 22	17 31	17 39	17 47	17 56
4.2	17 64	17 72	17 81	17 89	17 98	18 06	18 15	18 23	18 32	18 40
4.3	18 49	18 58	18 66	18 75	18 84	18 92	19 01	19 10	19 18	19 27
4.4	19 36	19 45	19 54	19 62	19 71	19 80	19 89	19 98	20 07	20 16
4.5	20 25	20 34	20 43	20 52	20 61	20 70	20 79	20 88	20 98	21 07
4.6	21 16	21 25	21 34	21 44	21 53	21 62	21 72	21 81	21 90	22 00
4.7	22 09	22 18	22 28	22 37	22 47	22 56	22 66	22 75	22 85	22 94
4.8	23 04	23 14	23 23	23 33	23 43	23 52	23 62	23 72	23 81	23 91
4.9	24 01	24 11	24 21	24 30	24 40	24 50	24 60	24 70	24 80	24 90
5.0	25 00	25 10	25 20	25 30	25 40	25 50	25 60	25 70	25 81	25 91
5.1	26 01	26 11	26 21	26 32	26 42	26 52	26 63	26 73	26 83	26 94
5.2	27 04	27 14	27 25	27 35	27 46	27 56	27 67	27 77	27 88	27 98
5.3	28 09	28 20	28 30	28 41	28 52	28 62	28 73	28 84	28 94	29 05
5.4	29 16	29 27	29 38	29 48	29 59	29 70	29 81	29 92	30 03	30 14
5.5	30 25	30 36	30 47	30 58	30 69	30 80	30 91	31 02	31 14	31 25
5.6	31 36	31 47	31 58	31 70	31 81	31 92	32 04	32 15	32 26	32 38
5.7	32 49	32 60	32 72	32 83	32 95	33 06	33 18	33 29	33 41	33 52
5.8	33 64	33 76	33 87	33 99	34 11	34 22	34 34	34 46	34 57	34 69
5.9	34 81	34 93	35 05	35 16	35 28	35 40	35 52	35 64	35 76	35 88

TABLE A 1. SQUARES AND SQUARE ROOTS OF NUMBERS (Continued)

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
6 0	36 00	36 12	36 24	36 36	36 48	36 60	36 72	36 84	36 97	37 09
6 1	37 21	37 33	37 45	37 58	37 70	37 82	37 95	38 07	38 19	38 32
6 2	38 44	38 56	38 69	38 81	38 94	39 06	39 19	39 31	39 44	39 56
6 3	39 69	39 82	39 94	40 07	40 20	40 32	40 45	40 58	40 70	40 83
6 4	40.96	41.09	41.22	41.34	41.47	41.60	41.73	41.86	41.99	42.12
6.5	42.25	42.38	42.51	42.64	42.77	42.90	43.03	43.16	43.30	43.43
6.6	43 56	43 69	43 82	43 96	44 09	44 22	44 36	44 49	44 62	44 76
6 7	44.89	45 02	45 16	45 29	45 43	45 56	45 70	45 83	45 97	46 10
6 8	46 24	46 38	46 51	46 65	46 79	46 92	47 06	47 20	47 33	47 47
6.9	47 61	47.75	47.89	48.02	48.16	48.30	48.44	48.58	48.72	48 86
7.0	49 00	49.14	49.28	49.42	49.56	49.70	49.84	49.98	50.13	50.27
7 1	50 41	50.55	50 69	50 84	50 98	51 12	51 27	51 41	51 55	51 70
7 2	51 84	51 98	52 13	52 27	52 42	52 56	52 71	52 85	53 00	53 14
7.3	53 29	53 44	53 58	53.73	53.88	54.02	54.17	54 32	54 46	54.61
7 4	54.76	54.91	55.06	55.20	55.35	55.50	55.65	55.80	55.95	56.10
7 5	56 25	56.40	56 55	56 70	56.85	57 00	57 15	57.30	57.46	57 61
7 6	57 76	57 91	58 06	58 22	58 37	58.52	58 68	58.83	58 98	59 14
7 7	59 29	59 44	59 60	59 75	59 91	60 06	60 22	60 37	60 53	60 68
7 8	60 84	61 00	61 15	61 31	61 47	61 62	61 78	61 94	62 09	62 25
7 9	62.41	62 57	62.73	62.88	63 04	63.20	63.36	63.52	63.68	63 84
8 0	64 00	64 16	64 32	64.48	64 64	64 80	64 96	65 12	65 29	65 45
8 1	65 61	65 77	65 93	66 10	66 26	66 42	66 59	66 75	66 91	67 08
8 2	67 24	67 40	67 57	67 73	67 90	68 06	68 23	68 39	68 56	68 72
8 3	68 89	69 06	69 22	69 39	69 56	69 72	69 89	70 06	70 22	70 39
8.4	70 56	70.73	70.90	71.06	71 23	71.40	71 57	71.74	71.91	72.08
8 5	72 25	72.42	72.59	72.76	72.93	73.10	73.27	73 44	73 62	73 79
8 6	73 96	74 13	74 30	74 48	74 65	74 82	75 00	75 17	75 34	75 52
8 7	75 69	75 86	76 04	76 21	76 39	76 56	76 74	76 91	77 09	77 26
8 8	77 44	77 62	77 79	77 97	78 15	78 32	78 50	78 68	78 85	79 03
8.9	79.21	79.39	79.57	79.74	79.92	80 10	80.28	80.46	80.64	80 82
9.0	81.00	81 18	81 36	81 54	81 72	81 90	82.08	82 26	82.45	82 63
9 1	82 81	82 99	83 17	83 36	83 54	83 72	83 91	84 09	84.27	84 46
9 2	84 64	84 82	85 01	85 19	85 38	85 56	85 75	85 93	86.12	86 30
9.3	86 49	86 68	86 86	87 05	87 24	87 42	87 61	87 80	87.98	88.17
9.4	88.36	88.55	88.74	88.92	89.11	89.30	89.49	89.68	89.87	90.06
9.5	90 25	90.44	90 63	90 82	91.01	91.20	91.39	91.58	91.78	91 97
9 6	92 16	92 35	92 54	92 74	92 93	93 12	93 32	93 51	93 70	93 90
9 7	94 09	94 28	94 48	94 67	94 87	95 06	95 26	95 45	95 65	95 84
9.8	96 04	96.24	96 43	96 63	96 83	97 02	97 22	97 42	97 61	97 81
9.9	98.01	98.21	98.41	98.60	98.80	99.00	99.20	99.40	99.60	99.80

The square root of a number may be found by finding the first three places from this table, dividing the number by the three-place root and averaging the quotient and divisor. Repetition will give the square root to any desired accuracy. For example, $\sqrt{80.00}$ is just less than 8.95. First, $80.00/8.95 = 8.9385475$. Second, $(8.95 + 8.9385475)/2 = 8.9442738$, which is close to the more precise value of 8.9442719.

TABLE A.2. INDIVIDUAL TERMS OF THE BINOMIAL DISTRIBUTION FOR $n = 10$

P	$x = 10$	$x = 9$	$x = 8$	$x = 7$	$x = 6$	$x = 5$
05	00000	.00000	00000	.00000	.00000	00006
.10	00000	00000	.00000	00001	.00014	.00149
.15	00000	00000	00001	00013	00125	00849
.20	00000	00000	00007	00079	00551	02642
.25	00000	.00003	.00039	00309	01622	.05840
.30	.00001	.00014	.00145	00900	03676	.10292
.35	.00003	.00051	.00428	02120	.06891	15357
.40	.00010	.00157	.01062	04247	11148	20066
.45	.00034	.00416	.02289	.07460	15957	.23403
.50	.00098	.00977	.04395	.11719	20508	.24609

P	$x = 4$	$x = 3$	$x = 2$	$x = 1$	$x = 0$
05	00096	.01048	07463	31512	59874
.10	.01116	.05740	19371	38742	.34868
.15	.04010	.12983	27590	34743	19687
.20	.08808	20133	30199	26844	10737
.25	14600	.25028	.28157	18771	05631
.30	.20012	.26683	23347	12106	02825
.35	.23767	25222	.17565	07249	01346
.40	25082	.21499	12093	04031	00605
.45	.23837	.16648	07630	02072	00253
.50	.20508	11719	04395	.00977	.00098

TABLE A 3. AREAS OF THE NORMAL DISTRIBUTION

An entry in the table is the fraction of the area under the whole curve which is between $Z = 0$ and a positive value of Z . This area is shown in black on the following graph. For example, the area between $Z = 0$ and $Z + 1.82$ is .4656. Since the normal curve is symmetrical, the area between $Z = 0$ and $Z - 1.82$ is also .4656. The area between -1.82 and $+1.82$ is .9312. The area in the tails of the curve beyond ± 1.82 is .0688.

[illegible]

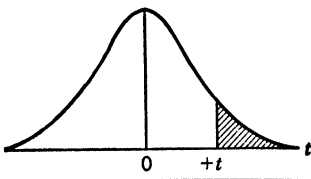


TABLE A.4. PROBABILITIES OF THE *t* DISTRIBUTION*

The quantity tabled is the value of *t* which for a given number of degrees of freedom is exceeded with the given probability. The area is shown in the diagram.

Degrees of freedom	Probability				
	.10	.05	.025	.01	.005
1	3 078	6 314	12 706	31 821	63 657
2	1 886	2 920	4 303	6 965	9 925
3	1 638	2 353	3 182	4 541	5 841
4	1 533	2 132	2 776	3 747	4 604
5	1 476	2 015	2 571	3 365	4 032
6	1 440	1 943	2 447	3 143	3 707
7	1 415	1 895	2 365	2 998	3 499
8	1 397	1 860	2 306	2 896	3 355
9	1 383	1 833	2 262	2 821	3 250
10	1 372	1 812	2 228	2 764	3 169
11	1 363	1 796	2 201	2 718	3 106
12	1 356	1 782	2 179	2 681	3 055
13	1 350	1 771	2 160	2 650	3 012
14	1 345	1 761	2 145	2 624	2 997
15	1 341	1 753	2 131	2 602	2 947
16	1 337	1 746	2 120	2 583	2 921
17	1 333	1 740	2 110	2 567	2 898
18	1 330	1 734	2 101	2 552	2 878
19	1 328	1 729	2 093	2 539	2 861
20	1 325	1 725	2 086	2 528	2 845
21	1 323	1 721	2 080	2 518	2 831
22	1 321	1 717	2 074	2 508	2 819
23	1 319	1 714	2 069	2 500	2 807
24	1 318	1 711	2 064	2 492	2 797
25	1 316	1 708	2 060	2 485	2 787
26	1 315	1 706	2 056	2 479	2 779
27	1 314	1 703	2 052	2 473	2 771
28	1 313	1 701	2 048	2 467	2 763
29	1 311	1 699	2 045	2 462	2 756
30	1 310	1 697	2 042	2 457	2 750
40	1 303	1 684	2 021	2 423	2 704
60	1 296	1 671	2 000	2 390	2 660
120	1 289	1 658	1 980	2 358	2 617
∞	1 282	1 615	1 960	2 326	2 576

* Table A.4 is abridged from Table III of Fisher and Yates, "Statistical Tables for Biological, Agricultural, and Medical Research," published by Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

TABLE A.5. CONFIDENCE INTERVAL FOR THE MEDIAN*

n	C.C. $\geq .95$			C.C. $\geq .99$		
	k	n - k + 1	Exact C.C.	k	n - k + 1	Exact C.C.
6	1	6	9844			
7	1	7	9922			
8	1	8	9961	1	8	.9961
9	2	8	9805	1	9	9980
10	2	9	9893	1	10	9990
11	2	10	9941	1	11	9995
12	3	10	9807	2	11	9968
13	3	11	9888	2	12	9983
14	3	12	9935	2	13	9991
15	4	12	9821	3	13	9963
16	4	13	9894	3	14	9979
17	5	13	9755	3	15	9988
18	5	14	9846	4	15	9962
19	5	15	9904	4	16	9978
20	6	15	9793	4	17	9987
21	6	16	9867	5	17	9964
22	6	17	9915	5	18	9978
23	7	17	9827	5	19	9987
24	7	18	9887	6	19	9967
25	8	18	9784	6	20	9980
26	8	19	9855	7	20	9953
27	8	20	9904	7	21	9970
28	9	20	9822	7	22	9981
29	9	21	9879	8	22	9959
30	10	21	9786	8	23	9974
31	10	22	9853	8	24	9987
32	10	23	9900	9	24	9965
33	11	23	9825	9	25	9977
34	11	24	9878	10	25	9955
35	12	24	9795	10	26	9970
40	14	27	9808	12	29	9968
45	16	30	9882	14	32	9967
50	18	33	9836	16	35	9967
55	20	36	9850	18	38	9968
60	22	39	9863	20	41	9969
65	25	41	9768	22	44	9970
70	27	44	9793	24	47	9972
75	29	47	9815	26	50	9974

SOURCE: Reproduced from *Sankhya: Indian Journal of Statistics*, vol. 4, part 4, pp. 556-557, March, 1940, with special permission of the Statistical Publishing Society, Calcutta 35, India.

* See Sec. 9.11 for explanation of use.

TABLE A.6. TOLERANCE FACTORS FOR NORMAL DISTRIBUTIONS*

P n	C.C. = 0.95				C.C. = 0.99			
	0.75	0.90	0.95	0.99	0.75	0.90	0.95	0.99
2	22 858	32 019	37.674	48 430	114 363	160 193	188.491	242 300
3	5 922	8.380	9.916	12 861	13 378	18.930	22.401	29.055
4	3 779	5 369	6 370	8 299	6 614	9 398	11.150	14 527
5	3.002	4 275	5.079	6 634	4 643	6 612	7 855	10 260
6	2 604	3.712	4.414	5.775	3.743	5.337	6 345	8 301
7	2.361	3 369	4.007	5.248	3 233	4 613	5 488	7 187
8	2 197	3.136	3 732	4 891	2.905	4 147	4 936	6 468
9	2.078	2 967	3 532	4 631	2.677	3.822	4 550	5 966
10	1 987	2.839	3.379	4 433	2.508	3.582	4 265	5 591
11	1 916	2 737	3.259	4 277	2 378	3 397	4.045	5 308
12	1 858	2 655	3.162	4 150	2 274	3.250	3.870	5 079
13	1 810	2 587	3 081	4.044	2 190	3 130	3 727	4 893
14	1.770	2 529	3 012	3 955	2 120	3 029	3 608	4 737
15	1 735	2.480	2 954	3 878	2 060	2 945	3 507	4 605
16	1.705	2 437	2 903	3 812	2 009	2 872	3 421	4.492
17	1 679	2 400	2 858	3 754	1 965	2.808	3 345	4 393
18	1 655	2 366	2 819	3 702	1 926	2 753	3 279	4 307
19	1 635	2 337	2 784	3 656	1 891	2 703	3 221	4 230
20	1 616	2 310	2 752	3 615	1 860	2 659	3.168	4 161
21	1 599	2 286	2 723	3 577	1 833	2 620	3.121	4 100
22	1.584	2 264	2 697	3 543	1 808	2 584	3 078	4.044
23	1 570	2 244	2.673	3 512	1 785	2 551	3 040	3 993
24	1 557	2 225	2 651	3 483	1 764	2 522	3 004	3.947
25	1.545	2.208	2 631	3 457	1 745	2.494	2 972	3 904
30	1 497	2 140	2.549	3.350	1.668	2 385	2.841	3.733
40	1 435	2 052	2 445	3 213	1 571	2.247	2 677	3 518
50	1 396	1 996	2.379	3 126	1.512	2.162	2 576	3 385
60	1 369	1 958	2 333	3.066	1.471	2 103	2 506	3 293
70	1 349	1 929	2 299	3 021	1 440	2 060	2 454	3 225
80	1.334	1 907	2.272	2 986	1.417	2.026	2 414	3.173
90	1.321	1 889	2 251	2 958	1.398	1.999	2 382	3.130
100	1 311	1 874	2 233	2 934	1 383	1 977	2 355	3.096
110	1 302	1 861	2 218	2 915	1 369	1.958	2 333	3.066
120	1 294	1 850	2.205	2 898	1.358	1.942	2 314	3.041
130	1 288	1.841	2.194	2.883	1.349	1.928	2 298	3.019
140	1 282	1 833	2 184	2 870	1 340	1.916	2 283	3 000
150	1 277	1 825	2 175	2 859	1 332	1.905	2 270	2 983
160	1 272	1 819	2 167	2 848	1 326	1 896	2 259	2.968
170	1 268	1 813	2 160	2 839	1 320	1 887	2 248	2.955
180	1 264	1 808	2 154	2 831	1 314	1 879	2 239	2 942

TABLE A.6. TOLERANCE FACTORS FOR NORMAL DISTRIBUTIONS (*Continued*)

P n	C.C. = 0.95				C.C. = 0.99			
	0.75	0.90	0.95	0.99	0.75	0.90	0.95	0.99
200	1 258	1 798	2 143	2 816	1 304	1 865	2 222	2 921
225	1 251	1 788	2 131	2 800	1 294	1 851	2 205	2 898
250	1 245	1 780	2 121	2 788	1 286	1 839	2 191	2 880
275	1 240	1 773	2 113	2 777	1 279	1 829	2 179	2 864
300	1 236	1 767	2 106	2 767	1 273	1 820	2 169	2 850
350	1 229	1 757	2 094	2 752	1 263	1 805	2 151	2 827
400	1 223	1 749	2 084	2 739	1 255	1 794	2 138	2 809
450	1 219	1 743	2 077	2 729	1 248	1 785	2 127	2 795
500	1 215	1 737	2 070	2 721	1 243	1 777	2 117	2 783
750	1 202	1 719	2 049	2 692	1 225	1 751	2 086	2 742
1000	1 195	1 709	2 036	2 676	1 214	1 736	2 068	2 718
∞	1 150	1 645	1 960	2 576	1 150	1 645	1 960	2 576

* Reproduced with permission from C. Eisenhart, M. W. Hastay, and W. A. Wallis, "Techniques of Statistical Analysis," McGraw-Hill Book Company, Inc., New York, 1947, Chap. 2.

TABLE A.7. CRITICAL VALUES OF THE RANK CORRELATION COEFFICIENT*

The entry in the table is that value of the rank correlation coefficient r' which, for the given sample size, is exceeded with probability .05 or .01.

n	.05	.01	n	.05	.01
4	1 000		16	425	601
5	900	1 000	18	399	564
6	829	913	20	377	534
7	711	893	22	359	508
8	643	833	24	343	485
9	600	783	26	329	465
10	564	746	28	317	448
12	506	712	30	306	432
14	456	645			

* This table is adapted from Edwin G. Olds, Distributions of Sums of Squares of Rank Differences for Small Numbers of Individuals, *Annals of Mathematical Statistics*, vol. 9, pp. 133-148, 1938, with the permission of the author and the Institute of Mathematical Statistics.

TABLE A.8. PROBABILITIES OF χ_p^2 FOR THE PHASE TEST IN TIME SERIES
($n > 12$)*

The probability tabled is the probability that an observed χ_p^2 will equal or exceed the specified value for $n > 12$.

χ_p^2	Probability	χ_p^2	Probability
5 448	10	8 50	024
5 50	098	8 75	021
5 674	09	8 836	02
5 75	087	9 00	019
5 927	08	9 25	017
6 00	077	9 50	015
6 163	07	9 75	013
6 25	069	10 00	012
6 50	061	10 25	010
6 541	06	10 312	01
6 75	054	10 50	009
6 898	05	10 75	008
7 00	048	11 00	007
7 25	043	11 25	006
7 401	04	11 50	006
7 50	038	11 755	005
7 75	034	12 00	004
8 00	030	13 00	003
8 009	03	14 00	002
8 25	027	15 085	001

* Table A 8 is abridged from W. Allen Wallis and Geoffrey H. Moore, A Significance Test for Time Series Analysis, *Journal of the American Statistical Association*, vol. 36, no. 215, p. 404, September, 1941, with the permission of the Association.

TABLE A.9. PROBABILITIES OF THE χ^2 DISTRIBUTION*

The tabled value is the value of χ^2 which, for a given number of degrees of freedom, is exceeded with the given probability.

D.F.	Probability				
	.10	.05	.025	.01	.005
1	2 71	3 84	5.02	6.63	7 88
2	4 61	5.99	7.38	9 21	10 60
3	6 25	7.81	9 35	11 34	12 84
4	7 78	9.49	11 14	13 28	14 86
5	9.24	11 07	12 83	15 09	16 75
6	10 64	12 59	14.45	16 81	18 55
7	12 02	14.07	16 01	18 48	20 28
8	13 36	15 51	17 53	20 09	21 96
9	14 68	16 92	19 02	21 67	23 59
10	15 99	18.31	20 48	23 21	25 19
11	17 28	19 68	21 92	24 73	26 76
12	18 55	21 03	23 34	26 22	28 30
13	19 81	22 36	24 74	27 69	29 82
14	21 06	23 68	26 12	29 14	31 32
15	22 31	25 00	27 49	30 58	32.80
16	23 54	26 30	28 85	32 00	34 27
18	25 99	28 87	31 53	34 81	37 16
20	28 41	31 41	34 17	37 57	40 00
24	33 20	36 42	39 36	42 98	45 56
30	40 26	43 77	46 98	50.89	53 67
40	51 81	55 76	59 34	63 69	66 77
60	74.40	79 08	83 30	88.38	91.95
120	140 23	146 57	152 21	158 95	163 64

* Table A.9 is abridged from Table IV of Fisher and Yates, "Statistical Tables for Biological, Agricultural, and Medical Research," published by Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

For larger values of n , the expression $\sqrt{2\chi^2} - \sqrt{2n - 1}$ may be used as a normal deviate with standard deviation equal to 1, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve.

TABLE A.10. CRITICAL VALUES OF THE CORRELATION COEFFICIENT*

The entry in the table is the value of the correlation coefficient r which, for a given number of degrees of freedom, is exceeded with the given probability.

D.F.	Probability		D.F.	Probability	
	.05	.01		.05	.01
1	997	1.000	24	388	.496
2	950	990	25	381	.487
3	878	959	26	374	.478
4	811	917	27	367	.470
5	.754	.874	28	.361	.463
6	707	834	29	355	.456
7	666	.798	30	349	.449
8	632	765	35	.325	.418
9	.602	.735	40	304	.393
10	.576	.708	45	288	.372
11	553	684	50	273	.354
12	532	661	60	250	.325
13	514	.641	70	232	.302
14	497	623	80	217	.283
15	.482	606	90	205	.267
16	.468	.590	100	.195	.254
17	456	.575	125	174	.228
18	444	561	150	159	.208
19	433	549	200	138	.181
20	423	537	300	113	.148
21	413	526	400	098	.128
22	404	515	500	088	.115
23	396	505	1,000	062	.081

* Portions of Table A.10 were abridged from Table VI of Fisher and Yates, "Statistical Tables for Biological, Agricultural, and Medical Research," published by Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

TABLE A.11. PROBABILITY LEVELS OF S FOR τ IN RANK CORRELATION*
 Probability that S (for τ) attains or exceeds a specified value. (Shown only for positive values. Negative values obtainable by symmetry.)

S	Values of n				S	Values of n		
	4	5	8	9		6	7	8
0	0 625	0 592	0 548	0 540	1	0 500	0 500	0 500
2	0 375	0 408	0 452	0 460	3	0 360	0 386	0 431
4	0 167	0 242	0 360	0 381	5	0 235	0 281	0 364
6	0 042	0 117	0 271	0 306	7	0 136	0 191	0 300
8		0 042	0 199	0 238	9	0 068	0 119	0 242
10		0 0083	0 138	0 179	11	0 028	0 068	0 190
12			0 089	0 130	13	0 0083	0 035	0 146
14			0 054	0 090	15	0 0014	0 015	0 108
16			0 031	0 060	17		0 0054	0 078
18			0 016	0 038	19		0 0014	0 054
20			0 0071	0 022	21			0 036
22			0 0028	0 012	23			0 023
24				0 0063	25			0 014
26				0 0029	27			0 0083
28				0 0012	29			0 0046
30					31			0 0023
32					33			0 0011

* Reproduced from "Rank Correlation Methods" by M. G. Kendall, Charles Griffin & Co., Ltd., by permission of author and publisher.

APPENDIX B: GLOSSARY OF SYMBOLS

The more commonly used symbols are listed below with a short explanation of their usage and a section number, in parentheses, illustrating this use. Symbols which appear only once in a rather special situation (for example, in the econometric model of Sec. 13.6) are not listed. In some instances, the same symbol is used in two or more contexts. For example, k is used as the multiple of the standard deviation (7.12), as the number of cells in a frequency distribution (8.1), and as a subscript to x in setting confidence limits for the median (9.11). In none of these uses is k a symbol to be particularly remembered (as, for example, is $\bar{X}$ for the sample mean), and little confusion should arise because of this multiple usage.

- a Sample intercept, estimate of A , in $y' = a + bx$ (10.4)
- A Region of acceptance (3.1)
- A Universe intercept constant in linear regression equation $Y = A + Bx$ (10.3)
- α Probability of error I (3.2); as a subscript, it indicates that value of the variable which cuts off α per cent in one tail of a distribution, for example, Z_α (4.6, 9.4), t_α (9.5)
- AQL Acceptable quality level (4.8)
- b Sample regression coefficient, estimate of B , in $y' = a + bx$ (10.4)
- B Universe regression coefficient in $Y = A + Bx$ (10.3)
- β Probability of error II (3.2), as a subscript, it indicates that value of the variable which cuts off β per cent of the distribution in one tail, for example, Z_β (4.6, 9.4)
- c Number of parameters estimated from a sample in computing degrees of freedom, for example, $k - c - 1$ (8.3)
- C.C. Confidence coefficient (5.2); as a subscript, it indicates the plus and minus values of a variable which include C.C. per cent of the distribution between them, for example, $Z_{C.C.}$ (9.8), $t_{C.C.}$ (9.8)
- χ^2 Chi-square (8.1, 15.1)
- χ^2_r χ^2 variable used in phase test for cycle (11.6) and seasonal (12.2)
- C.V. Coefficient of variation (7.11)
- d Half the length of a desired confidence interval (9.9)
- d Difference between ranks in a rank correlation (10.11)
- $d_{c.c.}$ Deviation between the population and sample cumulative relative frequency distributions (8.6)
- D.F. Degrees of freedom (9.5)
- $\Delta x, \Delta y, \Delta t$ Increment or change in value (10.4, 11.2)
- E Expected frequency in χ^2 test (8.1)
- E_i Events in event space (2.3)
- $E(x)$ Expected value of x (2.3)
- f_i Frequency of occurrence (1.8, 2.4); also f (7.8)
- $f(x)$ Probability distribution (2.3); binomial distribution (2.4); normal distribution (4.1)
- F F distribution (15.3, 15.4)
- h Used as a subscript to identify a stratum, for example, n_h, σ_h, N_h (6.8)

- H_0 Hypothesis to be tested (3.1); the subscript 0 is attached to parameters specified by H_0 , for example, P_0 (3.1), μ_0 (9.4), B_0 (10.5)
- H_1 Alternative hypothesis (3.1); the subscript 1 is attached to parameters specified by H_1 , for example, P_1 (3.1), μ_1 (9.4), B_1 (10.5)
- k Multiple of standard deviation in Tchebycheff's inequality (7.12)
- k Number of groups or cells in a frequency distribution (8.1)
- k A subscript to x indicating confidence limits for the median (9.11)
- K Multiple of sample standard deviation in a tolerance interval (9.13)
- LCL Lower control limit (11.4)
- LTPD Lot tolerance per cent defective (4.8)
- M Number of positive scores in rank correlation for a trend (13.1)
- μ The parameter designation of the arithmetic mean (2.7); a subscript attached indicates the variable, for example, μ_x (9.2), $\mu_{\bar{x}}$ (9.2), μ_b (10.5), μ_s (13.1)
- $\hat{\mu}$ Maximum-likelihood estimate of μ (9.8)
- $\underline{\mu}$ Lower confidence limit of μ (9.8)
- $\bar{\mu}$ Upper confidence limit of μ (9.8)
- n Sample size (2.1a)
- N Number of samples (1.7, 1.8, 2.7)
- N Size of population (6.8, 8.2)
- O Observed frequency in a χ^2 test (8.1)
- OC Operating-characteristic curve (3.5)
- p , Relative frequency of occurrence (1.8), sample percentage (5.1)
- $\hat{p}$ Maximum-likelihood estimate of P (5.1)
- p_i , Price of i th commodity in the j th time period (14.2)
- P Population proportion in one category (2.4)
- $\underline{P}$ Lower confidence limit of P (5.2)
- $\bar{P}$ Upper confidence limit of P (5.2)
- Pr Probability, for example, $\text{Pr}(x_i)$, $\text{Pr}(E_i)$ probability of x_i or E_i (2.3); binomial probabilities (2.4), normal distribution probabilities (4.1)
- q_i , Quantity purchased of the i th commodity in the j th time period (14.2)
- Q Population percentage in other category ($Q = 1 - P$) (2.4)
- Q_1, Q_2, Q_3 Quartiles (7.7)
- r A subscript indicating value of variable which defines region of rejection (3.4), for example, x_r (3.4), $\bar{X}_r$ (9.4), b_r (10.5), S_r (13.1)
- r Sample correlation coefficient (10.7)
- r' Rank correlation coefficient (10.11)
- R Region of rejection (3.1)
- $R(t)$ Random component of a time series (12.1)
- ρ Population correlation coefficient (10.9)
- ρ' Population rank correlation coefficient (10.11)
- s Standard deviation estimated from a sample; a subscript is attached to indicate the variable, for example, s_x (7.7, 9.5), s_y (7.13, 10.7), s_t (8.2), $s_{\bar{x}}$ (9.5), $s_{x'}$ (9.10), s_b (10.5)
- s^2 Variance estimated from a sample (7.7)
- $s_{y \cdot x}$ Sample standard deviation of y values around the regression $y' = a + bx$ (10.4); also $s_{y \cdot t}$ in trend analysis (13.1)
- S Event space (2.3)
- S Total score in rank correlation coefficient τ (13.1)
- $S(t)$ Seasonal component of a time series (12.1)
- σ Parameter value of the standard deviation (2.8); a subscript is attached to indicate the variable, for example, σ_x (9.2), $\sigma_{\bar{x}}$ (9.2), σ_s (13.1)

- σ^2 Parameter variance (2.8)
- σ_{y-x} Universe standard deviation of y values around regression $Y = A + Bx$ (10.3)
- Σ Summation sign (2.1)
- t Designation of a time variable (8.2)
- t t distribution (9.5)
- $T(t)$ Trend component of a time series (12.1, 13.1)
- τ Rank correlation coefficient (13.1)
- u Observed number of phases in phase test for cycle (11.6)
- U Expected number of phases in phase test for cycle (11.6)
- UCL Upper control limit (11.4)
- x_i Random variable (1.8); also used are X (1.6) and X_i (2.1)
- $\tilde{x}$ Median of a sample (7.7, 9.11)
- $\bar{X}$ Arithmetic mean of a sample (2.2, 9.1); the bar over a variable indicates the sample mean, for example, $\bar{Y}$ (7.13, 10.4), $\bar{t}$ (8.2)
- X_p Percentile, for example, X_{90} (7.7)
- X' Estimated population total (9.10)
- y Dependent variable in regression analysis (10.3) and in time series analysis (11.2, 11.6)
- y' Mean value of y in sample regression line $y' = a + bx$ (10.4) and adjusted figures in time series analysis (12.2)
- Y Universe mean value of y in a linear relationship $Y = A + Bx$ (10.3)
- Z Tabulated normal distribution variable (4.1); binomial approximation (4.3), $\bar{X}$ (9.4)

INDEX

- Acceptable quality level (AQL), 92
- Acceptance sampling, industrial quality control, 92
- Addition theorem of probabilities, 32, 33
- Alpha, 55
 - meaning of, to fix α , 56
 - relation to producer's risk, 92
- Analysis of variance, in design of marketing experiment, 365
 - test, for difference between two population means, 363
 - for equality of several population means, 365
- Area sample, 127
- Arithmetic mean (*see* Mean)
- Array, 133
 - mean from, 150
 - median from, 148
 - mode from, 150
 - percentiles from, 149
 - quartile from, 149
 - standard deviation from, 152
- Assignable cause of control charts, 275
- Autoregressive mechanism of a cycle, 326
- Averages (*see* Mean; Median; Mode)

- Beta, definition of, 55
 - relation to consumer's risk, 92
- Binomial distribution, 39
 - confidence-interval table for, 102
 - derivation for $n = 2$ and $n = 3$, 41
 - equation of, 40
 - mean of, 47
 - normal approximation to, 77
 - probability model of, 39
 - scoring variable for, 19
 - standard deviation of, 49
 - tables, reference to, 42
 - use of (*see* Statistical tests)
- Binomial distribution, variance of, 49
 - (*See also* Frequency distribution; Normal distribution)
- Bivariate normal distribution, 252
- Bus-motor failures, analysis of, 171
 - frequency distributions, 178

- Census of population, age definition of, 121
- Central limit theorem, 203
- Chi-square (χ^2), introduction to, 169
 - use in statistical tests (*see* Statistical tests)
- Class intervals, 136
- Cluster sampling, 125
- Coding data for simplified computation, 161
- Coefficient of variation, 159
- Computation, aids to, 27
 - coding for simplification of, 161
- Computation machines, use of, 28
- Confidence interval, for population mean, 208
 - for population median, 216
 - for population percentage, 100
 - for population relative frequency distribution, 179
 - for regression estimates, 243
 - table for binomial distribution, 102
- Consumer's risk, 92
- Continuous variable, 36
- ✓ Control charts, 275
- Cowles Commission for Research in Economics, time series models of, 330
- Correlation, 226
 - bivariate normal distribution, 252
 - method of least squares, 250
 - rank, 253
- Correlation coefficient, equation of, 248
 - measure of improvement in standard deviation, 248

- Correlation coefficient, use of, cautions on, 252
 - in testing for universe relationship, 250
- Cumulative frequency distribution, 155
- Cumulative relative frequency distribution, estimation of population, 179
- Curve fitting, examples of, 168
 - normal distribution, 171
 - testing goodness of fit, 168
- Cyclical fluctuations, examples of, 264
 - explanation of, 325
 - as percentage deviation from trend, 319
 - phase test for, 284, 293, 322
- Data, primary, 118
 - secondary, 120
- Decision problems, everyday, 67
 - statistical (*see* Statistical tests)
- Degrees of freedom, of chi-square, 170, 174, 176, 359, 360
 - of F , 363, 365
 - of t , 199, 241
- Dependent variable, 225
- Design of experiments, example of
 - marketing problem, 365
- Discrete variable, 32
- Dispersion measures, 48, 49, 151-153, 156
- Elementary unit, 8
- Errors, of first kind, 22, 54
 - of second kind, 23, 54
 - in testing hypotheses, 53
- Estimation, of cumulative relative frequency distribution, 179
 - of linear regression, 234
 - of maximum likelihood (*see* Maximum-likelihood estimate)
 - of percentiles, 218
 - of population mean, by confidence intervals, 208
 - by sample mean, 207
 - of population median, 216
 - of population percentage, by confidence intervals, 100
 - by sample percentage, 98
 - properties of, efficient, 99
- Estimation, properties of, unbiased, 99
 - of totals or aggregates, 214
 - trend, moving average in, 314
- Evaluating sources of information, 121
- Event, 31
- Events, independent, 33
 - mutually exclusive, 33
- Expected value, 35
- Experiment, bead-box, 4
 - random-number, 16
- Exponential distribution, 175
- Extrapolation with linear regression line, 237
- Extreme values in frequency distribution, 140
- F (variance ratio), 363
 - test of equality, of several population means, 365
 - of two population means, 363
- Factorial, 40
- Failures, bus-motor, theory of, 171
 - exponential distribution, 175
 - normal distribution, 171
- First differences, test for trend, 283
- Frame, 16
- Frequency distribution, 44, 135
 - arithmetic mean from, 154
 - cumulative, 155
 - extreme values, 140
 - fitting curves to, 167
 - graphing, choice of scales, 146
 - distortions due to scale breaks, 147
 - equal class intervals, 144
 - open-end intervals, 146
 - unequal class intervals, 145
 - median of, 155
 - mid-points, 136-139
 - mode of, 156
 - percentiles of, 155
 - quartile of, 155
 - relationships among mean, median, and mode, 157
 - rules for making, 138
 - skewness, 144
 - standard deviation of, 156
 - (*See also* Binomial distribution; Normal distribution)
- Frequency polygon, 45, 145

- Functional relationships, 224
- Fundamental probability set, 31
- Goodness of fit (*see* Statistical tests)
- Government statistics, uses of sampling
 - in gathering, 123
- Greek letters, 25
- Histogram, 30, 45, 144
- Horse-racing problem, solution to, 57
- Hypotheses, as guides to actions, 192
 - nonstatistical, 57
 - statistical, 57
 - tests of (*see* Statistical tests)
- Hypothesis, 6
- Independent events, 33
- Independent variable, 225
- Index numbers, 332
 - Consumer Price Index, 347
 - economic theory of, 355
 - Federal Reserve Board Index of Industrial Production, 351
 - price indexes, simple relatives, 333
 - weighted indexes, 337
 - production indexes, 341
 - sources of, 355
- Inductive behavior, 67
- Industrial quality control, acceptance
 - inspection for fraction defective, 92
- Interpolation with linear regression line, 237
- Interquartile range, 151
- Interviewing, 119
- Journals in statistics, 120
- Law of large numbers, 35
- Least-squares method, 250
- Lot tolerance percent defective (LTPD), 92
- Mail questionnaires, 119
- Mathematical probability, attributes of, 31
- Mathematical probability, continuous
 - case, 36
 - discrete case, 31
 - graphic model, 32
- Maximum-likelihood estimate, example, 99
 - of population mean, 207
 - of population proportion, 99
- Mean, arithmetic, from array, 150
 - of binomial distribution, 47
 - estimation of (*see* Estimation)
 - from frequency distribution, 154
 - testing hypotheses on (*see* Statistical tests)
- Median, from array, 148
 - confidence-interval estimates of, 216
 - of frequency distribution, 155
- Mid-point of frequency distribution intervals, 136–139
- Mode, from array, 150
 - of frequency distribution, 156
- Mortality tables, 35
- Moving average in trend estimation, 314
- Multiplication theorem of probabilities, 33
- Mutually exclusive events of probability theory, 33
- Mutually exclusive frequency distribution intervals, 136
- National Bureau of Economic Research, 316, 329
- Nonlinear relationships, 226
- Nonrandom samples, 128
- Normal distribution, 73
 - approximation to binomial distribution, 77
 - equation for, 78
 - examples of, 80
 - graphs of, 78
 - notation, 77
 - area and probability relationships, 75
 - bivariate, 252
 - equation of, 73
 - fitting to frequency distribution, 171
 - graphs of, 74
 - historical role of, 167
 - parameters of, 74
 - standard, 75
 - tables, use of, 76

- Normal distribution, testing goodness of fit**, 174
 transformation to Z , 75
 use of, in testing hypotheses (*see* Statistical tests)
 (*See also* Binomial distribution, Frequency distribution)
- Notation**, 25
 difference between population parameters and sample estimates, 46, 48
 list of symbols, 383-385
- Observation**, 9
- Operating-characteristic curve**, 60
 for different alternatives in population proportion, 63
 effect of sample size on, 64
 of horse-race test, 65
 ideal, 91, 195
 numerical computation, statistical test, of population mean, 192, 201, 204
 of population proportion, 62, 85
 use in determining sample size for testing hypotheses, of population mean, 193
 of population proportion, 88
- Percentiles**, from array, 149
 of frequency distribution, 155
- Phase test for cycle**, 284, 293, 322
- Population**, 9
 census of, age definition, 121
 describing with mean and standard deviation, 159
 estimating distribution without specifying form, 179
 (*See also* Confidence interval, Estimation)
- Power function** (*see* Operating-characteristic curve)
- Prediction**, intervals for regression estimates, 243
 use in, of probability, 34
 of relative frequencies, 35
- Probability**, addition theorem, 32, 33
 attributes of, 31
 classical notions about, 30
 event space, 31
- Probability, fundamental set**, 31
 introduction to, 29
 mathematical, 31
 model, for binomial distribution, 39
 for a die, 32
 scale of, 30
 theorems of, 33
 use of, in prediction, 34
- Producer's risk**, 92
- Purposive sample**, 129
- Quality control**, acceptance sampling for percentages, 92
- Quality level**, 92
- Quartile**, from array, 149
 deviation, 151
 of frequency distribution, 155
- Questionnaires**, 119
- Quota sample**, 129
- Random fluctuations in time series**, 264
- Random numbers**, experiment in use of, 16
 table of, 12, 13
 use of, 11
- Random sampling**, cluster, 125
 simple, 10
 stratification, 124
 systematic, 117, 123
- Random variable**, 20, 32
- Randomness in prediction from probabilities**, 34
- Range**, 151
- Rank correlation**, for measuring relationship between two variables, 253
 test for trend, 311
- References to books in statistics**, 367
- Region of acceptance and rejection**, 52
- Regression**, introduction to, 226
 line of, 226
 linear relationship, estimation of, 234
 interpretation of, 235
 method of least squares, 250
 multiple, 257
 prediction intervals, 243
 standard error of estimate, 238
 testing for universe relationship, 239
 universe model of, 230
- Regression fallacy**, 229

- Relationships between two variables,
 - cautions, on causation and association, 227
 - on common elements, 228
 - on implicit third factors, 228
 - on regression fallacy, 229
 examples of, 223
- Relative frequencies, use in prediction, 35
- Relative variation, 159
- Research, classroom investigation of
 - weights, 108
 - elementary principles of, 115
 - statistical methodology, 108
- Sample, 6, 10
 - method of selection, 10
 - nonrandom, 128
 - point, 51
 - purposive, 129
 - random, 10
 - space, 40, 51
- Sample size, effect on operating-characteristic curve, 64
 - in estimation, of population distribution, 181
 - of population mean, 212
 - of population proportion, 105
- in testing hypotheses on population mean, with known standard deviation, 193
 - with unknown standard deviation, 205
- in testing hypotheses on population proportion, 88
- Sampling, acceptance, industrial quality control, 92
 - in gathering government statistics, 123
 - random (*see* Random sampling)
 - with replacement, 42
 - without replacement, 42, 187
- Sampling distribution, concept of, 40
 - use of (*see* Statistical tests)
- Sanborn Maps for Los Angeles, 126
- Scatter diagram, 222
- Seasonal variation, adjustment for, 299
 - criterion in egg-production process, 292
 - in egg production, 289
 - errors in prediction, 302
- Seasonal variation, examples of, 267
 - indexes of egg production, 300
 - adjusted, 299
 - unadjusted, 298
 - uses of, 301
 - meaning of adjusted figures, 301
 - measurement of, 295
 - shifting, 303
 - statistical test for, 289
- Semilogarithmic graph paper, 268
 - measuring percentage changes, 271
 - rates of change, 270
 - scaling, 271
 - for seasonal-variation measurement, 295
- Skewness, 144
- Slutsky-Yule effect, 287, 325
- Sources of information, 118, 121
- Standard deviation, from array, 152
 - of binomial distribution, 49
 - computational formula, 153
 - of frequency distribution, 156
 - introduction to, 48
- Standard error of estimate, 238
- Statistic, 6
- Statistical inference (*see* Statistical tests)
- Statistical journals, 120
- Statistical significance, 91
- Statistical tests, arithmetic mean, 183-206
 - from arbitrary population, 202
 - difference between two population means, 361
 - for equality of several population means, 363
 - examples, 183
 - judging sample means, 196, 201
 - from normal population, with known standard deviation, 188
 - with unknown standard deviation, 198
- best unbiased test, 84
- for correlation coefficient, 250
- for cycle, phase test, 284, 322
- goodness of fit, chi-square, 169
 - exponential distribution, 175
 - normal distribution, 171
- as probability tests, 62
- proportions, 51-65, 73-95
 - difference between two sample, 357
 - equality of several sample, 359

- Statistical tests, proportions, judging
 sample results, 92
 large samples using normal approximation, 82
 small samples using tables, 63
 of rank correlation, 255
 rank test for trend, 311
 sample size (*see* Sample size)
 for seasonal variation, 289
 time series (*see* Time series)
 for trend, first difference test, 283
 rank test for, 311
 slope of linear trend line, 309
 for universe regression relationship, 239
- Statistics, function of, 2
 references to books in, 367
- Stratified random sampling, 124
- Summation, 26
- Symbols, glossary of, 383-385
- Systematic random sampling, 117, 123
- t*, distribution of, as sampling distribution of arithmetic mean, 199
 table, 374
 use of, 199
 in testing hypotheses on slope of regression line, 241
- Tables, of confidence limits for percentages, 102
 d.c.c., for estimating population distribution, 180
 mathematical, 27
 of percentage improvement in standard deviation, 249
 random numbers, 12, 13
 reference to binomial distribution, 42
 statistical, 369-382
- Tchebycheff's inequality, 160
- Tests of hypotheses in relation to everyday decisions, 67
 (*See also* Statistical tests)
- Time series, alternative hypotheses, 274
 analysis viewpoint, 273, 279
 Cowles Commission for Research in Economics, 330
- Time series, cycle, explanation of, 325
 percentage deviation from trend, 319
 phase test for, 322
 examples of, 262
 National Bureau of Economic Research, 329
 random forces in, 267
 random model of, 280
 random series, analysis of, 281
 seasonal variation (*see* Seasonal variation)
 statistical tests, comments on, 323
 for cycle, 284
 for seasonal variation, 289
 for trend, 283, 308
 systematic forces, 267
 trend, least squares, 306
 linear, test for, 309
 moving average, 314
 rank test for, 311
 separation from random forces, 326
 (*See also* Cyclical fluctuations; Seasonal variations; Statistical tests; Trend)
- Tolerance interval, 218
- Trend, examples of, 263
 separation from random forces, 326
 (*See also* Time series)
- Universe, 9
- Variable, of binomial distribution, 19
 continuous, 36
 discrete, 32
 random, 20, 32
- Variance, analysis (*see* Analysis of variance)
 from array, 152
 of binomial distribution, 49
 computational formula, 153
 of frequency distribution, 156
 introduction to, 48

