UNIVERSAL
LIBRARY

OU_164086

UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

AN INTRODUCTION TO

Statistical Analysis

C. H. RICHARDSON PH. D.

Professor of Mathematics, Bucknell University

REVISED EDITION

HARCOURT, BRACE AND COMPANY NEW YORK

COPYRIGHT, 1934, 1935, 1944, BY
HARCOURT, BRACE AND COMPANY, INC.

All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

[f · 10 · 49]

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

It is the aim of this book to present the fundamental notions of statistical analysis in such a manner that they can be comprehended by students who have had but little training in mathematics and yet in such a way that they can be studied to advantage even by those who have had considerable mathematics. To supplement the mathematical preparation of the former group we have intermittently interrupted the continuity of the statistical procedure by inserting sections on certain topics of advanced algebra and analytic geometry such as sums and summations, some properties of the straight line, permutations, combinations, and the elementary theory of probability.

Many of the basic notions of statistical analysis are expressed by formulas, the derivations of which have been assumed — altogether too frequently — to be hidden in a maze of higher mathematics. For a number of years we have encountered a growing opinion in some circles — betrayed by clever innuendo and subtle insinuation when not definitely expressed — that *how to use* a formula and *what it means* are the primary desiderata in statistical analysis and that *how it is derived* and *what are its limitations* are of secondary importance. It is our conviction that a reader will not comprehend fully what a formula means and what are its limitations unless he knows whence it comes and what are the assumptions underlying its development.

Since the mathematical attainments necessary for an understanding of the development of many of our basic formulas include no more than a knowledge of algebra through the binomial theorem, the theory and use of logarithms, and the progressions — topics that are included in a well organized course of secondary algebra — we have included many derivations that come within the grasp of the ordinary student. The limited preparation in mathematics that we assume on the part of our readers requires that difficult derivations be generally omitted.

While the theory of statistical analysis is not easy, yet the difficulties are, in the main, due to the *newness* rather than to the *abstruseness* of the notions encountered. The concepts will, therefore, become more meaningful and less terrifying if the student will be required to solve many of the numerous exercises that have been provided in the text.

Statistical analysis boils down ultimately to numerical results: the methods and processes used in obtaining them and the methods and means for estimating their reliability. The earlier chapters of the book are concerned mainly with the methods, processes, and forms used in obtaining numerical results and the later chapters deal with estimating their reliability.

The plan used in the development of the text may be briefly described as follows: Each topic is introduced with a brief statement of "what it is all about." Then follows a brief statement of the underlying theory of the topic under consideration which leads directly and simply to a development of the necessary formulas and processes. The reader is then shown how to use the formulas and processes to obtain the desired numerical results. Finally, the limitations of the formulas and processes and the significance and the reliability of the computed results are given due emphasis. Thus a student learns *why* a formula is applied, *whence* it is derived, *how* it is used and *what* are its limitations; he learns not only how to obtain the numerical results but also how to measure their reliability.

The method of treatment of all the topics is decidedly elementary. The graphical method has been widely employed and the explanations have been purposely detailed in order that the book may be more readily understood. Since the book undertakes to develop skills in deriving statistical results as well as to assist in understanding their significance, numerous exercises have been placed at strategic points in the text. This feature of solving exercises after a major topic has been considered adds to the teachableness of the subject, facilitates an understanding of the principles, and aids the student in acquiring the useful skills for statistical computations and interpretation.

In general, the exercises are based upon actual rather than imaginary data in order that the study may proceed, if possible, with real life situations. The alert teacher can improvise "homemade"

exercises as he needs them. Throughout the text, it is supposed that a computing machine is at the disposal of the student; nevertheless many of the exercises can be done satisfactorily with a slide rule or a table of logarithms, powers, and roots.

No attempt has been made to make this text an exhaustive treatise on statistical analysis. Many topics, such as multiple correlation and frequency curves, have been studiously omitted. We have tried to keep in mind that we are writing an Introduction that would include the minimum essentials, at the same time hoping that this Introduction might inspire the reader to continue his study into the more advanced fields.

I wish at this time to renew my thanks to Professor James W. Glover and Professor Harry C. Carver of the University of Michigan for their most generous aid to me when I was under their instruction. I also hasten to express my gratitude to Professor C. H. Forsyth of Dartmouth College and Professor Ralph W. Tyler of Ohio State University, who critically read the manuscript and made numerous helpful suggestions.

For any errors, I alone am responsible. Although the text has been checked painstakingly, it is not to be hoped that a publication of this character will appear without some errors creeping in. For the notification of such errors I shall be most grateful.

PREFACE TO THE ENLARGED EDITION

It has been an unexpected gratification to the author and the publishers alike to find that an enlarged edition of the book is called for so soon after its publication. The only criticism of consequence that has been made of the book was due to our omission of *Index Numbers*. After sampling the opinion of many teachers, it was felt desirable to add a chapter devoted to that topic. The opportunity has been taken to recast certain paragraphs, and such errors as were noted have been corrected. To all who have assisted me with their suggestions or by directing my attention to errors, I wish to express my sincere gratitude.

C. H. RICHARDSON

Lewisburg, Pennsylvania
April 30, 1935

PREFACE TO THE REVISED EDITION

ABOUT ten years have elapsed since the first edition of this book was published and twelve to fifteen years since the material for the first edition was collected and prepared. During this time a tremendous appreciation of and respect for statistical techniques have developed. A considerable extension of the use of statistical techniques in business, in public administration, and in the social sciences is very much in evidence. Research workers in biology, in education, in psychology, in sociology, in agriculture, lean more heavily on statistical techniques than ever before. And, with the passing of time, there has come a demand for more than primer notions: a deeper understanding of basic ideas is mandatory. For example, it no longer suffices merely to compute a statistical constant or statistic: one must evaluate it, determine its worth.

Notable gains have been made during the past decade in the development of new and in the improvement of old techniques. Enriching the old areas and exploring new ones have challenged some of the best minds of the world. Creative minds in pure as well as in applied mathematics have attacked fundamental problems so vigorously that now the literature of the field is colossal.

Having been alert to these new developments and improvements, it is our wish to incorporate those that are appropriate into this new edition. In doing this we have sought to retain the main features of the first edition since the plan of its construction has met the approval of a wide audience of teachers and students. The two objectives, statistical description and statistical evaluation, have been kept in mind. In this edition we are not giving less attention to statistical description but we have been careful to give more emphasis to statistical evaluation and statistical induction. It is essential that the student be able to compute a statistic: it is just as essential that he know what he has when he has it, and to know, in terms of probability, what he can do with it. Consequently, we have made a great effort to make the techniques and computations meaningful. At the risk of being prolix, we have given rather full verbal discussions of important matters; our illustrative examples are numerous and their solutions detailed.

Along with the progress that has been made in improving old techniques and in establishing new ones, there has come an enlarged opportunity for the study of statistical analysis by more and more students of our colleges. Due to its wide application a knowledge of statistical methods is now a "must" in a program for a liberal education. Of course this growth has been influenced greatly by the desire of thinkers to replace as far as possible the subjective elements of their fields by objective procedures. On the whole, this substitution of objectivity for mere opinion has been healthful.

The thirteen chapters of this edition fall into two divisions, each division associated with a definite objective. The first ten chapters emphasize statistical description whereas the last three chapters emphasize statistical induction. A study of the entire book is consequently necessary if one would seek an understanding of what is now considered to be the essentials of elementary statistics, statistical description and statistical induction.

One new chapter, Multiple Correlation, has been added to the present edition. New sections pertaining to other topics have been inserted. Many sections have been completely rewritten, others greatly amplified. The numerical exercises have been multiplied and the algebra of statistics has been extended. The book has therefore not only been revised but greatly enlarged, thus providing a wider selection of topics for the teachers.

Many friends and teachers have rendered invaluable assistance with their sympathetic suggestions for the improvement of the book. These suggestions have come to me over the years. I wish that I might mention here each contributor personally but the list is too long. However, I do want to again express my thanks to my friend and former teacher, Professor A. R. Crathorne of the University of Illinois, whose generous and tactful suggestions have been invaluable. Also, I want to express my thanks to my colleague, Mr. Paul Benson, who has assisted with the proof and has made numerous helpful suggestions. Of course for any errors, I alone am responsible.

In this edition I am including ANSWERS to many of the exercises. Obviously, it is too much to expect that all of them are correct. For the notification of any errors I shall be very grateful.

Lewisburg, Pennsylvania
September 1, 1943

C. H. RICHARDSON

Contents

1. INTRODUCTION

SECTION		PAGE
1.	THE MEANING AND IMPORTANCE OF STATISTICS	1
2.	MATHEMATICAL AND NON-MATHEMATICAL ASPECTS OF STATISTICS	4
3.	VARIABLES AND FUNCTIONS	5
4.	SUMS AND SUMMATIONS	7
5.	REMARKS ON MEASUREMENT	14
6.	DECIMAL ACCURACY	14
7.	SIGNIFICANT FIGURES	15
8.	ROUNDING OFF NUMBERS	16
9.	ERRORS IN CALCULATIONS	16
10.	THE PROPAGATION OF ERRORS	18

2. TABULAR AND GRAPHICAL REPRESENTATION: FREQUENCY DISTRIBUTIONS

11.	INTRODUCTION	23
12.	CLASSIFICATION OF THE DATA	23
13.	THE CHOICE OF THE CLASS INTERVAL	30
14.	CLASS LIMITS	30
15.	GRAPHICAL REPRESENTATION	37
16.	GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS	37
17.	GRAPHICAL REPRESENTATION OF TEMPORAL DISTRIBUTIONS	43
18.	CUMULATIVE DISTRIBUTIONS AND CURVES	48
19.	TYPES OF FREQUENCY CURVES	51
20.	SUGGESTIONS FOR TABULAR AND GRAPHICAL PRESENTATION	53

3. MEASURES OF CENTRAL TENDENCY

21.	INTRODUCTION	59
22.	THE ARITHMETIC MEAN, M_X	60

SECTION PAGE
23. THE ARITHMETIC MEAN AS A MOMENT 62
24. A SHORT METHOD FOR COMPUTING THE ARITHMETIC MEAN 71
25. THE MEDIAN, M_d 76
26. THE MODE, M_o 80
27. THE GEOMETRIC MEAN, M_g 87
28. THE HARMONIC MEAN, M_h 92
29. DISCUSSION AND CRITICISM OF THE MEASURES OF CENTRAL TENDENCY 98
A. THE ARITHMETIC MEAN 99
B. THE MEDIAN 100
C. THE MODE 100
D. THE GEOMETRIC MEAN 101

4. MEASUREMENT OF DISPERSION

30. THE INADEQUACY OF MEASURES OF CENTRAL TENDENCY 111
31. THE RANGE 114
32. THE QUARTILE DEVIATION 115
33. THE MEAN DEVIATION 120
34. THE STANDARD DEVIATION 125
35. THE NORMAL CURVE 134
36. THE PROBABLE ERROR 137
37. THE SIGNIFICANCE OF THE MEAN AND THE STANDARD DEVIATION 141

5. SKEWNESS: EXCESS: MOMENTS

38. INTRODUCTION 150
39. THE MEANING OF SKEWNESS 150
40. THE MEASUREMENT OF SKEWNESS 151
41. EXCESS OR KURTOSIS 158
42. THE UNADJUSTED MOMENTS OF A DISTRIBUTION 159
43. THE ADJUSTED MOMENTS: SHEPPARD'S CORRECTIONS 163
44. COMPUTATION OF THE MOMENTS 164
45. RETROSPECT AND PROSPECT 169

6. INDEX NUMBERS

46. INTRODUCTION 174
47. RELATIVES 174

SECTION PAGE
48. DEFINITIONS AND NOTATION 177
49. UNWEIGHTED INDEX NUMBERS 178
50. WEIGHTING 184
51. WEIGHTED AGGREGATES 185
52. WEIGHTED AVERAGES OF RELATIVES 188
A. THE WEIGHTED ARITHMETIC MEAN OF RELATIVES 188
B. WEIGHTED GEOMETRIC MEAN OF RELATIVES 191
53. SUMMARY AND EXTENSION 194
54. BIAS 197
55. FISHER'S IDEAL INDEX 198
CONCLUSION 200

7. LINEAR TRENDS

56. INTRODUCTION 203
57. SOME CHARACTERISTIC PROPERTIES OF A STRAIGHT LINE 204
58. THE EQUATION OF A STRAIGHT LINE 206
59. FITTING A STRAIGHT LINE TO OBSERVED DATA 210
A. THE METHOD OF LEAST SQUARES 210
B. THE METHOD OF MOMENTS 219
60. THE STRAIGHT LINE WITH THE ORIGIN AT THE CENTROIDAL POINT 221
61. FITTING A STRAIGHT LINE TO A TIME SERIES 226

8. SIMPLE CORRELATION

62. MEASURES OF CONCENTRATION OF POINTS ABOUT THE LINE OF REGRESSION 232
63. THE BRAVAIS-PEARSON COEFFICIENT OF CORRELATION 237
64. COMPUTATION OF r FOR UNGROUPED DATA 241
65. OTHER FORMS OF r 244
66. SUMMARY AND EXTENSION OF THE THEORY OF CORRELATION 247
67. COMPUTATION OF r FOR GROUPED DATA 253
68. CORRELATION BY RANKS 263
69. CORRELATION AND CAUSATION 267

9. MULTIPLE CORRELATION

70. PRELIMINARY EXPLANATION 277
71. THE CASE OF THREE VARIABLES 278
72. CONTINUATION OF THREE VARIABLES 282

SECTION PAGE

73. COEFFICIENT OF MULTIPLE CORRELATION FOR THREE VARIABLES 286
74. DETERMINANTS 288
A. DETERMINANTS OF THE SECOND ORDER 288
B. DETERMINANTS OF THE THIRD ORDER 290
C. DETERMINANTS OF ANY ORDER 292
75. APPLICATIONS OF DETERMINANTS 293
THREE VARIABLES
76. PARTIAL CORRELATION 295
77. THE CASE OF FOUR VARIABLES 297
78. THE CASE OF n VARIABLES 303

10. NONLINEAR TRENDS: CURVE-FITTING

79. INTRODUCTION 306
80. THE PROCESS OF DIFFERENCING 307
81. FITTING A STRAIGHT LINE TO OBSERVED DATA 311
A. THE METHOD OF SELECTED POINTS 311
B. THE METHOD OF AVERAGES 313
C. THE METHOD OF LEAST SQUARES 314
82. THE EXPONENTIAL FUNCTION: $Y = ab^X$ 316
83. THE POWER FUNCTION: $Y = aX^b$ 323
84. THE PARABOLA: $Y = aX^2 + bX + c$ 330
85. OTHER USEFUL CURVES 333
A. THE HYPERBOLA: $Y = a + \frac{b}{X}$ 333
B. THE HYPERBOLA: $Y = \frac{X}{a + bX}$ 334
C. THE MODIFIED EXPONENTIAL: $Y = a + bc^X$ 334
D. THE MODIFIED POWER FUNCTION: $Y = c + aX^b$ 337
86. LIMITATIONS OF EMPIRICAL EQUATIONS 338
87. GRAPHICAL METHODS IN TREND ANALYSIS 340
A. ARITHMETIC PAPER 341
B. SEMI-LOGARITHMIC PAPER 342
C. LOGARITHMIC PAPER 346
88. GOODNESS OF FIT OF CURVES TO OBSERVED DATA: NONLINEAR CORRELATION 354
A. GOODNESS OF FIT 354
B. NONLINEAR CORRELATION 355

11. PERMUTATIONS, COMBINATIONS, AND PROBABILITY

SECTION		PAGE
89.	INTRODUCTION	362
90.	PERMUTATIONS	364
91.	NUMBER OF PERMUTATIONS	366
92.	COMBINATIONS	367
93.	RELATIVE FREQUENCY: EMPIRICAL PROBABILITY	370
94.	THEORETICAL RELATIVE FREQUENCY: A PRIORI PROBABILITY	372
95.	EXPECTATION	374
96.	SOME ELEMENTARY THEOREMS	374
	A. MUTUALLY EXCLUSIVE EVENTS	374
	B. INDEPENDENT EVENTS	375
	C. DEPENDENT EVENTS	376
97.	REPEATED TRIALS	377

12. THE POINT BINOMIAL AND THE NORMAL CURVE

98.	INTRODUCTION	383
99.	CHARACTERISTICS OF THE POINT BINOMIAL	384
	A. THE MODE	385
	B. THE MEAN, THE DISPERSION, THE SKEWNESS	387
100.	THE POINT BINOMIAL APPLIED TO FREQUENCY DISTRIBUTIONS	391
101.	THE NORMAL CURVE: INTRODUCTORY REMARKS	395
102.	DERIVATION OF THE EQUATION TO THE NORMAL CURVE	397
103.	SOME PROPERTIES OF $\phi(t)$	401
104.	ILLUSTRATIVE EXAMPLES	405
105.	ON THE SIGNIFICANCE OF RESULTS	409
106.	GRADUATION OF A DISTRIBUTION BY THE NORMAL CURVE	413
	A. GRADUATION BY ORDINATES	414
	B. GRADUATION BY AREAS	415

13. THE THEORY OF SAMPLING: MEASURES OF RELIABILITY

107.	INTRODUCTION	419
108.	THE PROBLEM OF THIS CHAPTER	420
109.	THE STANDARD DEVIATION IN CLASS FREQUENCIES	422
110.	AN EXPERIMENT IN SAMPLING	425

SECTION PAGE

111. THE DISTRIBUTION OF MEANS 429
A. THE MEAN OF THE MEANS 429
B. THE STANDARD DEVIATION OF THE MEANS 430
C. THE PROBABLE ERROR OF THE MEAN 432
D. THE SKEWNESS AND EXCESS OF THE DISTRIBUTION OF MEANS 439
112. THE RELIABILITY OF THE STANDARD DEVIATION 442
113. THE RELIABILITY OF THE DIFFERENCE BETWEEN TWO MEASURES 443
114. SMALL SAMPLES 450
115. CONCLUDING REMARKS ON SAMPLING 456
116. SUMMARY OF RELIABILITY FORMULAS 456

APPENDICES

A. SELECTED BOOKS FOR SUPPLEMENTARY READING 465
B. AREAS AND ORDINATES OF THE NORMAL CURVE 467
C. FOUR-PLACE LOGARITHMS AND ANTILOGARITHMS 471
ANSWERS TO EXERCISES 475
INDEX 495

Chapter 1

INTRODUCTION

1. THE MEANING AND IMPORTANCE OF STATISTICS

During the last half-century,[1] the thinking world seems to have awakened to an unusually deep appreciation of and respect for numerical facts. Even the untrained mind has confidence in a conclusion stated in numerical language and supported by numerical facts. Whether the affairs are of state or laboratory, we must have observed that quantitative facts concerning them are collected in boundless profusion. The social and biological sciences, which were qualitative a few decades ago, have now become largely quantitative. Masses of numerical data are collected by individuals, by corporations, by governments.

These masses of data, numerical facts, measurements, which are generally known as *statistics*, may more precisely be called *statistical data*. The special methods used in the explanation and the elucidation of quantitative data may be fittingly called *statistical methods*. The analysis which is peculiar to and forms the basis of our method we call *statistical analysis*.

The word *statistics* is generally used indiscriminately in two different senses: on the one hand to refer to statistical material, the group of numerical data; and on the other hand, to statistical analysis, which includes those technical operations that have to do with the explanation and the interpretation of the numerical data.

As we shall use the term, *statistics is the science which deals with the collection, the organization, the analysis, and the interpretation of masses of numerical facts.* It will be noted that this definition is broader in scope than that given by Yule and Kendall. They say,[2]

[1] This is not meant to imply that statistics is a new subject. See the Book of Numbers in the Bible. See also H. M. Walker, *Studies in the History of Statistical Method*, 1929.

[2] G. U. Yule and M. G. Kendall, *Introduction to the Theory of Statistics*, 12th ed., p. 3.

By *statistics* we mean quantitative data affected to a marked extent by a multiplicity of causes.

By *statistical methods* we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

By *theory of statistics* we mean the exposition of statistical methods.

Statistical methods are fundamentally the same whether employed in the analysis of physical phenomena, the study of educational measurements, the records of biological experiment, or the analysis of quantitative material in economics. All such data are "affected to a marked extent by a multiplicity of causes." True, the physicist, the chemist, the biologist, and possibly the psychologist attempt to eliminate many disturbing causes and to concentrate their attention upon one or two most powerful influences affecting their phenomena, yet many disturbances are always present. However, the same general procedure is followed by the educationist and the economist. Generally, it is one of continued summarization.

We shall therefore feel free to apply our analysis to numerical data whether they come from the astronomer or the agriculturist, the physicist or the economist, the biologist or the chemist. Wherever there is a mass of numerical data that admits of explanation, we shall consider its analysis our field of endeavor.

The fact that the human mind is incapable of comprehending a large number of impressions at one time is generally recognized. A mass of numerical data is an appropriate illustration. To grasp the meaning of a mass of numerical data we must reduce its bulk. The *organization of the data* is the first step in the summarizing process. It is a phrase that is used to describe the process of arranging the data in a compact form that facilitates computations and comparison. When they are so arranged, ordered, classified, — *organized*, — they are then in a form suitable for the analysis.

The process of abstracting the *significant* facts contained in a mass of numerical data and making clear and concise statements about the derived results constitutes a *statistical analysis* of the data. A statistical analysis, therefore, enables us to express the *relevant* information contained in the mass of data by means of a few numerical values known as *statistical constants* or *parameters*, each constant describing an important property of the mass of data. It is thus the purpose of statistical analysis to give a summarized and compre-

hensible numerical description of masses of numerical data. This is effected by computing a few constants pertaining to the data and understanding their meaning.

Toward this numerical description of the mass of data we may adopt two points of view. We may view the description of the given mass as an end in itself, or we may view it as a basis for generalization, as a basis for making estimates of the character measured pertaining to a larger group. The smaller group that is analyzed we call a *sample*, the larger group about which we make estimates we call the *parent population* or *universe*. The *interpretation of the data* is a phrase used when we adopt the larger point of view and make estimates, form judgments, or draw inferences of the universe from a study of the statistical properties of the sample.

Let us consider, for example, the scores of 100 freshmen at Bucknell University on a standardized Algebra test on which the highest attainable score was 50. Here are the scores in Table 1.

TABLE 1. SCORES OF 100 FRESHMEN ON AN ALGEBRA TEST

43	18	25	18	39	44	19	20	20	26
40	45	38	25	13	14	27	41	42	17
34	31	32	27	33	37	25	26	32	25
33	34	35	46	29	24	31	34	35	24
28	30	41	32	29	28	30	31	30	31
28	31	30	34	40	29	46	30	30	47
31	35	36	29	26	32	36	35	36	37
32	23	22	29	33	37	33	27	24	36
23	42	29	37	19	23	44	41	45	39
21	21	42	22	28	38	15	16	17	28

If we desire such information regarding the above scores as is found in the answers to the following questions, we must look through the entire table.

1. How many students obtained scores greater than 43?
2. How many obtained scores greater than 22 and less than 43?
3. How many obtained scores less than 23?
4. What is the lower boundary of the upper 20% of the scores?
5. What is the lower boundary of the upper 40% of the scores?

Such questions may be readily answered if we *organize* the data by arranging the scores into classes, as we have done in Table 2.

TABLE 2. ORGANIZATION OF THE DATA OF TABLE 1

Class	*Frequency, or the number of scores in the given class*
42.5–47.5	8
37.5–42.5	12
32.5–37.5	20
27.5–32.5	28
22.5–27.5	16
17.5–22.5	10
12.5–17.5	6
Total	100

The new table is called a *frequency table* for it gives the frequency (the number of scores) in the respective intervals. Evidently the organization of the data presents them in a form that is more suitable for statistical purposes than the disorganized form of Table 1 does.

EXERCISE. Make a list of several facts that you can immediately discover from Table 2. Answer the questions that we have listed on the preceding page.

It must not be supposed that the answers to the above questions constitute the analysis of the data. The analysis is contained in the following constants that we shall later learn to compute and interpret.

$M = 30.7$ $\qquad$ $\sigma = 7.85$

$M_d = 30.71$ $\qquad$ $Q_1 = 25.3$

$M_o = 30.5$ $\qquad$ $Q_3 = 36.25$

each expressed in the given unit of measure.

To undertake an *interpretation* at this time would take us too far afield.

2. MATHEMATICAL AND NON-MATHEMATICAL ASPECTS OF STATISTICS

As has been indicated in our definition, the steps involved in the solution of a statistical problem may be summarized as follows:

1. The collection of the data
2. The organization of the data
3. The analysis of the data
4. The interpretation and criticism of the results

The collection of the data and their organization are largely non-mathematical operations. In regard to this Bowley says, "Common sense is the chief requisite and experience the chief teacher." [1] However, we shall refer to these items in later chapters. The elementary analysis of the data involves in general no so-called higher mathematics. It is well to understand algebraic averages and some of the elementary principles of the algebra of summations. These will be considered in another section of this chapter. An understanding of the calculus is always helpful and at certain times highly desirable, but this preparation is not necessary for the elementary course. The student who desires a knowledge of the more refined methods of statistical analysis will find an understanding of the calculus and the theory of probability indispensable.[2] For the interpretation and the criticism of the results, one cannot know too much. Bowley says in this regard:

> For criticism of estimates and interpretation of results it is necessary to use formulae of more advanced mathematics, and it is obviously expedient to understand the methods by which these formulae are obtained to ensure their intelligent use.[3]

Since this book is essentially one of the methods of elementary analysis of statistical data, all technical questions that require a considerable knowledge of advanced mathematics will be omitted.

3. VARIABLES AND FUNCTIONS

A common property of any character with which statistics is concerned is that of variation or change. The grades of a class in geometry, the scores in an examination, the heights of a group, the number of petals on a group of buttercups, the production of wheat from year to year, even a group of measurements of the length of a room — all these show variation. In statistics the magnitudes of

[1] A. L. Bowley, *Elements of Statistics*, p. 14.
[2] E. V. Huntington, "Mathematics and Statistics," *American Mathematical Monthly*, December, 1919.
[3] Bowley, *op. cit.*, p. 14.

the character measured are frequently called *variates*. A variate, then, is a very special and specific use of the broader term *variable*, which signifies any quantity that changes in magnitude.

TABLE 3. GRADES OF A CLASS IN ALGEBRA

Grade X	*Frequency, or the number of students receiving the grade* $f(X)$
65	3
75	14
85	10
95	3
(*Total*)	30

In Table 3 there are two variables — the grades, X, and the frequencies, $f(X)$ — but the variates are the magnitudes of the grades.

We shall find it necessary and convenient to recognize two distinct classes of variates, continuous, and discrete or discontinuous.

A *continuous variate* is one whose magnitudes may differ by infinitesimal amounts between certain limits: for example, the weight of a man, the temperature of a place, the length of a bean pod, the height of a plant.

A *discrete or discontinuous variate* is one whose value must be described in integers; for example, the number of pupils in a class, the number of kernels on an ear of corn, the number of seeds in an apple, the number of culms on an oat plant. A discrete variate is sometimes called an *integral variate*.

As in other fields of mathematics, it will be convenient to recall another classification of variables, namely, the independent and the dependent. The *independent variable* is the one to which we assign values at pleasure, whereas the *dependent variable* is the one whose value depends upon that assigned to the independent variable.

In Table 3 the frequency of a given grade depends upon the grade we have assigned at pleasure. The frequency is, therefore, the dependent variable. The dependent variable is frequently called a *function of the independent variable* or argument. The independent variables will be represented in this text by X, x', x, or t; the functional, or dependent variable by y, $f(x)$, or $f(t)$, etc. We shall say

that y *or* $f(x)$ *is a function of* x *if* y *is dependent upon* x *and if to every value of* x *there corresponds a value of* y *or* $f(x)$. It will not be necessary to describe this correspondence by means of an equation.

4. SUMS AND SUMMATIONS

Statistics may be roughly defined as the study of averages. Since nearly all averages involve the evaluation of certain sums, it will be well at this point to acquaint ourselves with an abbreviated notation for sums and develop some useful formulas that will aid us in quickly evaluating later sums. We shall discover that a facility with this new notation will quicken our understanding of the later chapters.

In elementary algebra if we add a set of letters $x_1, x_2, \ldots, x_n$ we indicate the sum by:

$$x_1 + x_2 + x_3 + \cdots + x_n$$

In more advanced mathematics we would designate this sum compactly as $\sum_{i=1}^{n} x_i$. Thus:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

This should be read "sigma of x sub i (or summation of x sub i) when i assumes all integral values from 1 to n inclusive."

The Greek capital letter Σ (sigma) placed before a term signifies the sum of all terms of which that term is the general type.

Thus:

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \sum_{1}^{n} x^2$$

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \sum_{1}^{n} \frac{1}{x}$$

$$2^3 + 3^3 + 4^3 + 5^3 + 6^3 = \sum_{2}^{6} x^3$$

$$\log 5 + \log 7 + \log 9 + \cdots + \log (2n - 3) = \sum_{4}^{n} \log (2x - 3)$$

and in general

$$f(1) + f(2) + f(3) + \cdots + f(n) = \sum_{x=1}^{n} f(x) \qquad (1)$$

The values below and above the summation symbol Σ which are the initial and final values of the independent variable are the *limits* of the summation. The limits of the summation may have any values and the independent variable may change by other amounts than unity. Slight changes in the notation make these extensions possible. Thus:

$$x_5 + x_{10} + x_{15} + \cdots + x_{100} = \sum_{i=5,\,10,\ldots}^{100} x_i = \sum_{i=1}^{20} x_{5i}$$

$$f(65) + f(75) + f(85) + f(95) = \sum_{65,\,75,\ldots}^{95} f(x)$$

$$5f(5) + 10f(10) + 15f(15) + \cdots + 75f(75) = \sum_{1}^{15} 5xf(5x)$$

$$f(a) + f(a + b) + f(a + 2b) + \cdots + f(a + nb) = \sum_{r=0}^{x=n} f(a + xb)$$

If the quantity under the summation Σ does not contain a variable, all the terms are equal. As examples we have:

$$\Sigma 1 = 1 + 1 + 1 + 1 + \cdots + 1 = N$$
$$\Sigma c = c + c + c + c + \cdots + c = Nc$$

where N is the number of observations or measurements.

It frequently happens that there is no necessity for writing the independent variable or the limits of the summation below and above the summation symbol. When the context tells us what is meant, we shall resort to this plan. Thus, if the lower or upper limit is omitted, it is assumed to be 1 or n respectively.

EXERCISES

Write the series that are represented by the following symbols:

1. $\sum_{1}^{n} \frac{1}{x^2}$ **2.** $\sum_{1}^{n} 2^x$ **3.** $\sum_{1}^{n} (x - 3)$ **4.** $\sum_{x=1}^{10} {}_{10}C_x$

5. $\sum_{x=1}^{n} xa^x$ **6.** $\sum_{x=1}^{10} x\,{}_{10}C_x$ **7.** $\sum_{10,\,20,\ldots}^{100} xf(x)$ **8.** $\sum_{5,\,10,\ldots}^{80} x^2 f(x)$

Write in the abbreviated form using Σ.

9. $1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \cdots + n(n + 1)$

10. $(X_1 - \bar{M})^2 + (X_2 - \bar{M})^2 + (X_3 - \bar{M})^2 + \cdots + (X_n - \bar{M})^2$

We will now consider two useful and important theorems.

Theorem I. *The Σ (sigma) of an algebraic sum of several functions is equal to the algebraical sum of the sigmas of the several functions.* Symbolically, we state that:

$$\Sigma[f(x) \pm F(x) \pm w(x) \pm \text{etc.}] = \Sigma f(x) \pm \Sigma F(x) \pm \Sigma w(x) \pm \text{etc.}$$

Theorem II. *The Σ (sigma) of a constant times a function is equal to the constant times the sigma of the function.* Symbolically, we state that:

$$\Sigma cf(x) = c\Sigma f(x)$$

This means of course that the constant factor, c, may be placed to the left or to the right of the Σ at pleasure.

Proof of Theorem I. (for two functions)

By the definition (1):

$$\begin{aligned}&\Sigma[f(x) \pm F(x)]\\ &= f(1) \pm F(1) + f(2) \pm F(2) + \cdots + f(n) \pm F(n)\\ &= [f(1) + f(2) + \cdots + f(n)] \pm [F(1) + F(2) + \cdots + F(n)]\\ &= \Sigma f(x) \pm \Sigma F(x)\end{aligned}$$

The proof is easily extended to any number of functions.

The proof of Theorem II will be left as an exercise for the student.

Example. From the identity

$$x^2 - (x-1)^2 = 2x - 1$$

we shall prove:

$$\sum_1^n x = \frac{n(n+1)}{2}$$

From the preceding definitions and theorems we have

$$\Sigma[x^2 - (x-1)^2] = 2\Sigma x - \Sigma 1 = 2\Sigma x - n$$

and this means, by (1):

$$\left.\begin{aligned}&1^2 - 0^2\\ &+ 2^2 - 1^2\\ &+ 3^2 - 2^2\\ &\cdots\cdots\cdots\\ &\cdots\cdots\cdots\\ &+ n^2 - (n-1)^2\end{aligned}\right\} = 2\Sigma x - n$$

or

$$n^2 = 2\Sigma x - n$$

Hence, we obtain:

$$\Sigma x = \frac{n(n+1)}{2}$$

EXERCISES

1. Write out in full what is meant by $\Sigma x = \frac{n(n+1)}{2}$. State in words.

2. Using the identity

$$x^3 - (x-1)^3 = 3x^2 - 3x + 1$$

prove that:

$$\sum_1^n x^2 = \frac{n(n+1)(2n+1)}{6}$$

3. Prove:

$$\sum_1^n (2x-1) = n^2.$$

4. Apply the result of Number 3 to find the sums:

(1) $11 + 13 + 15 + \cdots + 87 = \sum_1^{44}(2x-1) - \sum_1^{5}(2x-1)$

(2) $127 + 129 + 131 + \cdots + 195$

The above definitions, theorems, and exercises are concerned with the abstract algebra of summation. The numbers that have appeared as illustrations enjoy a degree of regularity not found in observed measurements. Thus, such series as: $1, 2, 3, \ldots, 25$; $1^2, 2^2, 3^2, \ldots, 16^2$ are not often found in actual measurements. So this abstract algebra is apparently not valuable in dealing with observed data. Actually, we shall find this abstract algebra of summation very helpful in developing statistical theory and frequently helpful in dealing with real measurements.

The numbers we meet in numerical problems are real measurements or scores that come from actual observation and they do not generally proceed with regularity. We should understand thoroughly how the algebra of summation applies to such measurements. Ten men in a class in Statistical Analysis gave their weights: 128, 131, 137, 143, 144, 146, 147, 149, 155, 170 pounds. Obviously these numbers do not proceed from the small to large values with the regularity of the numbers: $1^3, 2^3, 3^3, \ldots, 12^3$. However, we can apply our Σ notation to such irregular series as the weight data.

Let us arrange our data as in the adjacent vertical columns where we use the upper case X, capital X, to indicate a measurement of weight. The first measurement we indicate by X_1, the second by X_2, and so on. We are then able to express the sum of the weights by the

summation sign, Σ. Out of curiosity we found the average weight of the ten men.

Weight X
$X_1 = 128$
$X_2 = 131$
$X_3 = 137$
$X_4 = 143$
$X_5 = 144$
$X_6 = 146$
$X_7 = 147$
$X_8 = 149$
$X_9 = 155$
$X_{10} = 170$
$\sum_{i=1}^{10} X_i = 1450$

$$\text{Average weight} = \frac{1450}{10} = 145 \text{ lbs.}$$

In this case we note that it is the subscript i that varies, and with the weight of each man is associated a subscript. It is generally not necessary to indicate so precisely the indexes of the summation. It suffices to know that X refers to the characteristic measured, weight of a man, and ΣX represents the sum of the weights of the men. Consequently we could write

$$\text{Sum of the weights} = \Sigma X = 1450 \text{ pounds}$$

and not be misunderstood.

To save labor in statistical computations we find it convenient to effect simple transformations upon the variates. Thus, for the weight data we can work with smaller numbers if we refer our weights to some conveniently chosen number, say 100, instead of to 0. Since X represents the measurements referred to zero as origin, we must choose a new letter, say U, to represent

TABLE 4. WEIGHTS OF 10 MEN

Weight X	$U = X - 100$
$X_1 = 128$	$U_1 = 28 = X_1 - 100$
$X_2 = 131$	$U_2 = 31 = X_2 - 100$
$X_3 = 137$	$U_3 = 37 = X_3 - 100$
$X_4 = 143$	$U_4 = 43 = X_4 - 100$
$X_5 = 144$	$U_5 = 44 = X_5 - 100$
$X_6 = 146$	$U_6 = 46 = X_6 - 100$
$X_7 = 147$	$U_7 = 47 = X_7 - 100$
$X_8 = 149$	$U_8 = 49 = X_8 - 100$
$X_9 = 155$	$U_9 = 55 = X_9 - 100$
$X_{10} = 170$	$U_{10} = 70 = X_{10} - 100$

Adding, $\Sigma U = 450 = \Sigma X - 10(100)$

$$\Sigma X = 10(100) + 450$$

$$\text{Average weight} = \frac{\Sigma X}{10} = 100 + 45 = 145 \text{ lbs.}$$

the measurements referred to 100 as origin. We pose the question: Can we find ΣX by finding and using ΣU?

The first weight whose X is 128 will have a U of 28; the second weight has $X = 131$ and $U = 31$; and so on. Obviously we have

$$U = X - 100$$

for each of the measurements. The detail is shown in Table 4.

In practice we do not go into such detail. We abbreviate our work a great deal as is indicated in Table 5. We follow a few systematic steps:

(1) We decide upon the transformation we wish to use. For the weight data we use $U = X - 100$.

(2) We complete the table to agree with the chosen transformation. That is, we find the U that corresponds to a given X.

(3) We derive a formula to agree with the chosen transformation. Thus, when $U = X - 100$, we have

$$\begin{aligned} X &= U + 100 \\ \Sigma X &= \Sigma(U + 100) = \Sigma U + \Sigma 100 \\ \Sigma X &= \Sigma U + 10(100) \end{aligned}$$

since N, the number of measurements, is 10.

(4) We substitute the values found from the table, step (2), into the formula we derive in step (3).

TABLE 5. WEIGHTS OF 10 MEN

Weight X	$U = X - 100$
128	28
131	31
137	37
143	43
144	44
146	46
147	47
149	49
155	55
170	70
	$450 = \Sigma U$

Solution

$$U = X - 100$$
$$X = U + 100$$
$$\Sigma X = \Sigma U + N(100)$$
$$\frac{\Sigma X}{N} = \frac{\Sigma U}{N} + 100$$

$$\text{Average weight} = \frac{450}{10} + 100$$

$$= 145 \text{ pounds}$$

We now submit two more illustrations that involve simple transformations. In each case our problem is to find ΣX by using ΣU.

X	$U = \frac{X}{125}$
125	1
250	2
375	3
500	4
625	5
	$15 = \Sigma U$

$U = \frac{X}{125}$ or $X = 125U$

$\Sigma X = \Sigma 125U = 125\Sigma U$
$\Sigma X = 125(15) = 1875$

X	$U = \frac{X - 384}{128}$
128	-2
256	-1
384	0
512	1
640	2
768	3
	$3 = \Sigma U$

$U = \frac{X - 384}{128}$ or $X = 128U + 384$

$\Sigma X = \Sigma(128U + 384) = \Sigma 128U + \Sigma 384$
$\Sigma X = 128\Sigma U + N(384)$
$\Sigma X = 128(3) + 6(384) = 2688$

EXERCISES

1. Complete the following tables and find the values of the quantities suggested.

a

X	$x = X - 81$	x^2
85		
94		
73		
66		
87		
95		
80		
72		
95		
63		
$\Sigma X = 810$ Av. $= 81$		

$\Sigma x = (\ )$
$(\Sigma x)^2 = (\ )$
$\Sigma x^2 = (\ )$
$\sqrt{\Sigma x^2} = (\ )$

b

X	$U = X - 50$	U^2
72		
75		
60		
53		
75		
92		
73		
60		
85		
55		

$\Sigma U = (\ )$ $\Sigma U^2 = (\ )$
Show that $\Sigma X = 10(50) + \Sigma U$
$\Sigma X = (\ )$
Can you find ΣX^2 by using ΣU and ΣU^2?
$\Sigma X^2 = (\ )$

c

X	Y	X^2	XY	Y^2
2	11			
4	8			
6	7			
8	5			
10	4			

$\Sigma X^2 = (\ )$
$\Sigma Y^2 = (\ )$
$(\Sigma X)(\Sigma Y) = (\ )$
$\Sigma XY = (\ )$

d

X	Y	$x = X - 6$	$y = Y - 7$	x^2	xy
2	11				
4	8				
6	7				
8	5				
10	4				
30	35				
Average 6	7				

$\Sigma x = (\ )$ $\Sigma y = (\ )$
$\Sigma x^2 = (\ )$ $\Sigma xy = (\ )$
$\frac{\Sigma xy}{\Sigma x^2} = (\ )$

5. REMARKS ON MEASUREMENT

It is almost a commonplace that nearly all the numerical data used by the statistician are necessarily approximations, true usually to two, three, or more figures. The observer who collects the original data and the statistician who undertakes to analyze and interpret them are frequently different individuals. The statistician must accept the measurements that are given him and should seek to obtain results that are consistent with the data.

The degree of approximation of a measurement depends upon the skill and the carefulness of the operator and upon the kind of instrument used. Of course it may happen that a measurement is exactly the true value of the quantity measured, but the measurer can never know when this is so. Since statistics has to do primarily with observed measurements, which are admittedly approximations, and with processes that are also approximative, it is obvious that any numerical result computed from them will in like manner be an approximation.

6. DECIMAL ACCURACY

The average student may have some difficulty in grasping the idea that accuracy is a relative matter and absolute precision of measurement an impossibility. He is accustomed to think of 9.7 as meaning the same thing as 9.70 and even 9.70000000 . . . to an unlimited number of decimal places.

If 9.7 does not mean the ideal number 9.7000000 . . ., what does it mean? For the sake of clarity of understanding and precision of statement, the scientist has adopted the convention that "9.7" means between 9.65 and 9.75. If we record the length of a line as 18 inches we mean that it lies between 17.5 inches and 18.5 inches. When we say that the distance to the moon is 240,000 miles, we mean that that distance is between 235,000 and 245,000 miles.

A measurement recorded as 18 inches means that the measurement is correct to the nearest inch or to units. A measurement [1] recorded as 9.7 cm. means that the number is correct to the nearest tenth of a centimeter. The number is sometimes written 9.7 ± 0.05 in which the expression 0.05 should be read "with a possible error of 0.05."

Similarly, a recorded value of 9.70 would mean from 9.695 to 9.705 and might be written 9.70 ± 0.005.

Unless otherwise specified, a score for a continuous variate should be interpreted as extending from half a unit of the last place of the measurement below to half a unit above the recorded entry. A similar assumption regarding discrete data avoids confusion in the analysis. Hence we shall assume that a measurement for a discrete variate extends from half a unit below to half a unit above the recorded score.

7. SIGNIFICANT FIGURES

In the expression 9.7 cm., both the 9 and the 7 mean something or *are significant.* In the expression 97 mm., there are likewise two significant figures. There are five significant figures in each of the numbers 203.05, 263.10, 0.0076389, 500.00, but only two in the number 93,000,000 which gives the approximate number of miles from the earth to the sun.

When the distance from the earth to the sun is given as 93,000,000 miles, in the light of the convention that we discussed in the preceding section, the statement might be interpreted to mean that the distance is between 92,999,999.5 and 93,000,000.5 miles. Since the figures 9 and 3 alone are to be regarded as significant, the exact distance is between 92,500,000 and 93,500,000 miles. This confusion can be prevented by writing the number in the *standard* form 9.3×10^7,

[1] If a measurement may be written $a \pm e$, we call e the possible error in a, the measurement.

the number of significant figures being indicated by the factor at the left which has *one* figure before the decimal point.

We determine the significant digits in a number by reading the number from left to right, *commencing with the first digit not zero and ending with the last digit accurately specified.* The position of the decimal point has no influence on the number of significant digits.

Thus 34 has two significant figures; 7.3, two; 406, three; 7,003, four; 8.0, two; 0.40, two; 9.00, three; 0.006, one; 0.0050, two; and 2.4×10^6, two.

8. ROUNDING OFF NUMBERS

Sometimes we are furnished with numbers recording measurements that are given with a greater accuracy than we can use, or care to use. We accordingly *round them off* to the accuracy desired.

A number is rounded off by dropping one or more digits at the right. When the digit dropped is 5 or more, increase the preceding digit by unity; when it is less than 5, retain the preceding digit unchanged.

The following numbers are rounded according to the above rule:

Numbers	*Rounded Values*
4.5647	4.565; 4.57; etc.
0.49781	0.498; 0.50; etc.
17.65	17.7
17.75	17.8

9. ERRORS IN CALCULATIONS

As magnitudes determined by measurement are not exact, it is important to make clear the meaning of the term *error* as it is used in statistics.

In the first place, errors are not necessarily what we usually think of as mistakes or blunders. The latter arise from carelessness or incompetency in transcribing figures or reading values from a scale. An *absolute error* in observation is the difference between a given measurement and the true value of the quantity measured. Therefore, an error means a deviation, a difference, but not a mistake.

The *relative error* in a measurement is the ratio of the absolute error to the true value of the quantity. It may be closely approximated by finding the ratio of the possible error to the given measurement.

It is usually expressed as a percentage. Thus if a measurement of height is given as 68.5 inches, there is a possible error of 0.05 inches and an approximate relative error of $0.05/68.5 = 0.0007$, which equals 0.07 per cent. If a physician reports the weight of a man as 163 pounds with a possible error of 0.5 pound, the approximate relative error may be written as $0.5/163 = 0.003 = 0.3$ per cent.

The relative errors in the two distances 9.3×10^7 and 9.30×10^7 are approximately:

$$\frac{0.05}{9.3} = 0.005 = 0.5\% \quad \text{and} \quad \frac{0.005}{9.30} = 0.0005 = 0.05\%$$

This illustrates the fact that the relative error depends upon the number of significant figures in and not upon the position of the decimal point in a recorded measurement.

EXERCISES

1. How many significant figures are in the following numbers?

(1) 2.375 (2) 0.0347 (3) 0.0030 (4) $5.63(10^6)$ (5) $5.6300(10^2)$

2. What is the rule to be observed when rounding off numbers?

3. A line is measured and its length is recorded as 118.63 feet. What does this statement mean? What is the approximate relative error in the measurement?

4. A line is measured and its length is recorded as 125.65 feet. What does this statement mean? What is the approximate relative error in the measurement?

5. The population of a city is given as 2.5 million. What is the approximate percentage error?

6. The population of a city is given as 340 thousand. What is the approximate percentage error?

7. The value of π correct to five significant figures is 3.1416. Determine the percentage error when π is approximated by $3\frac{1}{7}$.

8. The values of all mineral production in continental United States in 1929, correct to the nearest million dollars, was \$5,165,000,000. Write this value in the *standard* form. Find the approximate percentage error in the given estimated value.

9. Prove:

$$\sum_{1}^{n} 2x = n(n+1).$$

10. Use the result of Number 9 to find the sums:

(1) $30 + 32 + 34 + \cdots + 96$.

(2) $128 + 130 + 132 + \cdots + 164$.

11. Find the sum of the following numbers correct to two decimal places: 2.4286, 12.673, 127.87, 35.583:

(1) By retaining the significant figures of the numbers and rounding off the sum to two places;

(2) By rounding off each number to two places and finding the sum.

This exercise illustrates the rule: "When several approximate numbers are to be added, it is best to round them at once to the number of decimal places in the least accurate measurement."

12. Find the sum of the following numbers correct to two decimal places: 3.4285, 16.743, 253.78, 36.583:

(1) By retaining the significant figures of the numbers and rounding off the sum to two places;

(2) By rounding off to two places each number and finding the sum.

10. THE PROPAGATION OF ERRORS

In general, statistical computations are more concerned with relative than with absolute errors. We shall include here the more important theorems that relate to relative errors and expect the reader who desires a wider knowledge to consult the splendid work by Scarborough.[1]

Theorem I. *The possible error in the sum or the difference of two measurements is equal to the sum of the possible errors in the individual measurements.*

Suppose a and b are the readings of the two measurements and that e_1 and e_2 are the numerical values of their errors. The true values are therefore $a + e_1$ and $b + e_2$, where e_1 and e_2 may be either positive or negative. The correct value of the sum of the measurements lies between the limits:

$$(a + e_1) + (b + e_2) = (a + b) + (e_1 + e_2)$$

and

$$(a - e_1) + (b - e_2) = (a + b) - (e_1 + e_2)$$

Hence the possible error in the sum, $a + b$, is $e_1 + e_2$.

The correct value in the difference of the measurements lies between the limits:

$$(a + e_1) - (b - e_2) = (a - b) + (e_1 + e_2)$$

and

$$(a - e_1) - (b + e_2) = (a - b) - (e_1 + e_2)$$

[1] J. B. Scarborough, *Numerical Mathematical Analysis*, p. 2.

Hence the possible error in the difference, $a - b$, is $e_1 + e_2$.

Example. The sides of a rectangular field are measured to be $127' \pm 0.2'$ and $231' \pm 0.4'$. Find the possible error in the sum of the two sides.

We have:

$$a = 127,\ b = 231 \qquad e_1 = 0.2,\ e_2 = 0.4$$
$$a + b = 358 \qquad e_1 + e_2 = 0.6$$

Hence the possible error is $0.6'$ and the true value of the sum of the two sides is between $358 - 0.6$ and $358 + 0.6$ feet.

Theorem II. *The relative error in the product of two measurements is equal to the sum of the approximate relative errors of the individual measurements.*

With the same notation as above, the product will lie between:

$$(a + e_1)(b + e_2) = ab + ae_2 + be_1 + e_1e_2$$

and

$$(a - e_1)(b - e_2) = ab - ae_2 - be_1 + e_1e_2$$

Since e_1 and e_2 are both small when compared to the other terms of the products, we shall ignore the term e_1e_2. We then have the possible error in the product to be approximately $ae_2 + be_1$.

Hence the relative error in the product is approximately

$$\frac{ae_2 + be_1}{ab} = \frac{e_1}{a} + \frac{e_2}{b}$$

which is the sum of the approximate relative errors.

Example. Find the absolute and the relative errors in the computed area of the rectangle whose sides are $127' \pm 0.2'$ and $231' \pm 0.4'$.

The possible error in the product is approximately

$$127(0.4) + 231(0.2) = 97 \text{ square feet}$$

and the true value of the area is somewhere between

$$(127)(231) - 97 \text{ and } (127)(231) + 97,$$

that is, between 29,240 and 29,434 square feet.

The relative error in the area is approximately:

$$\frac{0.2}{127} + \frac{0.4}{231} = 0.0032 = 0.32\%$$

Theorem III. *The relative error in the quotient of two measurements is equal to the sum of the approximate relative errors of the measurements.*

The quotient will evidently lie between:

$$\frac{a + e_1}{b - e_2} = \frac{a}{b} + \frac{ae_2 + be_1}{b(b - e_2)}$$

and

$$\frac{a - e_1}{b + e_2} = \frac{a}{b} - \frac{ae_2 + be_1}{b(b + e_2)}$$

Since e_2 is small compared with b, we may, for purposes of approximation, replace $b + e_2$ and $b - e_2$ by b; whence the possible error in the quotient is approximately:

$$\frac{ae_2 + be_1}{b^2}$$

Hence the relative error in the quotient is given approximately by

$$\frac{ae_2 + be_1}{b^2} \div \frac{a}{b} = \frac{e_1}{a} + \frac{e_2}{b}$$

which is the sum of the approximate relative errors.

Example. Find the possible and the relative errors when 625 ± 0.7 is divided by 36 ± 0.2.

We have:

$$a = 625, \quad e_1 = 0.7$$
$$b = 36, \quad e_2 = 0.2$$

The possible error in the quotient is given approximately by

$$\frac{625(0.2) + 36(0.7)}{36^2} = 0.12$$

and the true value of the quotient will therefore lie between

$$\frac{625}{36} - 0.12 \quad \text{and} \quad \frac{625}{36} + 0.12$$

that is, between 17.24 and 17.48.

The relative error in the quotient is given approximately by:

$$\frac{0.7}{625} + \frac{0.2}{36} = 0.0066 = 0.66\%$$

EXERCISES

Make each of the following computations and state the result so as to show a measure of the error involved.

1. $(125 \pm 0.2) + (238 \pm 0.3)$.

2. $(215 \pm 0.2)(115 \pm 0.3)$.

3. $(163 \pm 0.2)/(25 \pm 0.4)$.

4. What is the possible error in the area of a rectangle whose length and width are recorded as 50.4 ft., and 30.6 ft.?

5. a. Show that if e is the error in the side of a square whose recorded length is a, then the error in the area is approximately $2ae$.

b. Show that the relative error in the area is approximately twice the relative error in the edge.

6. Show that the relative error in the area of a circle is approximately twice the relative error of the radius.

7. The distance from the earth to the sun is given as 93,000,000 $\pm$ 500,000 miles, and the thickness of a watch spring is given as 0.014 $\pm$ 0.0005 inches. Which is the more accurate measurement?

8. Show that the relative error in the volume of a cube is approximately three times the relative error of the edge.

9. Show that the relative error in the volume of a sphere is approximately three times the relative error of the radius.

10. Are statistical data always approximate? If each of 10 men pays an income tax of \$87, is their total contribution \$870 approximate?

11. Find the value of $\Sigma(12x^2 - 4x + 3)$.

12. Find the sum of $3 \cdot 7 + 4 \cdot 9 + 5 \cdot 11 + \cdots$ to n terms.

13. Find $11^2 + 12^2 + 13^2 + \cdots + 50^2$.

14. Prove:

$$\sum_1^n 2x(3x + 1) = 2n(n + 1)^2$$

15. Use the result of Number 14 to find the sums:

(1) $12 \cdot 19 + 14 \cdot 22 + 16 \cdot 25 + \cdots + 32 \cdot 49$.

(2) $36 \cdot 55 + 38 \cdot 58 + 40 \cdot 61 + \cdots + 80 \cdot 121$.

16. Prove that

$$\sum_1^n 2x(2x + 1) = \frac{n(n + 1)(4n + 5)}{3}$$

17. Use the result of Number 16 to find the sums:

(1) $12 \cdot 13 + 14 \cdot 15 + 16 \cdot 17 + \cdots + 96 \cdot 97$.

(2) $48 \cdot 49 + 50 \cdot 51 + 52 \cdot 53 + \cdots + 90 \cdot 91$.

18. Find in terms of n the value of $\Sigma x(x + 1)$

19. Find in terms of n the value of

$$1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + \cdots \text{ to } n \text{ terms.}$$

20. Find identities and prove that:

a. $\sum_{1}^{n} x^3 = \left[\frac{n(n+1)}{2}\right]^2$

b. $\sum_{1}^{n} x^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$

21. Find $\sum_{50}^{100} x^3$.

22. Find $1 \cdot 2^2 + 2 \cdot 3^2 + 3 \cdot 4^2 + \cdots$ to n terms.

23. The estimated value of anthracite coal produced in Pennsylvania in 1929 was $\$3.935(10^8)$ and the estimated quantity produced was $7.664(10^7)$ tons. What was the estimated value per ton and what was the relative error in the estimate?

24. The estimated production of tobacco in the United States in 1929 was $1.5(10^9)$ pounds and the estimated price received was 10.0 cents per pound. What was the estimated value of the crop? What was the percentage error in the estimated value?

25. The estimated production of potatoes in the United States in 1929 was $3.57(10^8)$ bushels, and the estimated price per bushel was 131.4 cents. What was the estimated value of the crop? What is the relative error in the estimated value?

26. The estimated production of potatoes in the United States in 1929 was $3.57(10^8)$ bushels and the estimated acreage was $3.37(10^6)$. What was the estimated yield per acre? What is the relative error in the estimated yield?

27. A teacher's salary of \$350 a month was decreased 10 per cent and later increased 5 per cent. What is his present salary?

28. City A increased in population from 25,750 to 35,890 in a decade, and City B increased from 255,000 to 350,000 during the same decade. Which city had the greater percentage increase?

29. The general price level rose 80 per cent, then declined $33\frac{1}{3}$ per cent. How much was it then above its starting point?

30. A man whose salary was actually \$275.00 a month was reported to be receiving \$300.00 a month. What was the percentage error in the report?

31. A teacher's salary of \$200 a month was decreased 25 per cent and then increased 25 per cent. What is his present salary?

32. The value of the exports from the United States to a neighboring country in 1934 was 15 per cent less than the value in 1933, but in 1935 was 10 per cent greater than the value in 1934. Compare the value in 1935 with that in 1933.

33. The number of registered passenger automobiles in the United States in 1929 was $2.3122(10^7)$ and the estimated population the same year was $1.22(10^8)$. What was the estimated number of people for each passenger automobile? What was the relative error in the estimate?

Chapter 2

TABULAR AND GRAPHICAL REPRESENTATION: FREQUENCY DISTRIBUTIONS

11. INTRODUCTION

Almost without exception the object of a statistical analysis is to form a judgment of a very large universe by means of a study of a small part of it. The large universe we call *the parent population,*[1] and the part of it that we use as a basis for generalization we call a *sample.*

In some cases it is impossible to measure the entire parent population and in other cases it is impracticable to do so. Suppose a physician was interested in the blood pressure of American men between thirty and thirty-one years of age. He could never expect to get complete data for all the men in the parent population. Not only would it be impossible; it would be unnecessary, expensive, and a waste of time and energy. An excellent judgment could be made by the study of a properly selected sample.[2]

Our first task therefore is to secure the data of a properly selected sample, and then proceed to the analysis. The analysis will give us a summarized numerical description of the sample from which we may, if we desire, form certain judgments of the parent population.

12. CLASSIFICATION OF THE DATA

When a mass of data has been assembled it is necessary to classify the material in some compact and orderly form before it can be effectively analyzed. This procedure is known by statisticians as *tabulation.* It is merely the arrangement of the data into tables, or in a tabular form. The data in the original form are *ungrouped;* when they are summarized into a table they are *grouped.*

The following table, Table 6, gives the scores made in college

[1] Statistically speaking, any mass of data is a *population.*

[2] Chapter 13 will deal more specifically with the problem of sampling.

algebra by 125 first-year students at Bucknell University. These scores constitute a sample selected at random from a larger population. The grades are given to the nearest integer on the centigrade scale. This means, we recall, that a grade recorded as 92 might represent any mark between 91.5 and 92.5. We note that the lowest recorded score is 48 and the highest score is 97, giving a *range* of 97 − 48 = 49. The possible range is from 47.5 to 97.5, or 50.

TABLE 6. SEMESTER GRADES OF 125 STUDENTS IN COLLEGE ALGEBRA AT BUCKNELL UNIVERSITY
(Grades recorded to the nearest integer)

93	83	77	75	70
88	69	68	71	63
86	58	53	50	95
79	89	87	84	78
82	81	78	81	74
80	75	76	77	75
73	48	76	69	55
74	62	95	90	84
75	87	65	70	68
76	70	55	63	79
65	80	97	91	64
68	70	79	86	83
80	57	60	65	79
80	76	82	75	60
75	77	62	59	92
85	73	74	77	70
68	65	70	72	69
90	85	85	81	80
77	67	66	67	63
77	73	74	75	73
69	81	80	72	72
85	82	77	73	73
74	74	75	72	70
71	75	76	76	77
74	75	71	70	72

If these grades in college algebra are arranged in the order of magnitude the *array* will be more suitable for study than in the haphazard arrangement in Table 6, yet even the grades arranged

in this manner will still be unwieldy for a close analysis. A really compact form may be obtained by arranging the measures into *classes* of equal width, for example, 47.5–52.5, 52.5–57.5, etc., wherein the *class interval* or *class width* is 5 points. The number of items or measures occurring in each class (called the *class frequency*) is then determined by tallying.

The traditional method of tallying is to record the frequencies by marks until four have been made, then to make a cross mark for the fifth score. This procedure makes up the *preliminary sheet.*

The procedure described above for tallying offers no facilities for checking. If a repetition of the classification leads to a different result, we have no means of tracing the error. If the number of observations is large, it is better to enter the values on cards, one card to each measure, then sort the cards into the classes we desire. We can then check each pack, thereby placing each measure in the proper class.

The tabular arrangement — illustrated by Table 7 — consisting of a series of classes and a corresponding set of frequencies is called a *simple frequency distribution.* We designate the total frequency by N.

TABLE 7. SEMESTER GRADES OF 125 STUDENTS IN COLLEGE ALGEBRA
PRELIMINARY SHEET

Class	*Tally*	*Frequency*
92.5–97.5	////	4
87.5–92.5	𝍸 /	6
82.5–87.5	𝍸 𝍸 //	12
77.5–82.5	𝍸 𝍸 𝍸 ////	19
72.5–77.5	𝍸 𝍸 𝍸 𝍸	
	𝍸 𝍸 𝍸 //	37
67.5–72.5	𝍸 𝍸 𝍸 𝍸 ////	24
62.5–67.5	𝍸 𝍸 /	11
57.5–62.5	𝍸 /	6
52.5–57.5	////	4
47.5–52.5	//	2
Total		125 = N

The *organization* of the data has thus been effected and the data are now prepared for the next step, the *analysis.*

TABLE 8. SEMESTER GRADES OF 125 STUDENTS IN COLLEGE ALGEBRA
(Grades recorded to the nearest integer)

Form (a)			Form (b)	
Class	*Class Mark* X	*Frequency* $f(X)$	*Class Mark* X	*Frequency* $f(X)$
92.5–97.5	95	4	95	4
87.5–92.5	90	6	90	6
82.5–87.5	85	12	85	12
77.5–82.5	80	19	80	19
72.5–77.5	75	37	75	37
67.5–72.5	70	24	70	24
62.5–67.5	65	11	65	11
57.5–62.5	60	6	60	6
52.5–57.5	55	4	55	4
47.5–52.5	50	2	50	2
Total		$125 = N$	*Total*	$125 = N$

In the preparation of Table 8 we were cognizant that the data are continuous and are *recorded to the nearest integer.* A score recorded as 79, for example, really fell somewhere over the interval 78.5 to 79.5. Consequently we found it convenient to represent the end values of the class intervals to *tenths.* If the data had been recorded to tenths, we could have expressed the two figures defining each class to *hundredths.*

The two figures that define a class are called the *class limits* of the class. In some tabular representation of classes, the defining numbers of the class are *true class limits* or *class boundaries.*[1] The class boundaries can easily be determined as *each boundary is half way between the largest item in the lower class and the smallest item in the next higher class.* Thus in Form (a) above the largest measure in the lowest class is 52 and the smallest value in the next higher class is 53. The *class* boundary is half way between 52 and 53, that is at 52.5. The other boundary points in Form (a) can be found in a similar manner.

The difference between the lower boundary of one class and the lower boundary of the next higher class is the *class interval* or *class width.* The class interval is also the difference between the upper

[1] Some authors call class boundaries *closed class limits.*

boundaries of two adjacent classes. The upper boundary of one class is the lower boundary of the next higher class, and the lower boundary of one class is the upper boundary of the next lower class. That is, for continuous data adjacent classes should "join up" or be contiguous. The number half way between the upper and lower boundaries of a class is the *class mark*. Thus

$$\text{Class mark} = \frac{\text{Upper boundary} + \text{Lower boundary}}{2}$$

A class boundary is half way between the class marks of two adjacent classes. The class boundaries of a class can be found by adding to and subtracting from the class mark one half the class width. With this in mind, Form (b) is a mere abridgment of Form (a).

Form (a), using class boundaries, is a widely used method of indicating the classes of a simple frequency distribution. It is suitable to discrete as well as to continuous data, and we recommend it as our favorite method. However other methods for defining the classes are found in the literature of the subject. We shall present and discuss some well known forms to which the data of Table 8 may be applied.

Form (c)

Class	*Class Boundaries*	*Class Mark Continuous*	*Class Mark Discrete*
93 a.u. 98	92.5–97.5	95	95
88 a.u. 93	87.5–92.5	90	90
etc.	etc.	etc.	etc.

In Form (c), "93 a.u. 98" means "93 and under 98." That is, in this class are found the measures as large as 93 but less than 98. The classes in Form (c) are defined by *class limits* but not by class boundaries. For clearness, we give the class boundaries which in turn assist us in finding the class marks. Form (c) is suitable for continuous and discrete data, but in using this form the student must recall that a score of 93 means any number in the interval 92.5 to 93.5 and thus the lower boundary is 92.5. Similarly, a score of 88 has a lower boundary at 87.5. The class marks are now easily determined.

Occasionally the classes are denoted by the smallest and largest

measures of a given class, and the class interval may *appear* to range from the smallest to the largest measurement for each class. For continuous variates, this method of defining the class does not show the full range of the class and leaves gaps at the ends of the class. In this, as in all forms of class representation, *the statistician must ascribe to each class the true class limits or the class boundaries, and*

Form (d)

Class	*Class Boundaries*	*Class Mark Continuous*	*Class Mark Discrete*
93–97	92.5–97.5	95	95
88–92	87.5–92.5	90	90
etc.	etc.	etc.	etc.

the true class mark. Thus in Form (d) in the given classes we indicate the class limits by the smallest and largest values that may fall in a given class. We have included, for emphasis and for clearness, the class boundaries.

Occasionally we find in the literature a tabular representation similar to Form (e). This form states ambiguously what Form (c) states more definitely. It is unsafe for tallying scores for the reason

Form (e)

Class	*Class Boundaries*	*Class Mark Continuous*	*Class Mark Discrete*
93–98	92.5–97.5	95	95
88–93	87.5–92.5	90	90
etc.	etc.	etc.	etc.

that it is easy to mis-tally boundary scores. Thus, to which class would a score of 93 belong? Again, we have included the class boundaries for sake of clearness, also the class marks.

In later chapters we shall find it necessary to locate certain division points on the X-scale: quartiles, deciles, percentiles. To find these points we shall need *true class limits* or *class boundaries.*

The determination of true class marks is also very important as

many of our statistical constants, such as the arithmetic mean and the standard deviation, are found from the class marks of the classes. In fact to save labor in computation, we shall find it necessary to assume that the items are uniformly distributed over the given intervals and that the class frequencies are concentrated at the class marks.

From this discussion of the several forms it is evident that, inasmuch as the *class boundaries* must eventually be found to aid in the analysis of the data, *we can save ourselves confusion and time by adopting class boundaries in the beginning of our problem.* This procedure we have followed and it is one we highly recommend.

It should be emphasized that when a score is tabulated in the proper class interval, it loses its identity. Of course it falls somewhere within the boundaries of the interval, but in computation we do not use it again. For computational purposes in effecting the numerical analysis, it is necessary that we concentrate the class frequency at the mid-point of the class interval. Thus, in our computations on Table 8, we replace the scores 93, 95, 97, 95 of the class 92.5 – 97.5 by four scores each of value 95, the mid-value of the class. Similarly, we replace the scores 88, 90, 89, 90, 91, 92 of the class 87.5 – 92.5 by six scores each of value 90, the mid-value of the class. And so on for the other classes.

While our assumption that the scores are evenly distributed over the interval is seldom verified by observed data, yet if the sample is sufficiently numerous the assumption leads only to a very slight error. Some such assumption must be made, and experience and statistical theory recommend the assumptions of evenness of measures over the interval and the concentration of the class frequency at the mid-point of the class interval.

Example 1. If 10 scores in integral variates are evenly distributed over the interval 72.5 – 77.5, what are the scores?

Since the scores are integers, 10 in number, and must be evenly distributed over the interval, they would have the values 73, 73, 74, 74, 75, 75, 76, 76, 77, 77.

Example 2. Are the values 73, 73, 73, 73, 73, 77, 77, 77, 77, 77 evenly distributed over the interval 72.5 – 77.5?

No. While the statistical results are essentially the same as if the entire 10 scores are situated at the class mark, 75, these values are not evenly distributed over the given interval.

Example 3. If 20 measurements, rounded to the nearest half-inch, are evenly distributed over the interval 72.25 – 82.25, what are their values?

Their values are: 72.5, 73.0, 73.5, 74.0, . . ., 82.0.

Would two of each of the following 10 measurements be satisfactory: 73, 74, 75, . . ., 82?

Example 4. What measurements would satisfy for the preceding example if the interval were 72 a.u. 82? What is the upper boundary of the class?

The values would be: 72.0, 72.5, 73.0, . . ., 81.0, 81.5. The largest value in the class is 81.5 and the smallest value in the next higher class is 82.0. The class boundary is the value half way between them, namely 81.75.

EXERCISES

1. Suppose the data are dinner checks from a cafeteria. Show that two checks for each of the values 93, 94, 95, 96, 97 cents would give the same total as 10 checks of 95 cents each.

2. Suppose the temperature at Lewisburg is recorded to the nearest tenth of a degree and that a 5 degree class interval has been selected. If the class limits are 60.0 – 64.9, 55.0 – 59.9, 50.0 – 54.9, etc., what are the class boundaries and the class marks of the three classes?

3. A group of intelligence quotients (continuous data) are arranged with the class intervals as follows: 75 – 79, 80 – 84, etc. What are the class boundaries and the class marks?

4. What values are contained in the interval 75 – 79 if the data are discrete? If the data are continuous and recorded to the nearest integer?

13. THE CHOICE OF THE CLASS INTERVAL

In the choice of a class interval, the following brief suggestions may be helpful:

1. The number of classes should, in general, not be less than 10 nor more than 30, seldom more than 25.
2. If possible, the class intervals should be uniform in width.
3. In general there should be no class intervals without definite limits. Intervals of the type "all over" and "all under" are to be avoided when possible.[1]
4. To facilitate computation, class intervals of multiples of 5 or 10 are convenient.

14. CLASS LIMITS

The lowest limit of the lowest class may be chosen in many positions. This choice and that of the class interval will practically

[1] Many of the tables found in the data sent out by the United States Government are of this type.

determine the limits of the other classes. We rather hesitate to state many rules for their selection; much must be left to the judgment and resourcefulness of the student. The following suggestions should prove helpful:

1. To facilitate computation, the mid-points should be integers. We shall find that carrying out this suggestion is frequently impossible.
2. Certain types of data are loaded at special points. For example, college marks on a centigrade scale are loaded at 60, 65, 70, 75, etc. Distributions in which *age* is the independent variable are usually loaded at 20, 25, 30, etc. When the data display such a peculiarity, these loaded points should be chosen as mid-points of the class intervals. This is especially to be kept in mind, since in the analysis of our distributions we shall assume that all measures of a class are concentrated at the mid-point of the class.
3. Class limits should be unambiguous and mutually exclusive.

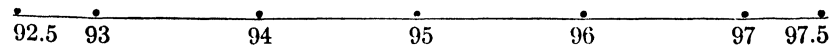
The class limits can be decided accurately only when the accuracy of the data is known. A score for either type of variate is assumed to extend from half a unit in the last place of measurement below to half a unit above the entry recorded. If the data are accurate to tenths, the class limits should be expressed to hundredths; if heights are measured to the nearest quarter of an inch, the class limits should be arranged in eighths of an inch.

In many of the exercises that follow in the text, it will not be possible to carry out the suggestion of the preceding paragraph, because the original observers were not meticulously careful to state the accuracy of the original measurements. When this is the case we shall have to make some reasonable assumptions and proceed along the line suggested by them.

EXERCISES

1. The following diagram for the first class of Form (a), Table 8, shows that the mid-point of the class (the class mark) is 95. Make similar diagrams to explain the other Forms.

DIAGRAM 1



2. Suppose you were asked to construct a frequency table of the grades in Table 6 (p. 24) with the class marks at 97.5, 92.5, etc. and with a class interval of 5, what would you say?

3. An educational research department recently sent the author a score card for some data (to the nearest integer). The *Class* column was marked thus: 0–4, 5–9, 10–14, etc. What are the class marks? the true class limits?

4. The daily wages of 100 men were recorded to the nearest cent. Complete the table finding the class boundaries and the class marks.

Class	*Class Boundaries*	*Class Mark*	$f(X)$
\$2.25–2.49			5
2.50–2.74			11
2.75–2.99			23
3.00–3.24			29
3.25–3.49			17
3.50–3.74			9
3.75–3.99			6
Total			100

5. The weights of 1,000 male students (in pounds) were recorded to the nearest half pound. Complete the table.

Class Boundaries	*Class Mark*	$f(X)$
	105.25	4
	115.25	12
	125.25	20
	etc.	etc.

6. The heights of 1,000 male students (in inches) were recorded to the nearest tenth of inch. Complete the table.

Class	*Class Boundaries*	*Class Mark*	$f(X)$
60.8 a.u. 62.8		61.75	1
62.8 a.u. 64.8		63.75	3
64.8 a.u. 66.8		65.75	11
etc.		etc.	etc.

7. The ages at marriage of 100 women were distributed as shown in the table. Find the class boundaries.

Class	*Class Boundaries*	*Class Mark*	$f(X)$
15–19		17	4
20–24		22	28
25–29		27	23
etc.		etc.	etc.

8. The number of pedicels per cluster of a certain plant resulted in the following distribution. Find the class boundaries.

Class	*Class Boundaries*	*Class Mark*	$f(X)$
12–19		15.5	8
20–27		23.5	52
28–35		31.5	176
etc.		etc.	etc.

9. A distribution of heights of students (in centimeters) was arranged as follows:

Height (centimeters)	*Class Mark*	$f(X)$
155–157	156	4
158–160	159	8
161–163	162	26
etc.	etc.	etc.

What are the class boundaries?
Can you guess at the accuracy of the original measurements?
Do you agree with the class marks?

10. Suppose you are given 500 grades (in per cent) in English to distribute. The lowest grade is 20% and the highest grade is 90%. You decide upon a class width of 5%. Which of the two groupings, A or B, would be preferable?

A

Class	X	$f(X)$
19.5–24.5	22	
24.5–29.5	27	
etc.	etc.	

B

Class	X	$f(X)$
17.5–22.5	20	
22.5–27.5	25	
etc.	etc.	

11. In his book, "The Fundamentals of Statistics," Professor L. L. Thurstone tabulates the scores made on an intelligence test by 140 freshmen at Swarthmore College. His table follows:

Scores	*Class Mark*	$f(X)$
40–49	45	1
50–59	55	5
60–69	65	12
70–79	75	21
80–89	85	23
90–99	95	23
100–109	105	25
110–119	115	14
120–129	125	11
130–139	135	4
140–149	145	1
Total		140

What are the class boundaries?

Using our assumptions, what are the class marks of the classes?

What are the tacit assumptions that Professor Thurstone makes regarding the extreme scores in the classes?

12. The following data pertain to the ages of unemployed male workers in Boston in 1930. Professor R. C. White in his "Social Statistics," page 215, takes the classes and the class marks as shown.

Age (Years)	*Class Mark*	$f(X)$
10–14	12.5	
15–19	17.5	
20–24	22.5	
etc.	etc.	

What are the class boundaries?

Do you agree with the class marks?

What assumptions does Professor White evidently make regarding the largest and smallest ages of a class?

13. In his "Statistical Methods, Revised," Professor F. C. Mills exhibits on page 105 a distribution of the weekly earnings of workers in open-hearth furnaces in the Pittsburgh district in 1935. A portion of the table is shown here.

Class Interval (in dollars per week)	*Mid-point*	*Frequency*
$0– 3.99	2	67
4– 7.99	6	290
8–11.99	10	437
etc.	etc.	etc.

What are Professor Mills' assumptions regarding the values that are placed in the given classes?

According to our assumptions what would be the mid-points of the classes?

14. In Davies and Yoder, "Business Statistics," pages 110 and 114 we find the following distribution:

L_1–L_2	X	f
10–12	11	3
12–14	13	15
14–16	15	20
16–18	17	10
18–20	19	2

What are the assumptions of the authors regarding the values that are placed in the several classes?

According to our assumptions what would be the values of X?

15. In his book, "Statistics for Students of Psychology and Education," Professor Herbert Sorenson on page 43 exhibits a distribution of scores obtained on an objective test in educational psychology. Here is a portion of the table.

Scores by Intervals	X	f
80–84	82.5	3
75–79	77.5	5
70–74	72.5	7
etc.	etc.	etc.

What are Professor Sorenson's assumptions regarding the scores in the given intervals? (See page 44 of his text.)

What are the class boundaries and the values of X according to our assumptions?

16. In his book, "The Mathematical Part of Elementary Statistics," Professor B. H. Camp gives on page 8 a distribution of wage data, a portion of which we show here. What are Professor Camp's assumptions regarding the scores in the given intervals? What are the class boundaries of the classes?

Class	*Mid-value*	*f*
$4.50–5.99	5.245	43
6.00–7.49	6.745	99
etc.	etc.	etc.

The illustrations found in these Exercises certainly show that authorities differ in their interpretations of class limits, class marks, class boundaries, et cetera. In reading the literature of our field we must be alert, therefore, to the assumptions, either tacit or expressed, that guide the procedure. Further, we must be charitable and seek to understand what are the assumptions that are guiding an author's steps, and realize that there is more than one way of doing a simple statistical task.

The problem we are discussing is simply this: what is meant by a recorded score of 74? of 74.6? of 74.67? We assume that if a score is recorded 74, its value is between 73.5 and 74.5; if it is recorded 74.6, its value is between 74.55 and 74.65; and so on. On the contrary, many statisticians assume that if a score is recorded 74, its value ranges from 74 to but not including 75; if a score is recorded 74.6, it ranges from 74.6 to but not including 74.7; and so on. They also use another method of description. They assume that a recorded score of 74 ranges from 74 to 74.99; a recorded score of 74.6 ranges from 74.6 to 74.699; and so on.

The mathematician, accustomed to rigor in his thinking, generally prefers our method of description, namely, that the classes be determined rigidly by class boundaries, whereas the worker in an applied field may be willing to sacrifice some rigor. This is one of the controversial questions in statistical procedure so let us not assume that we have the full truth. After all, it is not a matter of extreme importance whether the scores on an English test average 74.26% or 74.16%. It is essential, however, that we impose refinements when the data warrant them. It is just as essential that we do not give a false impression of accuracy in our procedures.

15. GRAPHICAL REPRESENTATION

When the data have been organized into a suitable table, they are now ready for the first step in the analysis, that of presenting the data graphically. Graphical presentations display outstanding facts and bring into bold relief relationships that otherwise would be difficult to comprehend or possibly would not be noted at all. A column of figures may overwhelm us; the same data in graphic form may tell an easily understood story. Relative quantities especially can be grasped through visual means with a comprehensiveness that is not possible by pure analysis.

While the ultimate basis of graphical presentation is mathematical, yet the practical work of constructing the charts can be accomplished without a profound knowledge of the true mathematical basis. Charts and graphs, then, can enable us to discover simply and quickly many facts and mathematical relationships about numerical data without the use of more difficult methods of analysis. The careful statistician, however, will be very cautious to verify by the more precise methods of analysis the suggestions that he receives from the graph.

It is not our intention to present in this book a detailed account of the many graphical procedures that are used today. We shall explain certain important principles of graphic presentation and leave it to the reader who desires a more comprehensive knowledge to consult the excellent volumes that are accessible.[1]

16. GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

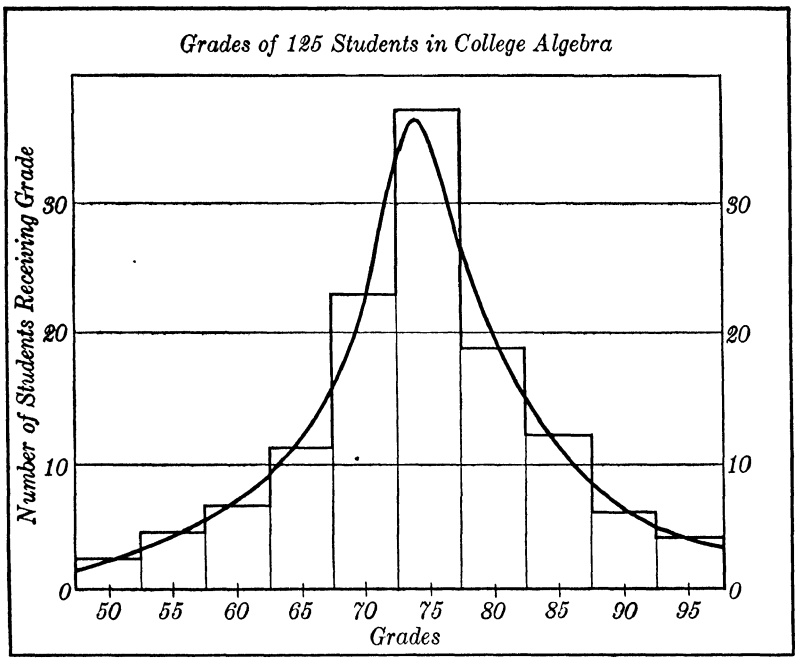
Probably the best graphical representation of a simple frequency distribution is furnished by a *column diagram* or *histogram*. It is constructed by erecting upon the class intervals rectangles whose altitudes are proportional to the frequencies. Suitable scales must be chosen so that the graph of the data can be made to fit the data, be of sufficient size to be readily interpreted, and be of such proportions that it will be agreeable to our artistic tastes. The left-hand side of the first rectangle is plotted at the lower boundary of the

[1] Excellent references for graphical presentation are listed in the bibliography in Appendix A of this volume.

lowest class and the right-hand side of the last rectangle is plotted at the upper boundary of the highest class. Chart 1 shows the histogram for the distribution of grades in college algebra previously tabulated in Table 8 (p. 26).

The student will note that each rectangle contains an *area* that is proportional to and *represents* the frequency of the class and that

CHART 1



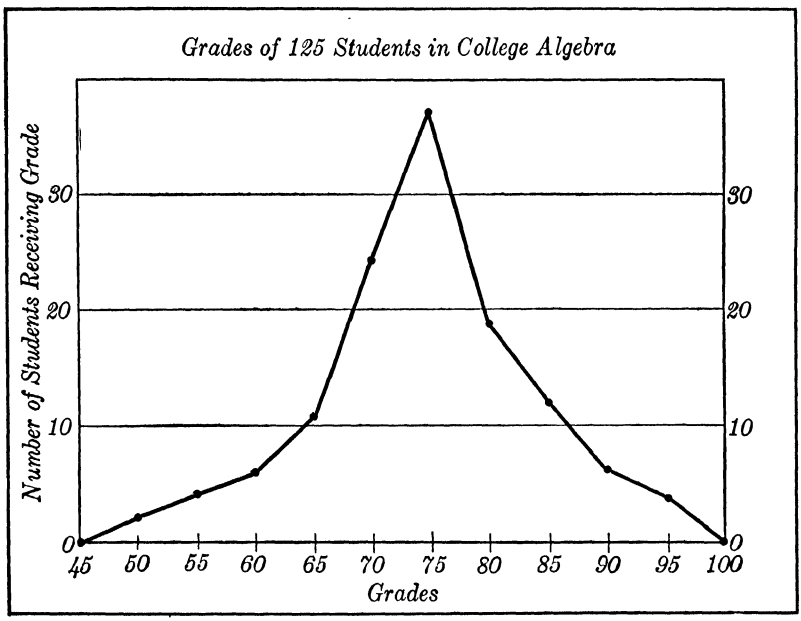
the total area equals the total frequency times the class width. If the class width is taken as the unit, the total area equals the total frequency.

Another method of representing graphically a frequency distribution is by what is called a *frequency polygon.* Its construction is very much like the plotting of curves and line diagrams in elementary algebra. In form (b) of Table 8 (p. 26), each pair of values X, $f(X)$, defines a point. Plotting the several points and connecting them by a broken line, we obtain the frequency polygon. The last points at

either end must be joined to the base at the center of the next class interval. The observing student will note that the vertices of the frequency polygon are merely the mid-points of the tops of the rectangles of the histogram, and that the *ordinates represent* the frequencies. Chart 2 shows the frequency polygon for the grades in college algebra displayed in tabular form in Table 8 (p. 26).

The fact that the polygon extends beyond the limits of the table suggests that if the grades of a larger group of students were taken,

CHART 2



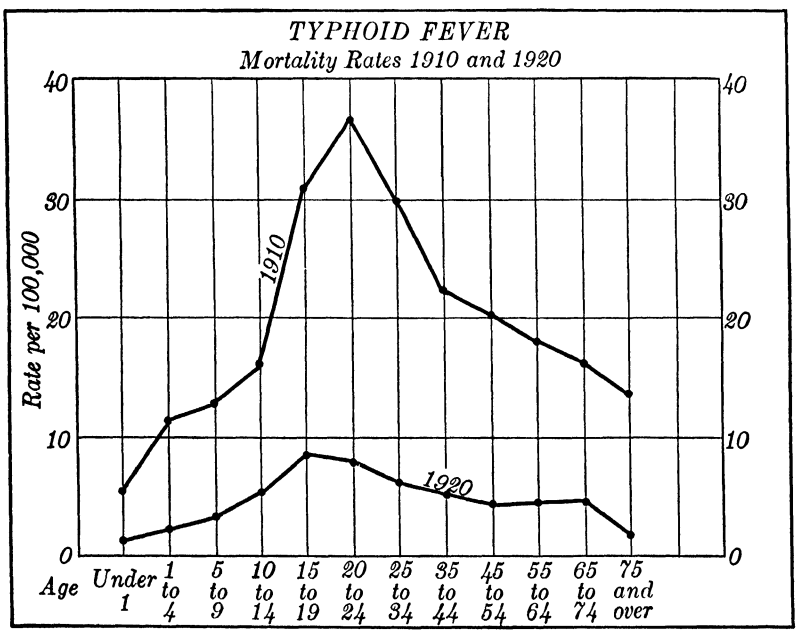
a few would have been found with grades less than any in our sample and a few with grades larger than any in our sample. Both the histogram and the polygon show graphically the outstanding facts of the sample considered. If one is interested in the sample only, this representation is sufficient. However, the purpose of the investigation would usually be to answer certain questions regarding the larger group, *the parent population*, of all the grades at this institution in college algebra. The frequency polygon for the parent population would resemble very closely a smooth curve.

If the class interval be made smaller and smaller and the total frequency, N, be increased without limit, the limit approached by the histogram and the frequency polygon is termed a *frequency curve.* In Chart 1, a frequency curve has been drawn.

It should be borne in mind that this frequency curve brings out the *general tendencies of the parent population* by means of what we have assumed to be a representative sample. This curve meets the base line near the same points at which the frequency polygon meets it; it rises, slowly at first, to a maximum, then recedes again to the base line. The curve should be so drawn that the total area under the curve is equal to the total area of the histogram.

The graphical representation of the data of Table 9 brings out in bold relief the outstanding facts that would possibly not be noted by a glance at the table. Since our primary aim here is the comparison of the two sets of mortality rates, we shall superimpose them on the same graph sheet, Chart 3, so that they may be readily compared.[1]

CHART 3



[1] The reader will note that the age divisions of Chart 3 are unequal.

TABLE 9. MORTALITY RATES PER 100,000 POPULATION FOR TYPHOID FEVER IN THE REGISTRATION STATES, 1910 AND 1920 [1]

Age	*Mortality Rate in 1910*	*Mortality Rate in 1920*
Under 1	5.5	1.1
1 to 4	11.9	2.4
5 to 9	13.0	3.5
10 to 14	16.6	5.6
15 to 19	31.2	8.5
20 to 24	37.1	8.0
25 to 34	30.4	6.2
35 to 44	22.1	5.5
45 to 54	20.4	4.7
55 to 64	18.5	4.7
65 to 74	16.9	4.7
75 and on	14.0	2.0

EXERCISES

1. The lengths of a sample of 75 beans were measured to the nearest tenth of a centimeter. The results are shown in the following distribution:

DISTRIBUTION OF LENGTH OF 75 BEANS

Length	*Class Mark* X	*Frequency* $f(X)$
1.45–1.55	1.5	2
1.55–1.65	1.6	4
1.65–1.75	1.7	6
1.75–1.85	1.8	8
1.85–1.95	1.9	12
1.95–2.05	2.0	20
2.05–2.15	2.1	11
2.15–2.25	2.2	9
2.25–2.35	2.3	2
2.35–2.45	2.4	1
	Total	75

Draw the histogram for these data. Connect the mid-points of the tops of the rectangles, complete at the extremes as previously directed, and thereby obtain the frequency polygon. What is the total area of the histogram? of the polygon?

[1] *Mortality Statistics, 1910–1920,* United States Bureau of the Census, p. 36.

2. If 1,024 throws are made with 10 coins, theoretically, the following results are "expected":

THEORETICAL FREQUENCIES IN COIN-TOSSING

Number of Heads Turning up X	*Frequency* $f(X)$	*Number of Heads Turning up* X	*Frequency* $f(X)$
0	1	6	210
1	10	7	120
2	45	8	45
3	120	9	10
4	210	10	1
5	252	*Total*	1,024

Draw the frequency polygon.

This distribution, we observe, is *symmetrical* with respect to a vertical line drawn through the point (5, 0). While symmetrical distributions never occur in observed data, they are closely approximated in biological and anthropometric measurements. Many educational measurements also result in series that possess remarkable degrees of symmetry.

3. As another example of a series of discrete variates, consider the distribution of the following table:

DISTRIBUTION OF RAYS IN TAIL FINS OF 703 FLOUNDERS[1]

Number of Rays X	*Number of Flounders* $f(X)$	*Number of Rays* X	*Number of Flounders* $f(X)$
47	5	55	111
48	2	56	74
49	13	57	37
50	23	58	16
51	58	59	4
52	96	60	2
53	134	61	1
54	127	*Total*	703

Draw the histogram and a frequency curve for these data.[2]

[1] Paul Riebesell, *Biometrik und Variationsstatistik*, p. 760.

[2] As with all discrete variates, this curve is defined only at the points determined by the data. We draw the curve merely to emphasize the characteristics of the distribution.

4. The data in the following table give the frequencies of the numbers of petals on a certain series of the plant named. They illustrate what is called the J-shaped distribution.

FREQUENCIES OF PETAL NUMBERS,
RANUNCULUS BULBOSUS

Number of Petals X	*Frequency* $f(X)$
5	133
6	55
7	23
8	7
9	2
10	2
Total	222

Plot the histogram and the frequency curve.

17. GRAPHICAL REPRESENTATION OF TEMPORAL DISTRIBUTIONS

The distributions we have thus far considered have dealt mainly with biological and educational data. They have not generally been primarily related to time. The tabular representations have, in general, shown few members at the extremes but they have shown a comparatively large number in the central portions of the tables. The graphical representations, whether by histogram, polygon, or curve, have possessed a common description, namely, low at each end with a maximum near the center. We shall call such distributions *mound-shaped.*

Of the distributions previously considered, some have shown a wide variation or *dispersion,* whereas others have shown moderate variation. Further, they have been more or less unsymmetrically distributed about any line or point. In other words, they have possessed a quality of *asymmetry* or *skewness.* The distribution of algebra grades seemed considerably *peaked* (leptokurtic) near the center. This quality of "peakedness" (or "flatness") is called *kurtosis* and *excess.*[1]

[1] *Kurtosis* by the British school; *excess* by the Scandinavian school.

In the chapters that follow we shall develop measures of these qualities of the distributions. Our present task is the organization and the graphical representation of the data; our next problem will be its algebraical and arithmetical analysis.

Another type of distribution frequently encountered in dealing with economic and mortality data is that in which *time* is the independent variable. Such distributions are called *temporal distributions* or *time series*.

We shall note that time series display a number of distinct types of movement such as long-time trends, seasonal variation, cyclical movements, etcetera. These types of movement call for close examination.

As a first example, consider the growth of population of the United States from 1790 to 1930 inclusive.

TABLE 10. POPULATION: CONTINENTAL UNITED STATES, 1790–1930 [1]

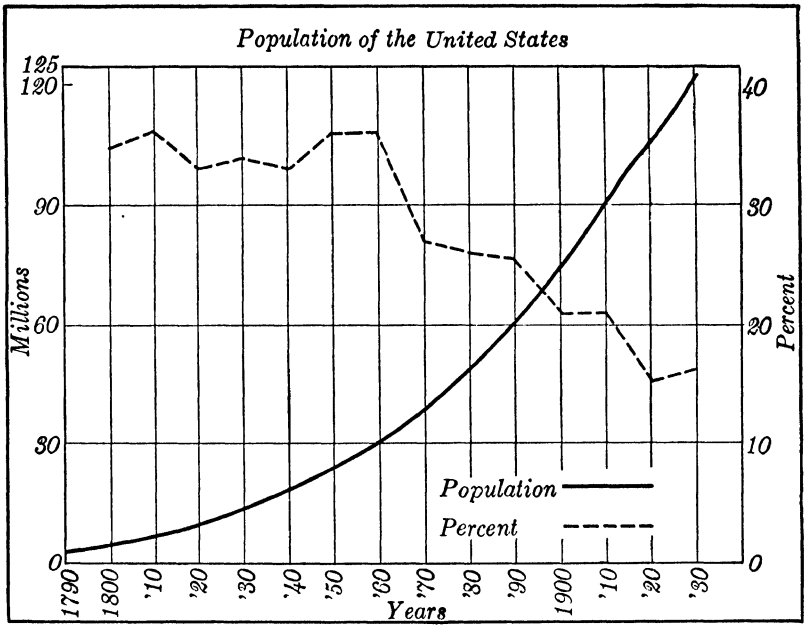
Census Year	*Population (thousands)*	*Per Cent of Increase over Preceding Census*	*Census Year*	*Population (thousands)*	*Per Cent of Increase over Preceding Census*
1790	3,929	—	1870	38,558	26.6 *
1800	5,308	35.1	1880	50,156	26.0 *
1810	7,240	36.4	1890	62,948	25.5
1820	9,638	33.1	1900	75,995	20.7
1830	12,866	33.5	1910	91,972	21.0
1840	17,069	32.7	1920	105,711	14.9
1850	23,192	35.9	1930	122,725	16.1
1860	31,443	35.6			

* Estimated rates are given here.

The graphical representation of these data is shown in Chart 4. We note that the population has enjoyed a steady growth. From 1790 to 1860 each census increased approximately one-third, usually somewhat more, over the preceding; from 1860 to 1890 the decade rates of growth were somewhat over one-fourth, and from 1890 to 1910 a little over one-fifth. Since 1910 the decade rates of increase have been about 15 per cent. Hence we see that, whereas the population has steadily increased, the rate of increase has been steadily decreasing.

[1] The data are taken from the *Fifteenth Census of the United States*, Bureau of the Census, Vol. I, Population, p. 6.

CHART 4



As a second example of time series, consider the following table which gives the production of lumber in the United States in billions of board feet for the given years.

TABLE 11. LUMBER PRODUCTION IN THE UNITED STATES[1]

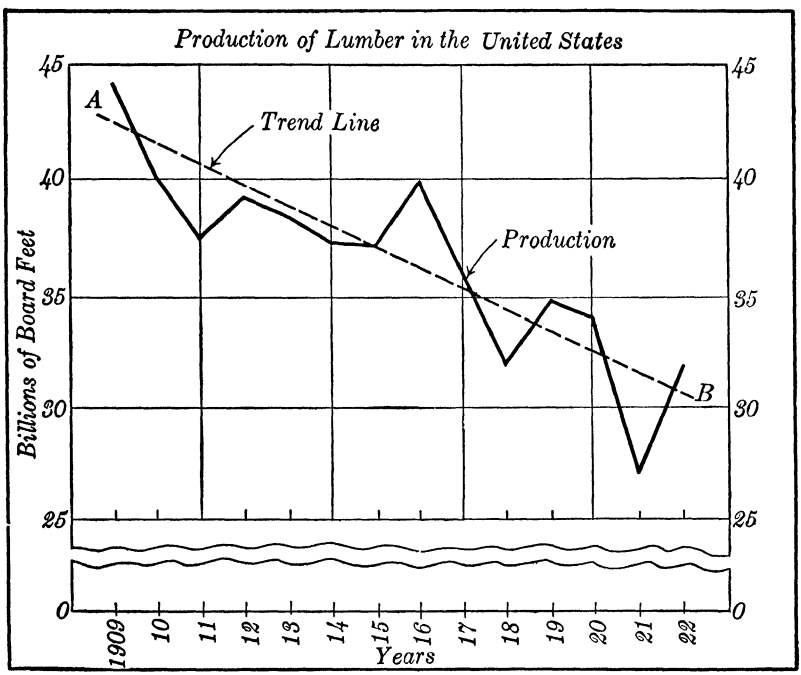
Year	*Reported Production (billions of board feet)*	*Year*	*Reported Production (billions of board feet)*
1909	44.5	1916	39.9
1910	40.0	1917	35.8
1911	37.0	1918	31.9
1912	39.2	1919	34.6
1913	38.4	1920	33.8
1914	37.3	1921	27.0
1915	37.0	1922	31.6

[1] The data are taken from *Statistical Abstract of the United States*, 1928, p. 689.

These data are represented graphically in Chart 5. This is a typical diagram for a historical series. The broken line which represents the production oscillates back and forth on either side of the line of trend which we have estimated graphically. All the points for the production polygon lie within a comparatively narrow strip of which the trend line is the center. Both the trend line and the production polygon emphasize the general diminishing of the production during the years in question. In a later chapter we shall discuss methods for a closer analysis of these data.

Chart 5 affords a good illustration of the possibilities of omitting unimportant areas. In order to give emphasis to the main facts of the data, we obey the instructions of the Joint Committee on Standards for Graphic Presentation[1] to the effect that the zero line should be shown by the use of a horizontal break in the diagram.

CHART 5



[1] A. C. Haskell, *How to Make and Use Graphic Charts*, 1919, p. 71.

The recommendations of this committee should be observed when a single set of data is exhibited. It may not be advisable to carry out the recommendations when two sets of data are placed upon the same graph sheet. We found it possible to do this on Chart 5, but for the data in Table 12, though it is possible, it is inadvisable.

One purpose of a graph is to emphasize outstanding facts, to make evident outstanding relationships. To accomplish this, the proper scales must be selected. The selection of the scales that will give due emphasis to the facts and relationships may not be the scales such that the zero lines for both sets of data can be shown on the diagram.

Consider the data of Table 12. Here we freely omit, without confusing the figure, the zero lines for both sets of data. This table gives the quantity of beef available for consumption per capita per

TABLE 12. BEEF: QUANTITY AVAILABLE PER CAPITA PER ANNUM STEERS: PRICE PER HUNDREDWEIGHT IN DOLLARS [1]

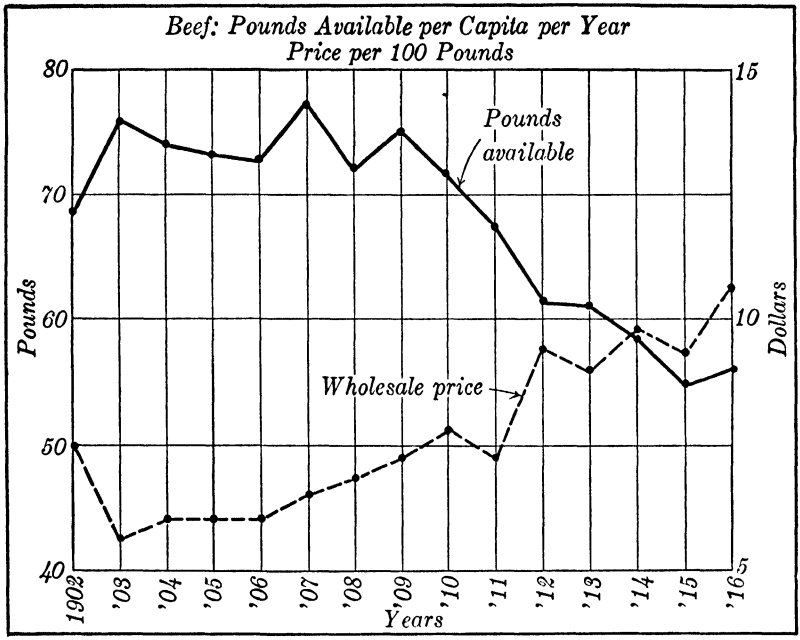
Year	*Beef Available (pounds)*	*Price per Cwt. (dollars)*	*Year*	*Beef Available (pounds)*	*Price per Cwt. (dollars)*
1902	68.5	7.47	1910	71.1	7.77
1903	76.0	5.57	1911	67.7	7.23
1904	73.6	5.96	1912	61.1	9.36
1905	73.0	5.97	1913	60.6	8.93
1906	72.6	6.13	1914	58.5	9.65
1907	77.5	6.54	1915	54.5	9.31
1908	71.5	6.82	1916	56.0	10.42
1909	75.4	7.34			

annum, and the wholesale price of steers per 100 pounds for the given years.

The graphical representation of these data is found in Chart 6. During this fifteen-year period we note that the quantity of beef available per capita per annum has generally decreased, whereas the wholesale price per hundredweight has almost steadily increased. That is, the general trend of the quantity available has shown a downward trend whereas the price has shown an upward trend. The trend for the quantity available seems to be curvilinear, while that for the price seems to be linear. These trends and their relationships will be further analyzed in Chapter 8.

[1] The data are taken from *Yearbook of Agriculture*, 1928, p. 962; United States Bureau of Labor Statistics, *Bulletin* No. 335, p. 38.

CHART 6



18. CUMULATIVE DISTRIBUTIONS AND CURVES

Frequently the chief interest in a frequency distribution is not so much in the items as they are distributed in the several classes as in the accumulated totals of certain of the classes. We may, for example, be chiefly interested in the number of students who receive "more than" or "less than" a given mark; in the number of employees who receive "more than" or "less than" a given wage; in the number of families who receive "more than" or "less than" a given income.

We are thus led to a discussion of *cumulative distributions* and to their graphical representations, known as *cumulative curves.*[1]

Consider, for example, the distribution of Table 13, which illustrates the formation of a "less than" distribution. The column denoted by *Cum.* $f(X)$ gives us the number of the given sample who receive an income less than a given amount, and the column denoted

[1] The cumulative curve is sometimes called an *ogive.*

TABLE 13. DISTRIBUTION OF THE ESTIMATED INCOME AMONG UNMARRIED WOMEN OF THE UNITED STATES IN 1910 [1]

Income (dollars)	*Number $f(X)$*	*Income less than (dollars)*	*Cum. $f(X)$*	*$\frac{\text{Cum. } f(X)}{N}$*
0– 200	10	200	10	0.006
200– 300	70	300	80	0.044
300– 400	560	400	640	0.354
400– 500	530	500	1,170	0.646
500– 600	280	600	1,450	0.801
600– 700	150	700	1,600	0.884
700– 800	110	800	1,710	0.945
800– 900	37	900	1,747	0.965
900–1,000	22	1,000	1,769	0.977
1,000–1,100	16	1,100	1,785	0.986
1,100–1,200	12	1,200	1,797	0.993
1,200–1,300	8	1,300	1,805	0.997
1,300–1,400	5	1,400	1,810	1.000
Total	1,810			

by $\frac{\text{Cum. } f(X)}{N}$ enables us to note the per cent of the total frequency, N, who receive less than a given amount. Thus, of the 1,810 incomes considered in the sample, 640 or 35 per cent received less than $400; 1,600 or 88 per cent received less than $700; 1,769 or 98 per cent received less than $1,000.

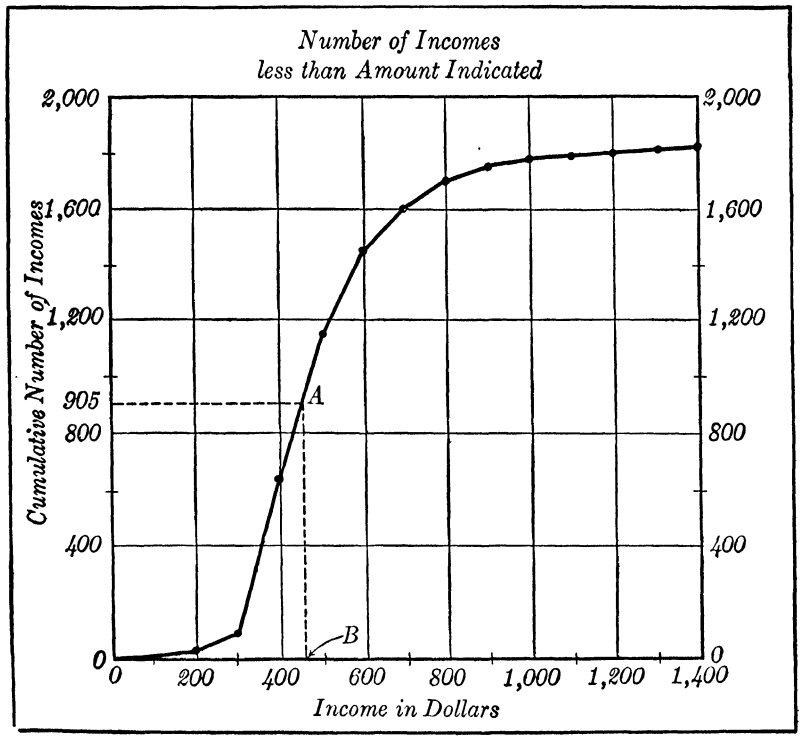
The diagram for the cumulative distribution of Table 13 is constructed by plotting the points (200, 10), (300, 80), etc., as in elementary algebra, and joining them by a broken line as in Chart 7. If a smooth curve be drawn through the points plotted we have a cumulative curve. The graph of $\frac{\text{Cum. } f(X)}{N}$ is precisely coincident with that of *Cum.* $f(X)$ if a proper scale is used for the ordinates.

The cumulative curve is useful for the process of interpolation, that is, for estimating values between those given in the table. Suppose, for example, we desire to know the *income* such that half of the 1,810 have incomes less than it and half have larger incomes. Such an income is called the *median* [2] income.

[1] W. I. King, *Wealth and Income of the People of the United States*, 1915, p. 224.

[2] The median will be discussed in Chapter 3.

CHART 7



We have:

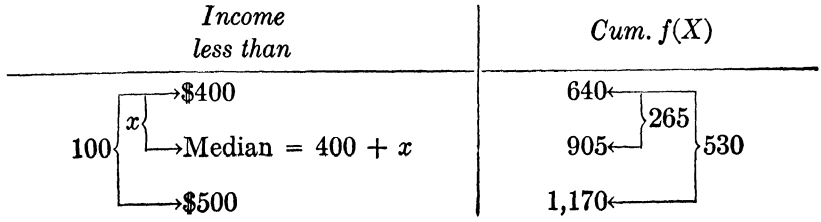
$$N = 1{,}810$$

$$\frac{N}{2} = 905$$

Hence our question is: What is the *income* when the *cumulative number of incomes* is 905?

Mark the point 905 on the vertical scale. Draw through this point a horizontal line which meets the cumulative polygon at A. Draw through A a vertical line which meets the horizontal axis at B for which the *income* is about \$450.

We can check this by simple proportion. From Table 13, we have:



$$\frac{x}{100} = \frac{265}{530}$$

$$x = \$50$$

$$\text{Median} = \$400 + x = \$450 \text{ (Approx.)}$$

In general the proportion is written:

$$\frac{\text{Partial difference in 1st column}}{\text{Total difference in 1st column}} = \frac{\text{Partial difference in 2nd column}}{\text{Total difference in 2nd column}}$$

EXERCISE

Estimate from Chart 7 the *income* such that it is exceeded by exactly three-fourths of the 1,810 incomes. Check your estimate by algebraical interpolation. (You should secure about $367 for your result.)

In a manner similar to the formation of the "less than" distribution we may form a "more than" distribution. While the "less than" distribution proceeds from the least variates and refers to the upper limits of the classes, the "more than" distribution proceeds from the greatest variates to the least and refers to the lower limits of the classes.

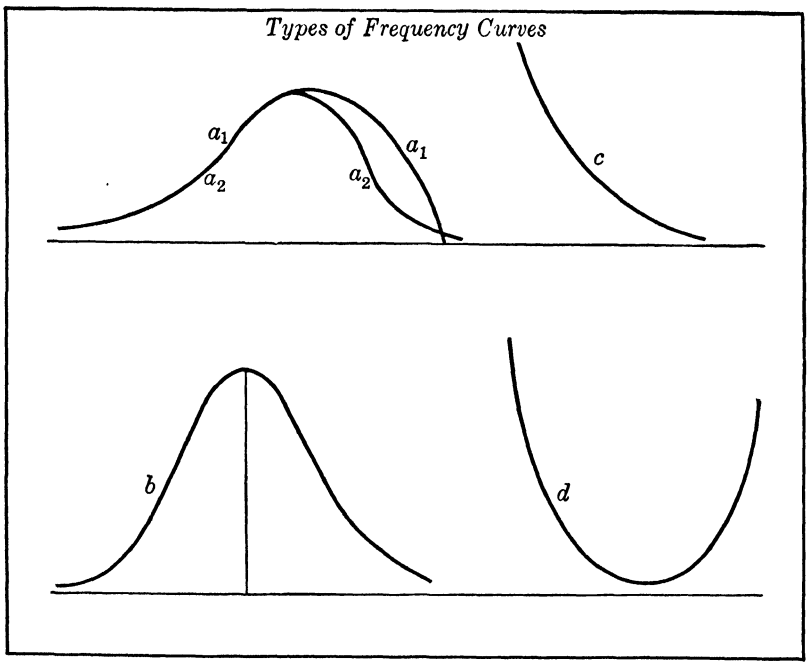
19. TYPES OF FREQUENCY CURVES

As the student proceeds with the graphical representation of frequency distributions he will be impressed with the fact that the graphs of data, even when collected from widely different fields, show certain common characteristics, and can therefore be described as belonging to certain general types; in fact, many of the frequency curves can be closely represented by equations.

The problem of representing the several types of frequency distributions by equations that will best fit the data belongs to the field of advanced statistics, and we desire merely to suggest it at this point. We do, however, wish to describe briefly the general types of frequency curves that are most common.

By far the most common of all is the *moderately asymmetrical* or *mound-shaped distribution.* It occurs in data collected from many fields, such as education, psychology, sociology, economics, biology. The frequencies of this general type increase more or less regularly up to a maximum, and then decrease in the same way. We also note in this type a piling up of cases near the center, that is, a *central tendency.* On Chart 8 we illustrate this type by curves a_1 and a_2.

CHART 8



The second type we may name the *symmetrical distribution,* in which the frequencies decrease uniformly on either side of a line through the center. It is frequently approached in form by data derived from coin- and dice-throwing experiments; from errors of observation in physical measurements; from biological measurements; from educational measurements. The graph for this second type is illustrated by curve *b* on Chart 8.

Many writers call this second type a *normal* or *bell-shaped distribu-*

tion. We much prefer to reserve the name *normal* for the special symmetrical distribution whose equation is

$$y = Ce^{-h^2x^2}$$

and which we shall discuss rather fully in Chapter 12. There are many equations for the curve of the general form we are describing here (see Exercises 6, 7, 8, and 9 at the end of this chapter) but a curve is *normal* only when its equation has the above form.

The third type is the J-*shaped distribution* in which the frequency constantly increases or constantly decreases. The greatest frequency is at one end of the distribution or the other. It is illustrated by curve *c* on Chart 8. [See Exercise 4, page 43.]

The fourth type we shall mention is the U-*shaped distribution* (see curve *d*, Chart 8). It is rarely met. The stock example, familiar to statisticians, is given in Exercise 11 at the end of this chapter.

20. SUGGESTIONS FOR TABULAR AND GRAPHIC PRESENTATION

The process of arranging data into columns and rows in an orderly manner is called *tabulation.* The essential characteristics of a good tabulation are clearness and compactness. While no hard and fast rules can be given to cover all cases of table construction, the following suggestions may be found helpful:[1]

1. The table should have a clear and concise title.
2. The columns and the rows should be arranged in an order that will facilitate comparisons.
3. The columns should have concise headings stating the units of measurement when necessary.
4. The forms should be set off by double lines at the top and the bottom, the sides remaining open.[2]
5. The totals may be placed above or below the detail which they summate.
6. If possible, the source of the data should be given.

In the construction of the charts, we should note especially that:

1. A boundary (picture frame) improves the appearance of the picture.
2. A clear title and subtitle should be in evidence.

[1] A splendid treatment of tabular representation is found in Horace Secrist, *An Introduction to Statistical Methods*, rev. ed., 1925, Chap. VI.

[2] A narrow, compact table may have side lines.

3. The scales should be so selected that the main facts are given due emphasis.
4. The horizontal and vertical scales, with suitable captions, should be easily interpreted.

EXERCISES

1. The following table gives the distributions of heights and weights of 1,515 first-year university men. What are the class boundaries of the several classes in each distribution? Construct the histograms and frequency curves for these distributions. To what general types do these distributions belong?

DISTRIBUTION OF HEIGHTS AND WEIGHTS OF 1,515 MEN

(a)
Heights in Inches

Class Mark X	*Frequency* $f(X)$
58	2
59	1
60	7
61	10
62	26
63	40
64	74
65	142
66	220
67	230
68	258
69	231
70	118
71	99
72	38
73	15
74	2
75	1
76	0
77	1
Total	1,515

(b)
Weights in Pounds

Class Mark X	*Frequency* $f(X)$
95.5	5
105.5	34
115.5	139
125.5	300
135.5	367
145.5	319
155.5	205
165.5	76
175.5	43
185.5	16
195.5	3
205.5	4
215.5	3
225.5	1
Total	1,515

2. The following table gives the distribution of head-breadths of 1,000 Cambridge men, the measurements being taken to the nearest tenth of an inch. Draw the histogram and the frequency polygon for these data. What is the general type of this distribution? Find the class boundaries.

DISTRIBUTION OF HEAD-BREADTHS OF 1,000 MEN [1]

Class Mark X	*Frequency* $f(X)$
5.5	3
5.6	12
5.7	43
5.8	80
5.9	131
6.0	236
6.1	185
6.2	142
6.3	99
6.4	37
6.5	15
6.6	12
6.7	3
6.8	2
Total	1,000

3. In the following table, the average price per bushel is that received by producers December 1.

AVERAGE YIELD AND AVERAGE PRICE OF WHEAT, 1919–1928 [2]

Year	*Average Yield per Acre (bushels)*	*Average Price per Bushel (cents)*
1919	12.8	214.9
1920	13.6	143.7
1921	12.8	92.6
1922	13.9	100.7
1923	13.4	92.3
1924	16.5	129.9
1925	12.9	141.6
1926	14.8	119.8
1927	14.9	111.5
1928	15.6	97.2

Make a chart of these data representing both the average yield and the average price on the same diagram. What can you say about the trends?

[1] The data are taken from *Biometrika*, Vol. I, p. 220.
[2] The data are taken from *Yearbook of Agriculture*, 1928, p. 670.

4. In the following table the average price per barrel is that of Baldwins at Boston.

TOTAL PRODUCTION OF APPLES IN THE UNITED STATES AND AVERAGE PRICE, 1910–1926 [1]

Year	*Production (millions of bushels)*	*Price per Barrel (dollars)*	*Year*	*Production (millions of bushels)*	*Price per Barrel (dollars)*
1910	142	3.68	1919	142	6.71
1911	214	2.56	1920	224	4.02
1912	235	2.28	1921	99	6.69
1913	145	3.95	1922	203	4.84
1914	253	2.08	1923	203	4.02
1915	230	2.36	1924	172	4.78
1916	194	3.44	1925	172	3.92
1917	167	4.40	1926	247	3.22
1918	170	5.94			

Make a chart of these data representing both the production and the price on the same diagram. Point out an important relationship that is emphasized by the graph.

5. Make a chart of the data of the following table. Describe the trend.

DIVORCES IN THE UNITED STATES [2]

Year	*Number of Divorces (thousands)*	*Year*	*Number of Divorces (thousands)*
1890	33.5	1916	112.0
1895	40.4	1922	148.8
1900	55.8	1926	180.0
1905	68.0		

Plot the curve for each of the following equations, and describe the general type to which each curve belongs.

6. $y = \dfrac{100}{2^x + 2^{-x}}$ **7.** $y = \dfrac{100}{x^2 + 2}$ **8.** $y = 50\,(2)^{-\frac{x^2}{2}}$

9. $y = 50\left(1 - \dfrac{x^2}{16}\right)$ **10.** $y = 50\left(1 + \dfrac{x}{3}\right)^4\left(1 - \dfrac{x}{4}\right)^2$

[1] *Loc. cit.*, p. 764.

[2] The data are taken from *Statistical Abstract of the United States*, 1930, p. 92.

11. Draw a histogram to represent the data of the following table.

FREQUENCIES IN DAYS OF ESTIMATED INTENSITIES OF CLOUDINESS AT BRESLAU, 1876–1885

Cloudiness	*Frequency*
0	751
1	179
2	107
3	69
4	46
5	9
6	21
7	71
8	194
9	117
10	2,089
Total	3,653

12. The following table gives the annual production of Portland Cement in the United States. *Statistical Abstract of the United States*, 1930, p. 785. Construct a broken-line diagram for these data. Is the general trend upward or downward? linear or curvilinear?

PORTLAND CEMENT PRODUCTION

Year	*Production (Millions of barrels)*	*Year*	*Production (Millions of barrels)*
1910	77	1920	100
1911	79	1921	99
1912	82	1922	115
1913	92	1923	137
1914	88	1924	149
1915	86	1925	161
1916	92	1926	165
1917	93	1927	173
1918	71	1928	176
1919	81	1929	171

13. The following table gives the annual production of cigarettes in the United States. Construct a broken-line diagram for these data. Is the trend of production linear or curvilinear?

CIGARETTE PRODUCTION

Year	*Annual Production (Billions)*	*Year*	*Annual Production (Billions)*
1920	47.4	1925	82.2
1921	52.1	1926	92.1
1922	55.8	1927	99.8
1923	66.7	1928	108.7
1924	72.7	1929	122.3

CUMULATIVE REVIEW

1. Name several fields of investigation that make use of the statistical method.

2. Name the four steps in the solution of a statistical problem and state briefly what each means.

3. What is meant by variation in statistical data?

4. Define continuous variates; discrete variates. Illustrate.

5. What is meant by "error in a measurement"? The relative error in a measurement? Illustrate.

6. Usually what is the object of a statistical analysis?

7. Distinguish between sample and parent population.

8. Can you think of a problem in which the primary object is a summarized numerical description of the sample only?

9. What letter do we use to designate the total frequency, $\Sigma f(X)$?

10. In the terminology of the text, what do the symbols X, $f(X)$, and N represent?

11. Give directions for constructing a histogram; a frequency polygon.

12. Prove: The total area of a histogram equals the total frequency, N, times the class width, w. That is,

$$\text{Area} = w\Sigma f(X) = wN.$$

13. What is an ogive? Mention several uses of the ogive.

Chapter 3

MEASURES OF CENTRAL TENDENCY

21. INTRODUCTION

It will be recalled that after the collection of the data the next step in the solution of a statistical problem is the organization of the data. The preceding chapter has been devoted to the problem of organization of the data and its graphical analysis. This brings us to the third step, the numerical analysis of the data. We shall find it necessary to devote several chapters to this important part of our problem.

The primary purpose (see Section 1) of a statistical analysis is to abstract the *relevant* information from a mass of numerical data and to express the results clearly and concisely. We accomplish this purpose by computing certain summarizing numbers, or *averages*, which are simply *statistical constants*, rigidly defined, and which are designed, as Professor Bowley says, "to enable the human mind to comprehend with a single effort the significance of the whole."

Averages may be used not only to give us a concise picture of a large group of numbers, but they may be used also to compare different groups, to obtain important facts about a large universe (the parent population) from the measurements of a sample, to measure the relationship between different groups.

The present chapter will be devoted to the averages which measure central tendency.[1] We shall give attention to five such measures: (1) the arithmetic mean, (2) the median, (3) the mode, (4) the geometric mean, and (5) the harmonic mean.

While we shall not at this time undertake to judge the relative merits of these measures, we may with propriety mention several criteria by which an average may be fairly judged. Yule has mentioned several properties that an average should possess.[2] He says that an average (1) should be rigidly defined, (2) should be based on all the observations, (3) should be readily comprehensible, (4) should be easily computed, (5) should be affected as little as possible by the

[1] Averages that measure other characteristics will be discussed in succeeding chapters.

[2] Yule and Kendall, *op. cit.*, p. 113.

fluctuations due to sampling,[1] and, finally, (6) should lend itself readily to algebraic treatment.

22. THE ARITHMETIC MEAN, M_X

The *arithmetic mean* of a group of numbers, essentially measurements, is their sum divided by their number. For example, the arithmetic mean of the numbers 3, 5, 8, 13, 6 is given by:

$$A.M. = \frac{3 + 5 + 8 + 13 + 6}{5} = 7$$

In algebraic form, if $X_1, X_2, X_3, \ldots, X_N$ is a set of N variates, their arithmetic mean is given by:

$$M_X = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\Sigma X}{N} \qquad (1)$$

It may happen that many of the variates may be equal. Suppose the grades of 23 students on a certain test were: 96, 92, 92, 85, 85, 85, 85, 76, 76, 76, 76, 76, 76, 65, 65, 65, 65, 60, 60, 60, 50, 50, 40. Of course we can find the arithmetic mean by the above formula and definition, but the arithmetic is simplified if we proceed as follows:

$$M_X = \frac{96(1)+92(2)+85(4)+76(6)+65(4)+60(3)+50(2)+40(1)}{23}$$

$$= \frac{1656}{23} = 72 \text{ centigrade units (c.u.)}$$

The numbers, 1, 2, 4, 6, 4, 3, 2, 1 are the *frequencies* of the grades. We can show this arithmetic mean by simply arranging the above in tabular form, thus giving the frequency distribution.

TABLE 14. FREQUENCIES OF GRADES OF 23 STUDENTS

Grade X	*Frequency* $f(X)$	$Xf(X)$
96	1	96
92	2	184
85	4	340
76	6	456
65	4	260
60	3	180
50	2	100
40	1	40
Total	23	1,656

[1] See Section 37 for an explanation.

$$M_X = \frac{1656}{23} = 72 \text{ c.u.}$$

In general, suppose that X_1 appears $f(X_1)$ times, that X_2 appears $f(X_2)$ times, and so on, and that X_n appears $f(X_n)$ times, then evidently:

$$M_X = \frac{\sum_{i=1}^{n} X_i f(X_i)}{\sum_{i=1}^{n} f(X_i)} = \frac{\Sigma X f(X)}{N} \tag{2}$$

where $N = \Sigma f(X)$ = the total frequency = the number of the measures. The table headings for formula (2) should be

X	$f(X)$	$Xf(X)$

as is illustrated by Table 14.

At this point a few words about our notation are in order. We have indicated our measurements, scores, etc. by the upper case X. Since the arithmetic mean is generally called the mean, we may naturally represent the mean by M. The subscript X gives emphasis to the fact that we are averaging N values of X. If the original items are indicated by Y, or by Z, the corresponding means may be represented by M_Y and M_Z respectively. We shall find it necessary to use the subscript only when dealing with problems of theory or when we wish to emphasize what variable we are averaging. Hence, in general we shall indicate the mean by M without the subscript.

In the preceding chapter, when discussing the formation of frequency distributions, our attention was directed to two important assumptions that we must make regarding our data. We assume:

1. That in any class the measures are uniformly distributed throughout the interval;
2. That the frequency of the class may be concentrated at its mid-point.

We shall see that in most cases the error due to grouping is relatively slight and that even this can frequently be adjusted by certain corrections.[1]

[1] Sheppard's Corrections, Section 43 of this volume.

It is evident, as indicated above, that the use of formula (2) for computing M requires columns for X the class mark, for $f(X)$ the frequency, and for $Xf(X)$. The column for the class intervals may or may not be included at the pleasure of the computer.

As another illustrative example, we shall compute M for the distribution of grades in college algebra previously exhibited in Table 8. (Note the application of assumption 2 above.)

TABLE 15. GRADES IN COLLEGE ALGEBRA: COMPUTING M

X	$f(X)$	$Xf(X)$
95	4	380
90	6	540
85	12	1,020
80	19	1,520
75	37	2,775
70	24	1,680
65	11	715
60	6	360
55	4	220
50	2	100
Total	125	9,310

$$M = \frac{9310}{125} = 74.48 \text{ c.u.}$$

The sum of the original grades in Table 6 is 9,313, thus giving the arithmetic mean from the ungrouped data to be 74.504. In either case if the values are rounded to one decimal place we have:

$$M = 74.5 \text{ c.u.}$$

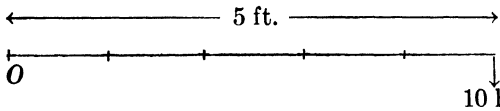
The extreme closeness of the two results is accounted for by our choosing as mid-points of the class intervals the values 60, 65, 70, 75, etc., at which the original data were heavily loaded.

23. THE ARITHMETIC MEAN AS A MOMENT

The term *moment* is one which the statisticians have borrowed from the subject of mechanics, where "the moment of a force is its

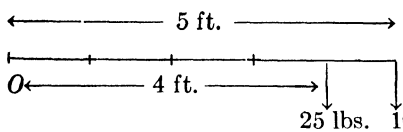
tendency to produce rotation." Thus if we have a weight of 10 pounds suspended from a horizontal bar at a point 5 feet from the fulcrum O,

DIAGRAM 2



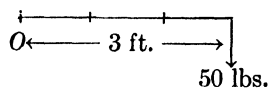
the *first moment* [1] of the force about O is $10 \times 5 = 50$ foot pounds, which is the tendency of the force to produce clockwise rotation about O. If we have two weights of 25 pounds and 10 pounds suspended at distances of 4 feet and 5 feet respectively from and on the same side of O, the total first moment of the two forces about O is $25 \times 4 + 10 \times 5 = 150$ foot pounds, which is the tendency of the two forces to produce clockwise rotation about O.

DIAGRAM 3



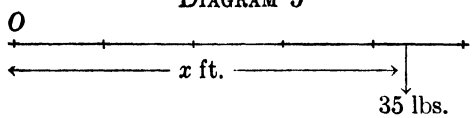
It is evident that a single force of 50 pounds suspended 3 feet from O on the same side would produce the same turning effect. But

DIAGRAM 4



where could we suspend both the 10-pound and the 25-pound weights (or a single 35-pound weight) in order that they would produce the same first moment as the 25-pound and the 10-pound weights when located as above?

DIAGRAM 5



[1] If the weight be multiplied by the square of the distance, the product is called the *second moment* of the force about O.

We evidently have the equation:

$$35x = 150$$

from which

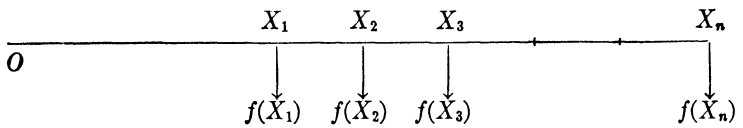
$$x = 4\tfrac{2}{7} \text{ ft.}$$

Let us now consider the frequencies as weights or forces acting at the distances from O determined by their class marks as the figure indicates. The total clockwise turning effect (the first moment) of all the frequencies is:

$$X_1 f(X_1) + X_2 f(X_2) + X_3 f(X_3) + \cdots + X_n f(X_n)$$

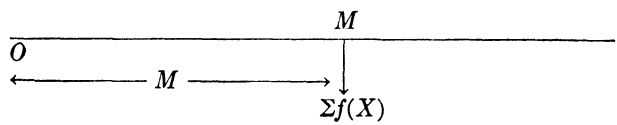
That is, the total first moment of the several frequencies is $\Sigma X f(X)$.

DIAGRAM 6



Now where can the sum of the frequencies, $\Sigma f(X)$ or N, be suspended in order that it may produce the same turning effect? Evi-

DIAGRAM 7



dently at the point M, since by (2):

$$M\Sigma f(X) = \Sigma X f(X)$$

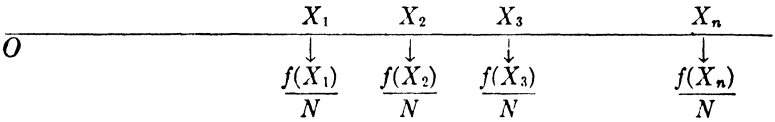
Hence ***M** is that point in the **X** scale at which the total frequency may be suspended so that the first moment of the total frequency about **O** equals the total first moment about **O** of the several frequencies.*

Let us look at this matter now from a *statistical* rather than from a *statical* point of view. The preceding discussion has really been concerned with statical moments. By Theorem II of Section 4 (p. 9), we can write:

$$M = \frac{\Sigma X f(X)}{N} = \Sigma X \frac{f(X)}{N}$$
$$= X_1\frac{f(X_1)}{N} + X_2\frac{f(X_2)}{N} + X_3\frac{f(X_3)}{N} + \cdots + X_n\frac{f(X_n)}{N}$$

If we suspend the quantities $\frac{f(X_i)}{N}$, $i = 1, 2, 3, \ldots, n$, at the points designated by the class marks, it is evident that M is the tendency of the several frequencies, when each is divided by N, to produce rotation about O; that is, ***M is the first statistical moment about O.***

DIAGRAM 8



The nth statistical moment of a frequency distribution about any point A is defined as:

$$\Sigma d_i^n \frac{f(X_i)}{N} = \frac{\Sigma d_i^n f(X_i)}{N}$$

where d_i is the distance from A to X_i and $f(X_i)$ is the frequency corresponding to X_i.

Another simple but important moment property of the arithmetic mean is contained in the theorem: *the first moment of a distribution about the arithmetic mean is zero.*

We shall indicate the deviation of any measure from the arithmetic mean by x. That is, $x_i = X_i - M$. Applying Theorems I and II of Section 4, and formula (2) we have

$$\text{First moment about M} = \Sigma x \frac{f(X)}{N} = \frac{1}{N}\Sigma(X - M)f(X)$$
$$= \frac{1}{N}[\Sigma X f(X) - M\Sigma f(X)]$$
$$= \frac{1}{N}[MN - MN] = 0$$

Of course the corollary immediately follows that

$$\Sigma xf(X) = 0$$

The arithmetic mean is then the "center of balance" or the "center of equilibrium" of the frequencies. That is, it is the point about which the frequencies suspended as weights will balance or be in equilibrium.

Let us examine the formula $\Sigma xf(X) = 0$ for its *algebraic* meaning to statistics. Each x is a *deviation* of a corresponding X *from the arithmetic mean:* that is, $x_i = X_i - M$. Since each x occurs $f(X)$ times, the quantity $\Sigma xf(X)$ gives the algebraic sum of the deviations from the arithmetic mean of measures grouped in a frequency distribution. Thus we have the theorem: *if N measures are arranged in a frequency distribution, the algebraic sum of the deviations from M is zero.* [See Exercise 5 of the next list.]

To illustrate these important properties, consider the following distribution:

TABLE 16

X	$f(X)$	$Xf(X)$	$x = X - 70$	$xf(X)$
82.5	1	82.5	12.5	12.5
77.5	3	232.5	7.5	22.5
72.5	8	580.0	2.5	20.0
67.5	10	675.0	− 2.5	− 25.0
62.5	4	250.0	− 7.5	− 30.0
Total	26	1820.0		00.0

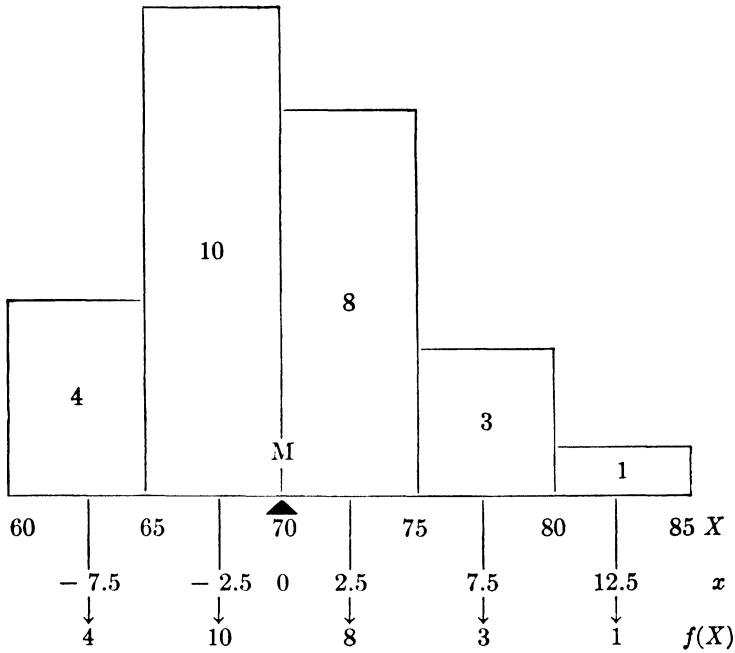
By formula (2) we find

$$M = \frac{1820}{26} = 70$$

The second part of the table follows immediately.

We may see what this theorem means graphically by considering the following diagram. We see that the counter-clockwise moment (turning effect) about $M(= 70)$ is balanced by the clockwise moment about this point for the clockwise moment equals + 55 and the counter-clockwise moment equals − 55. Thus, the point 70 is the center of equilibrium.

DIAGRAM 9



These moment considerations have led some authorities to call the arithmetic mean for grouped data, as defined by Formula (2), *the weighted arithmetic mean.* In contradistinction the arithmetic mean for ungrouped data, as defined by Formula (1), they call *the unweighted arithmetic mean.*

Note. The following list of exercises is given primarily to prepare the student for a facile reading of Section 24. The several exercises should be solved in detail.

EXERCISES

1.

X	$f(X)$
2.5	2
5.0	4
7.5	8
10.0	4
12.5	2

Draw a moment-diagram (see Diagram 8, page 65) for the data of the adjacent distribution. Compute the first statistical moment about 0 of these data.

2. Compute M for the distribution of lengths of beans described in Exercise 1 on page 41. In what unit is M measured?

3. Compute M for the distribution described in Exercise 3 on page 42. In what unit is M measured?

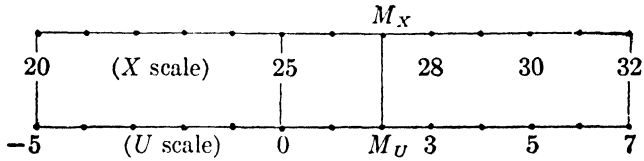
4. Complete the following:

$X_1 = 1$	$X_1 - M$	$= 1 - (4)$	$= (-3) = x_1$
$X_2 = 2$	$X_2 - M$	$= 2 - (\ )$	$= (\ \ ) = x_2$
$X_3 = 4$	$X_3 - M$	$= 4 - (\ )$	$= (\ \ ) = x_3$
$X_4 = 5$	$X_4 - M$	$= 5 - (\ )$	$= (\ \ ) = x_4$
$X_5 = 8$	$X_5 - M$	$= 8 - (\ )$	$= (\ \ ) = x_5$
$\Sigma X = (\ )$	$\Sigma X - 5M$		$= \ 0 \ = \Sigma x_i$
$M = (4)$			

5. In our notation x_i designates the deviation of X_i from M, i.e., $x_i = X_i - M$. Completing the adjacent table we arrive at the theorem: *the algebraical sum of the deviations of a group of numbers from their arithmetic mean is zero.*

$X_1 - M = x_1$	
$X_2 - M =$	
. .	
. .	
. .	
$X_N - M =$	
$\Sigma X - NM = \Sigma x$	
$\Sigma X - N(\ ) = \Sigma x$	
$0 = \Sigma x$	

6. We may frequently save labor in statistical computations by referring the numbers to some new origin. Consider the numbers X: 20, 25, 28, 30, 32. Referred to $X = 25$ as origin these numbers become U: -5, 0, 3, 5, 7.



$$M_U = \frac{\Sigma U}{5} = \frac{-5+0+3+5+7}{5} = \frac{10}{5} = 2$$

on the U scale which corresponds to 27 on the X scale. That is, $M_X = 27$.

7. The U and X in Number 6 are evidently connected by the relation $U = X - 25$, or $X = U + 25$. Replace X by this value, in formula (1), page 60, and show that

$$M_X = \frac{\Sigma X}{5} = 25 + \frac{\Sigma U}{5}$$

8.

X	$U = X - 25$
20	( )
25	( )
28	( )
30	( )
32	( )
	( ) $= \Sigma U$

Complete the table and find M_X from the formula derived in Number 7.

9. Find M_X of the numbers 315, 330, 345, 360, 375, 395, 400, by selecting the new origin at $X = 350$. Proceed as follows:

First: Derive the formula for M_X using $U = X - 350$ and $N = 7$.

Second: Prepare the table and substitute in the derived formula.

10. Find M_X of the numbers 228, 232, 234, 236, 238, 240, 243, 247, by selecting the new origin at $X = 240$.

11. Find M_X of the numbers 215, 230, 245, 260, 275, 295, 300, by selecting the new origin at $X = 250$.

12. Find M_X of the numbers 523, 534, 536, 538, 540, 543, 547, by selecting the new origin at $X = 540$.

13. Can you think of two simple ways to find M_X for the numbers 75, 150, 225, 375? Explain them.

Hint: Let $U = \dfrac{X}{75}$ or $X = 75U$, and proceed.

14. Find M_X of the numbers 128, 256, 384, 512, 640, 768. Note that each number is divisible by 128.

15. Prove that if $U = X - A$ or $X = U + A$, A being a constant, then, using (1),

$$M_X = A + \frac{\Sigma U}{N} = A + M_U$$

16.

Class	X	$f(X)$	$x' = \dfrac{X - 20}{5}$	$x'f(X)$
2.5 – 7.5	5	2		
7.5 – 12.5	10	6		
12.5 – 17.5	15	11		
17.5 – 22.5	20	16		
22.5 – 27.5	25	10		
27.5 – 32.5	30	6		
32.5 – 37.5	35	3		
Totals		54		

(1) Complete columns 4 and 5, and find $\Sigma x'f(X)$.

(2) Note that $x' = \dfrac{X - 20}{5}$ or $X = 5x' + 20$. Replace X by this value in formula (2), page 61, and show that

$$M_X = 20 + \frac{5\Sigma x'f(X)}{54}.$$

(3) Substitute and find M_X.

17. Prove that if $U = \dfrac{X}{k}$, k being a constant, then, using (1),

$$M_X = \frac{k\Sigma U}{N} = kM_U$$

18. If $$X = U + h \quad \text{or} \quad U = X - h,$$

h being a constant, show from equation (2) that:

$$M_X = h + \frac{\Sigma Uf(X)}{N}$$

This transformation is equivalent to moving the origin to the point $(h, 0)$, the unit of measure remaining the same.

19. Let $h = 75$, and use the conclusion in the preceding exercise to find M_X for the data in Table 15. The tabular diagram should be as follows:

M_X FOR TABLE 15 WHEN $h = 75$

X	$f(X)$	$U = X - 75$	$Uf(X)$
95	4	20	80
90	6	15	90
etc.	etc.	etc.	etc.
Total			

20. Let $$x' = \frac{X}{w} \quad \text{or} \quad wx' = X,$$

w being a constant, and show, using (2), that

$$M_X = \frac{w\Sigma x'f(X)}{N}$$

This transformation is equivalent to expressing the variates in class units, the origin remaining the same.

21. Let $w = 5$, and use the conclusion in the preceding exercise to find M_X for the data in Table 15. The tabular diagram should be as follows:

M_X FOR TABLE 15 WHEN $w = 5$

X	$f(X)$	$x' = \frac{X}{5}$	$x'f(X)$
95	4	19	76
90	6	18	108
etc.	etc.	etc.	etc.
Total			

22. Using $h = 54$, and the results of Exercise 18 above, find M_X for the distribution in Exercise 3, page 42.

24. A SHORT METHOD FOR COMPUTING THE ARITHMETIC MEAN

It frequently happens that the distribution under consideration has large values for X, large values for $f(X)$, or large values for both X and $f(X)$, and the consequent arithmetical work for computing M_X and other statistical constants becomes very tedious. In such cases it is convenient, sometimes necessary, to simplify the numbers so that we can save much labor. Three possible steps may be taken. We may:

1. Change the unit of measure as in Exercise 20, page 70.
2. Express the variates as measures from some new origin (frequently called the *provisional mean* or the *guessed mean*) as in Exercise 19, page 70.
3. Combine 1 and 2 to change the unit of measure *and* express the variates as measures from some new origin as in Exercise 16, page 69.

We shall derive the appropriate formula for the third possibility, and show that the others are special cases of it. To do this let: [1]

$$x' = \frac{X - h}{w} \quad \text{or} \quad X = wx' + h$$

where:

h = the distance in original units from O to the new origin O'

w = the class width

x' = the deviation of X from h expressed in *class units*

[1] See Figure 1, page 73.

Then applying Theorems I and II of Section 4 (p. 9), we have:

$$M = \frac{\Sigma Xf(X)}{N} = \frac{\Sigma(wx' + h)f(X)}{N} = \frac{\Sigma wx'f(X)}{N} + \frac{\Sigma hf(X)}{N}$$

$$= \frac{w\Sigma x'f(X)}{N} + \frac{h\Sigma f(X)}{N}$$

$$M = h + \frac{w\Sigma x'f(X)}{N} \tag{3}$$

since $\Sigma f(X) = N$.

The quantity $\frac{\Sigma x'f(X)}{N}$ is usually denoted in statistical work by b_x or by ν_1' (read: nu one prime), and is called the first moment about the arbitrary origin $(h, 0)$, expressed in *class units*. Hence we have:

$$M = h + wb_x = h + w\nu_1' \tag{3a}$$

where

$$b_x = \nu_1' = \frac{\Sigma x'f(X)}{N}$$

If $h = 0$, we get the results previously mentioned in Exercise 20 on page 70, and if $w = 1$, we get the results equivalent to those mentioned in Exercise 18 on page 70. We shall refer to the computation of M by (3) as "the short method of computing M."

Figure 1 on page 73 gives the graphical representation of this development. We shall refer to this figure many times, hence it should be well mastered.

We have here the frequency curve $Y = f(X)$ referred to the axes OX and OY. The point P whose coördinates are $(X, f(X))$ is any point on the curve.

$O' = (h, 0)$ is the arbitrary origin or guessed mean. It should be chosen *at a class mark* near the center of the distribution.

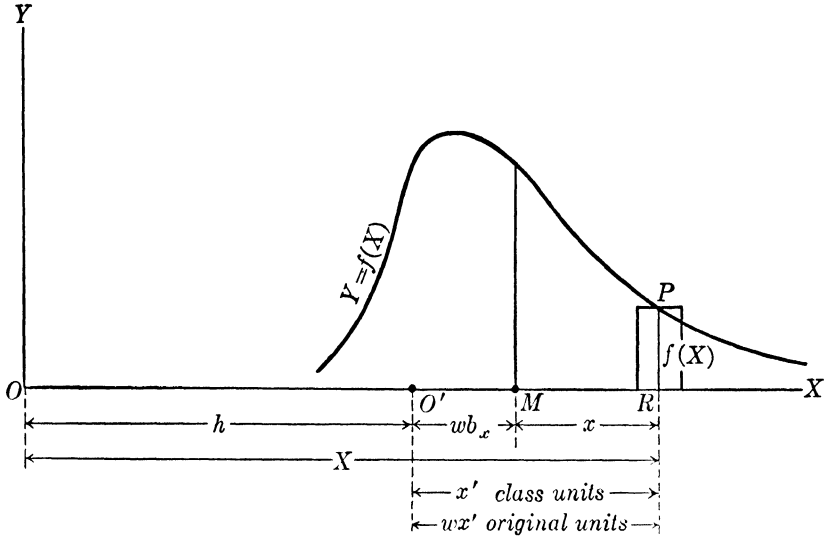
wb_x = the distance from O' to M.

Evidently:

$$X = h + wx'$$
$$OM = M = h + wb_x$$

The distance MR, which is the deviation of any measure from the mean, will be needed in the next chapter. It is represented by small x.

FIGURE 1



Let us apply formula (3) to compute M for the distribution of Table 15.

Let $h = 75$ and $w = 5$. We then have:

$$X = 5x' + 75 \quad \text{or} \quad x' = \frac{X - 75}{5}$$

and Table 15 becomes:

TABLE 17. M FOR TABLE 15 WHEN $h = 75$ AND $w = 5$

X	$f(X)$	$x' = \frac{X - 75}{5}$	$x'f(X)$
95	4	4	16
90	6	3	18
85	12	2	24
80	19	1	19
75	37	0	0
70	24	− 1	− 24
65	11	− 2	− 22
60	6	− 3	− 18
55	4	− 4	− 16
50	2	− 5	− 10
Total	125		− 13

$$b_x = \frac{\Sigma x' f(X)}{N} = \frac{-13}{125}$$

$$M = h + wb_x = 75 + 5\left(\frac{-13}{125}\right) = 74.48$$

EXERCISES

1. Using the short method, compute M for the distribution (a) of heights, Exercise 1, page 54.

2. Compute M of weights, Exercise 1 (b), page 54, by the short method.

3. Compute M for the distribution of head-breadths, Exercise 2, page 54, by the short method.

4. a. Show that M for the first N integers, $1, 2, 3, \ldots, N$ is $(N + 1)/2$.

b. Show that M for the first N odd integers, $1, 3, \ldots, (2N - 1)$ is N.

5. The salaries of 100 male employees of the Smith-Jones Machine Company were arranged into two groups of 40 and 60 men with mean weekly salaries of \$24.96 and \$36.47 respectively. What was the mean salary of the total group?

1st group	*2nd group*	*Total group*
$N_1 = 40$	$N_2 = 60$	$N = 100$
$M_1 = \$24.96$	$M_2 = \$36.47$	$M = (\ )$

6. Twenty-five employees of the Smith-Jones Machine Company earned \$764.38 in a week, and fifteen other employees earned \$638.92 during the same period. What was the mean weekly salary of the forty employees?

7. If in a series of N_1 observations, the arithmetic mean is M_1, and in a second series of N_2 observations, the arithmetic mean is M_2, show that for the entire group of $N = N_1 + N_2$ observations:

$$\text{Combined mean } M = \frac{N_1M_1 + N_2M_2}{N}$$

8. Generalize Exercise 7 above for n groups and show that:

$$\text{Combined mean } M = \frac{N_1M_1 + N_2M_2 + \cdots + N_nM_n}{N} = \frac{\Sigma N_iM_i}{N}$$

where $N = N_1 + N_2 + \cdots + N_n$

9. Prove: $M_{aX} = aM_X$. Illustrate.

10. Prove: $M_{aX+b} = aM_X + b$. Illustrate.

11. The sales record of a certain firm showed the following items: 800 articles at 10 cents; 400 articles at 25 cents; 300 articles at 50 cents. What was the average price per article?

12. The following data taken from Bulletin 435 of the U.S. Bureau of Labor Statistics, "Wages and Hours of Labor in the Men's Clothing

Industry, 1911–1926," give the weekly earnings in 1926 of Hand Sewers on Men's Coats in St. Louis, Cincinnati, and Cleveland. Compute M for each distribution.

Weekly earnings	*Number of Employees*		
	Cincinnati	*Cleveland*	*St. Louis*
\$0 a.u. \$2	1	1	0
2 " 4	1	2	2
4 " 6	2	1	2
6 " 8	2	1	4
8 " 10	6	3	11
10 " 12	14	4	13
12 " 14	27	10	28
14 " 16	15	12	28
16 " 18	19	14	29
18 " 20	15	22	21
20 " 22	9	28	13
22 " 24	7	33	6
24 " 26	7	26	2
26 " 28	4	14	2
28 " 30	2	18	4
30 " 32	2	13	1
32 " 34	3	3	1
34 " 36	1	4	1
36 " 38	3	0	2
38 " 40	0	0	1
Total	140	209	171

13.

Strength (lbs. per sq. in.)	*Number of Bricks*
230– 370	1
380– 520	1
530– 670	6
680– 820	38
830– 970	80
980–1120	83
1130–1270	39
1280–1420	17
1430–1570	2
1580–1720	2
1730–1870	0
1880–2020	1
Total	270

The data of the adjacent table give the transverse strength of bricks in pounds per square inch. They are taken from: *American Society of Testing Materials*, Vol. 33, Part I, p. 458. (Measurements made to nearest 10 pounds.)

Compute M.

25. THE MEDIAN, M_d

A second measure of central tendency, one that has a wide usage in statistical work, is the *median*. Roughly speaking, the median of a set of numbers is the middle one of the set when they are arranged in order of magnitude. Thus, if the set of numbers 33, 93, 45, 83, 72, 97, 21, 67, 91, 46, 82 be arranged in the order of their size: 21, 33, 45, 46, 67, 72, 82, 83, 91, 93, 97, the middle number, 72, is called the median number. Since there are eleven numbers in the above set, the sixth number is the median. In general, if there are N numbers in a set arranged in the order of their size (i.e., "arrayed"), the median number is the one that corresponds to $(N + 1)/2$. If N is even, obviously there is no middle number. In this case the median is commonly taken to be one-half the sum of the two middle numbers. For example, the median of the set 6, 7, 9, 12, 16, 20 is usually taken to be $(9 + 12)/2 = 10.5$.

If the measures are tabulated in a frequency distribution, we shall define the median as the point on the X-scale such that one-half the measures are below it and one-half are above it. On the histogram, frequency polygon, or frequency curve, it is that point on the X-axis at which, if an ordinate is erected, the area of the histogram, polygon, or curve will be bisected. The class interval in which the median is found is called the *median class*.

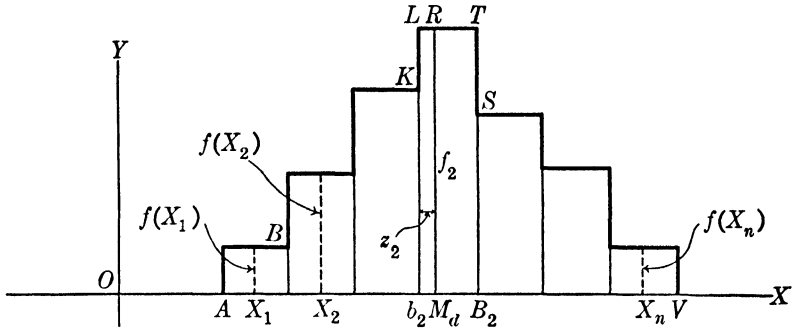
In Section 18 (p. 48) we had a little work in computing the median in simple situations. Let us now derive a formula for finding M_d by looking at the matter from a slighly different point of view.

Let: N = the total frequency
w = the class width
b_2 = the lower class boundary of the median class
B_2 = the upper class boundary of the median class
n_2 = the total frequency of all classes less than b_2
N_2 = the total frequency of all classes greater than B_2
f_2 = the frequency of the median class
z_2 = the distance from b_2 to the median
M_d = the median

Since w is the width of each rectangle and the altitudes are $f(X_1)$, $f(X_2)$, . . ., $f(X_n)$, the area of the histogram is:

$$\text{area} = wf(X_1) + wf(X_2) + \cdots + wf(X_n)$$
$$= w[f(X_1) + f(X_2) + \cdots + f(X_n)] = w\Sigma f(X) = wN$$

FIGURE 2



That is, area wN represents N measures, and therefore

$$\text{area } \frac{wN}{2} \text{ represents } \frac{N}{2} \text{ measures, and}$$

$$\text{area } wn_2 \text{ represents } n_2 \text{ measures.}$$

From the figure we have:

$$ABKb_2 + b_2LRM_d = \frac{wN}{2}$$

or

$$wn_2 + f_2z_2 = \frac{wN}{2}$$

from which we obtain

$$z_2 = \left(\frac{\frac{N}{2} - n_2}{f_2}\right)w$$

Hence the median is given by:

$$M_d = b_2 + z_2 = b_2 + \left(\frac{\frac{N}{2} - n_2}{f_2}\right)w \tag{4}$$

The student should note especially that the value of the median requires the class boundary, not the class mark, of the median class. Once the median class is determined we know immediately $N/2$, n_2, b_2, and f_2. Then computing M_d is decidedly simple.

Our first task, then, in computing M_d is to determine the median class. To do this we find $N/2$, begin at the *lower end of the scale*

and add the frequencies in the successive classes until the lower limit, b_2, of the class containing the median is reached. We then have the median class and, incidentally, n_2.

Next we find $N/2 - n_2$, and observe the frequency of the median class, f_2. We now have all the elements required by formula (4); hence, substituting the values, we find M_d.

Consider the data in Table 18 as an illustrative example.

TABLE 18. COMPUTING M_d FOR SEMESTER GRADES OF 125 STUDENTS IN COLLEGE ALGEBRA

Class	$f(x)$ [1]	
92.5–97.5	4	
87.5–92.5	6	
82.5–87.5	12	
77.5–82.5	19	
72.5–77.5	37 $= f_2$	
67.5–72.5	24	
62.5–67.5	11	
57.5–62.5	6	$\Sigma = n_2 = 47$
52.5–57.5	4	
47.5–52.5	2	
Total	125 $= N$	

We have:

$$w = 5$$
$$N = 125$$
$$\frac{N}{2} = 62.5$$
$$f_2 = 37$$
$$b_2 = 72.5$$
$$n_2 = 47$$

Hence by (4):

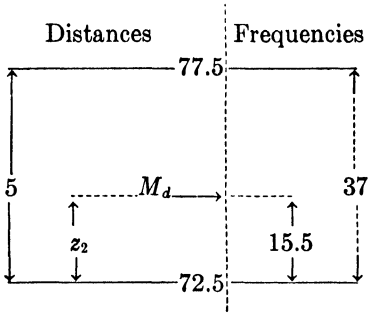
$$M_d = 72.5 + \left(\frac{62.5 - 47}{37}\right)5$$
$$= 74.595 = 74.6 \text{ (approx.)}$$

Employing the assumption made in Section 12 to the effect that the items of a class are uniformly or evenly distributed over the interval, we can find the median by simple interpolation and thus be freed from the tedium of remembering a formula.

Consider the data of Table 8 on page 26. We count from the lower values and determine 72.5–77.5 to be the median class. Below this class are found 2 + 4 + 6 + 11 + 24, or 47, scores. We need to move up the scale above 72.5 a distance z_2 until we obtain 15.5 scores from the 37 scores of the median class, and thus have 47 + 15.5 = 62.5, or $N/2$, scores. By simple proportion we set up the equation for determining z_2 and thus find M_d. The following diagram may assist in understanding the solution.

[1] From this point forward we shall designate the class frequency corresponding to X, x', or x by $f(x)$.

DIAGRAM 10



$$\frac{z_2}{5} = \frac{15.5}{37}$$

$$z_2 = \frac{5(15.5)}{37} = 2.095$$

$$M_d = 72.5 + z_2 = 74.595 \text{ c.u.}$$

EXERCISES

Compute the medians of the following distributions:

1. The distribution of Exercise 1, page 41.

2. The distribution of Table 13, page 49.

3. The distributions (a) and (b) of Exercise 1, page 54.

4. The distributions of Exercise 12, page 74.

5. The distribution of Exercise 13, page 75.

6. Refer to Figure 2, and by equating to $wN/2$ the area $VSTRM_dV$, show that the median is given by:

$$M_d = B_2 - \left(\frac{\frac{N}{2} - N_2}{f_2}\right)w$$

7.

Class	$f(x)$
30 a.u. 33	4
27 " 30	8
24 " 27	16
21 " 24	0
18 " 21	12
15 " 18	12
12 " 15	4
Total	56

According to the definition, what may be the median of the adjacent distribution? At what point would *you* take the median?

8. Compute the median for the distribution of Exercise 3, page 42. Since this is a distribution of discrete data, what interpretation can you give to this median?

26. THE MODE, M_o

A mere glance at Table 8 (p. 26) informs us that the class interval 72.5–77.5 has the greatest frequency. It is called the *modal class.* The class mark, 75, of the modal class is called the *crude mode.*

The *mode* may be roughly defined as the measure that occurs most frequently. The modal height of twelve-month-old white boys is about 29 inches, for there are more twelve-month-old white boys 29 inches high than for any other height. Any haberdasher will tell you that there are more calls for shirts of size 15 than for any other size; hence the modal size for shirts is 15. The mode is the typical measure, the fashionable measure, *la mode.* It is probably what the layman understands as the "average."

The *true mode* is easy to define but very difficult to determine. The true mode is the value of X at which the ideal frequency curve which best fits a set of data has a maximum. Of course the subject of fitting frequency curves in general is beyond the scope of this text, but we may state that the ideal curve for a given distribution is difficult to find.

The mode is roughly approximated by the mid-point of the class with the greatest frequency. We appropriately call this value the crude mode. We obtain a closer approximation to the true mode *by making a correction upon the crude mode.* This correction is made by a process of interpolation. Such interpolation is usually based upon the values that determine the modal class and its two adjacent classes which we choose to call the three "central" classes.

While it is true that for most mound-shaped distributions the mode is in the central part of the distribution, it is not unusual to encounter a mode near one of the extremes of the distribution. When this does occur the mode is certainly an important measure of central tendency.

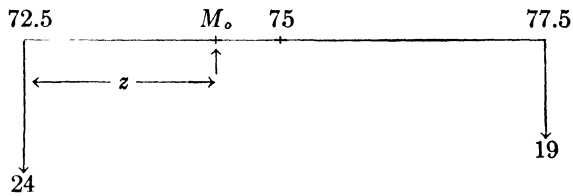
Of the several methods we shall use to determine an approximate mode, probably the method-of-the-parabola is the best. Although the mode, like the median, does not behave beautifully in the algebra of statistics and does not integrate conveniently in the description of the more complex features of statistical phenomena, it deserves a careful consideration. Let us now proceed to the problem of this section, how to find an approximate mode.

Consider the grades in College Algebra, Table 8 (p. 26). The modal class and the two adjacent classes are

X	$f(x)$
80	19
75	37
70	24

There is a well-defined modal class, namely, that with the class mark of 75. Further, since there are 24 members in the 70 class and 19 members in the 80 class, certainly the mode should be drawn from 75 toward 70 because of the added weight of the 70 class. Evidently the mode is located in the 72.5–77.5 interval, the point to be determined by the weights of the adjacent classes. Consider the following diagram.

DIAGRAM 11



Let the frequencies 24 and 19 be considered as weights suspended at the *ends* of the modal class interval. Let z be the amount that must be added to the lower boundary 72.5 to give the approximate mode. In order that the weights shall balance at M_o, we must have:

$$24z = 19(5 - z)$$

from which we obtain

$$z = 2.2$$

and

$$M_o = 72.5 + z = 74.7$$

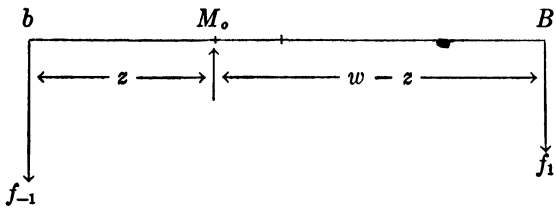
This illustrates the well-known method given by Professor W. I. King in his *Elements of the Statistical Method*, page 124. In general, let:

f_{-1} = the frequency of the class next lower than the modal class
f_1 = the frequency of the class next higher than the modal class
b = the lower boundary of the modal class

B = the upper boundary of the modal class
w = the class width
z = the amount which must be added to b to give M_o

If the frequencies are suspended as weights at the ends of the modal class interval, in order for the weights to balance at M_o, we must have:

DIAGRAM 12



$$f_{-1}z = f_1(w - z)$$

from which

$$z = \left(\frac{f_1}{f_{-1} + f_1}\right) w$$

and

$$M_o = b + z = b + \left(\frac{f_1}{f_{-1} + f_1}\right) w \tag{5}$$

It may be argued that, to be consistent with Section 22 (p. 60), the frequencies should be suspended at the *mid-points* of the respective class intervals. We shall give some exercises at the end of this section that will involve that very point.

A second, and possibly a closer, approximation to the mode can be found by passing a quadratic parabola through the three central points and finding the value of X for which Y or $f(x)$ has a maximum.

The student may recall from elementary or college algebra that

$$Y = aX^2 + bX + c$$

represents a parabola; that it has a maximum if a is negative, a minimum if a is positive, as in Figures 3 and 4. It can be shown in several ways that the coördinates of the bend points, m and M, are:

$$\left(-\frac{b}{2a}, \frac{4ac - b^2}{4a}\right)$$

FIGURE 3

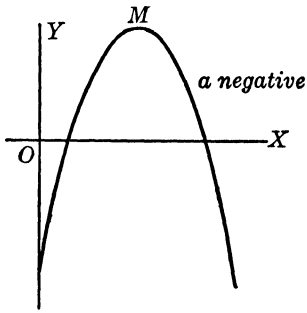
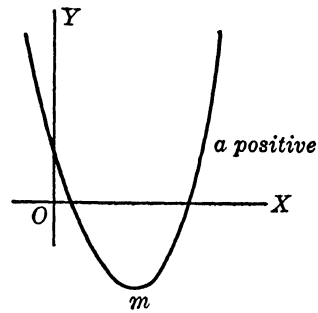


FIGURE 4



That is, if a is negative, the value of X for which $aX^2 + bX + c$ is a maximum is:

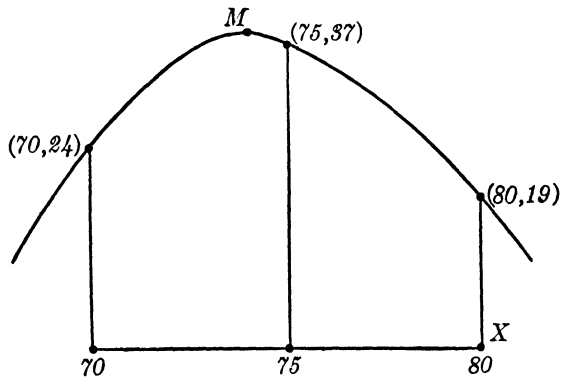
$$X = -\frac{b}{2a}$$

For example, $aX^2 + bX + c$ can be put into the form:

$$a\left(X + \frac{b}{2a}\right)^2 + \frac{4ac - b^2}{4a}$$

If a is negative, the largest value is obtained when $X + \frac{b}{2a} = 0$; that is, when $X = -\frac{b}{2a}$. Similarly, if a is positive, the smallest value is obtained when $X = -\frac{b}{2a}$.

FIGURE 5



Let us apply this method to the distribution of grades in college algebra the three "central" classes for which were given on page 81. When they are plotted they appear as in Figure 5.

We have the three points on the curve as shown. The equation of the curve is:

$$Y = AX^2 + BX + C$$

Substituting the coördinates, we have:

$$24 = A(70)^2 + B(70) + C$$
$$37 = A(75)^2 + B(75) + C$$
$$19 = A(80)^2 + B(80) + C$$

Solving for A, B, and C, we obtain

$$A = -0.62, \quad B = 92.5, \quad C = -3413$$

and the equation of the curve passing through the three given points is:

$$Y = -0.62X^2 + 92.5X - 3413$$

The value of X for which Y is greatest is

$$X = -\frac{B}{2A} = \frac{92.5}{1.24} = 74.597 \text{ c.u.}$$

and this is an approximate mode.

The algebra can be greatly simplified [1] by using the class width ($w = 5$) as a unit and by moving the origin to the point (75, 0) where $h = 75$, *the crude mode*. We then have:

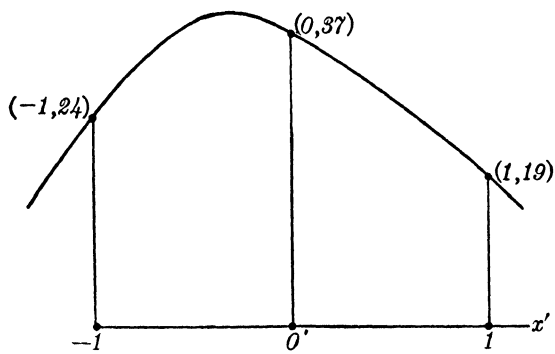
$$X = 5x' + 75 \quad \text{or} \quad x' = \frac{X - 75}{5}$$

and the equation of the curve is now of the form:

$$Y = ax'^2 + bx' + c$$

Figure 6 exhibits the (x', Y) coördinates of the three "central" points.

FIGURE 6



[1] See Exercises 40 and 41, page 110, for solutions based upon determinants.

Substituting the coördinates, we have:

$$24 = a(-1)^2 + b(-1) + c$$
$$37 = a(0)^2 + b(0) + c$$
$$19 = a(1)^2 + b(1) + c$$

Solving for a, b, and c, we have:

$$a = -15.5, \quad b = -2.5, \quad c = 37$$

The value of x' at the mode is

$$x' = -\frac{b}{2a} = -\frac{2.5}{31}$$

and the value of X at the mode is:

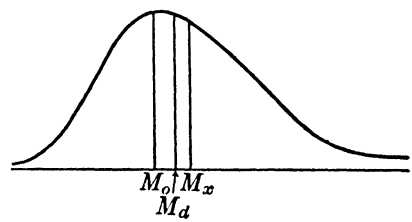
$$X = 5x' + 75 = -\frac{12.5}{31} + 75 = 74.597 \text{ as before.}^1$$

For mound-shaped distributions that are moderately asymmetrical and also possess a moderate peakedness near the center, as in Figure 7, the formula, due to Karl Pearson,

$$M_X - M_o = 3(M_X - M_d)$$

has been found to be approximately true. Since the median and the arithmetic mean are not difficult to compute, this formula may be used to advantage in finding M_o for certain types of distributions. Owing to the fact that the distribution of college algebra grades is very peaked, this formula cannot be expected to check very satisfactorily.

FIGURE 7



EXERCISES

Find the approximate modes by three different methods for each of the following distributions:

1. Of Exercise 1, page 41.
2. Of Exercise 3, page 42.
3. Of Exercise 1(a), page 54.
4. Of Exercise 2, page 54.
5. Assume that the class frequencies f_{-1} and f_1 of the classes adjacent to

[1] The method of determining an approximate mode by passing a quadratic parabola through three points gives the same result as the method of finite differences given by Czuber, *Die statistischen Forschungsmethoden*, p. 71, which is mentioned by Professor Rietz in the *Handbook of Mathematical Statistics*, p. 27.

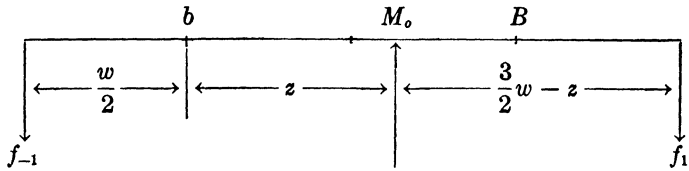
the modal class are suspended as weights at their class marks as in the figure, and show that, if the weights balance at M_o:

$$z = \left[\frac{3f_1 - f_{-1}}{2(f_1 + f_{-1})}\right] w$$

and

$$M_o = b + z = b + \left[\frac{3f_1 - f_{-1}}{2(f_1 + f_{-1})}\right] w$$

DIAGRAM 13



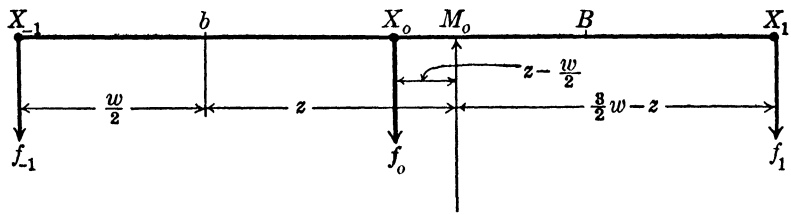
6. Assume that f_o, the frequency of the modal class, and f_{-1} and f_1, the frequencies of the classes adjacent to the modal class, are suspended as weights at their respective class marks, as in the figure, and show that if the weights balance at M_o:

$$z = \left[\frac{f_o + 3f_1 - f_{-1}}{2(f_{-1} + f_o + f_1)}\right] w$$

and

$$M_o = b + z = b + \left[\frac{f_o + 3f_1 - f_{-1}}{2(f_{-1} + f_o + f_1)}\right] w$$

DIAGRAM 14



7. Show that the value of M_o in Exercise 6 above is the arithmetic mean of the modal group and the two groups adjacent to it; that is, show that:

$$M_o = \frac{X_{-1}f_{-1} + X_o f_o + X_1 f_1}{f_{-1} + f_o + f_1}$$

Hint: $X_{-1} = b - \frac{w}{2}$, $X_o = b + \frac{w}{2}$, and $X_1 = b + \frac{3w}{2}$

27. THE GEOMETRIC MEAN, M_g, AND KINDRED TOPICS

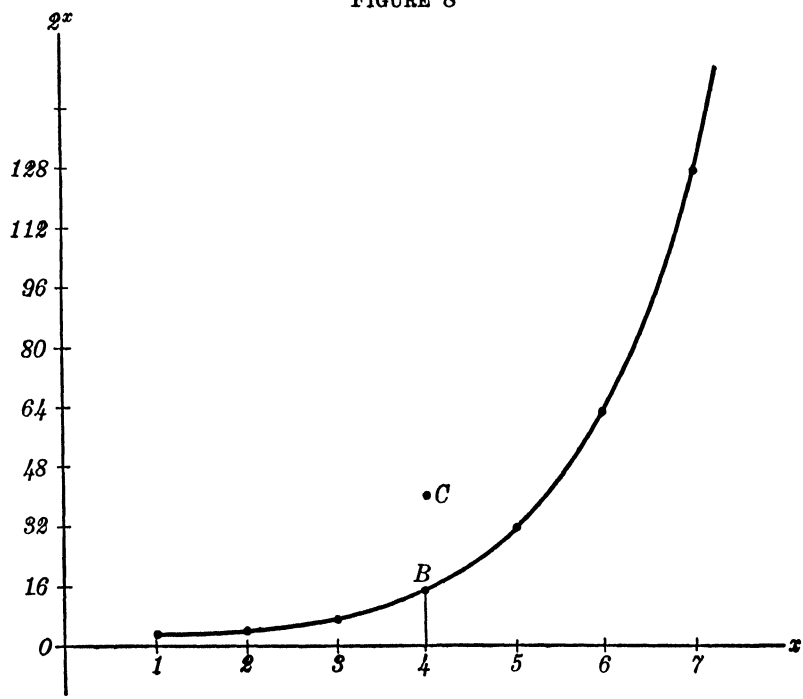
In the preceding pages of this chapter considerable attention has been devoted to the three measures of central tendency that are most widely used — the arithmetic mean, the median, and the mode.

TABLE 19

x	2^x
1	2
2	4
3	8
4	16
5	32
6	64
7	128

Not all data are most logically averaged by these measures. From several points of view the best average for the numbers 2, 4, and 8 is not $4\frac{2}{3}$, their arithmetic mean, but $4 = \sqrt[3]{2 \cdot 4 \cdot 8}$, their geometric mean. The most logical average for the set of numbers 2, 4, 8, 16, 32, 64, 128 is their geometric mean, the seventh root of their product, which is 16. It is a member of the group and, when the data are plotted, is in the curve or *trend* of the data. The geometric mean is represented by the point B, and the arithmetic mean by the point C $(4, 36\frac{2}{7})$.

FIGURE 8



We have learned in college algebra that series in which the quantities increase or decrease at each interval by a constant percentage of the value at the beginning of the interval are in geometric progression. In different words, if the ratio of any number to the preceding number is constant, the numbers are in geometric progression.

It is to such classes of numbers that the geometric mean most logically applies as an average. In observed data it is not expected that the ratio of any number to the preceding number will be absolutely constant; however, if the ratio is *approximately constant, in such data the geometric mean is preferable.*

The geometric mean, then, is widely used in averaging rates of increase or decrease, such as in the study of the growth of any statistical population, growth of skill in an individual, relative changes in the prices of commodities — in short, any data that approximately satisfy the previously stated criterion.

Consider the following table:

TABLE 20. POPULATION OF THE CONTINENTAL UNITED STATES[1]

Year	*Population X (millions)*	*Ratio of Each Item to the One above*
1910	92.0	...
1920	105.7	1.15
1930	122.8	1.16

Since in this particular period of twenty years the ratios are 1.15 and 1.16, essentially constant, we assume that the populations are in geometric progression, and their average would be their geometric mean, namely:

$$M_g = \sqrt[3]{(92.)(105.7)(122.8)}$$

To evaluate this we shall use logarithms, and write:

$$\log M_g = \tfrac{1}{3}[\log 92.0 + \log 105.7 + \log 122.8]$$
$$= \tfrac{1}{3}[1.9638 + 2.0241 + 2.0892]$$
$$= \tfrac{1}{3}[6.0771] = 2.0257$$

and $$M_g = 106.1 \text{ millions}$$

[1] The data are taken from the *Fifteenth Census of the United States*, Vol. I, Population, p. 6.

Further, if we assume that the decade rate of growth will continue for the next decade, then the population for example in 1940 will be $1.16(122.8) = 142.4$ millions.

We have noted that the decade rate of growth from 1910 to 1920 is 1.15 or 115 per cent. Suppose we are interested in the *annual rate of growth*, which we assume is constant during the decade, then we may interpolate the population for the years 1911, 1912, etc.

Let:

r = the annual rate of increase
P_0 = the population in 1910
P_1 = the population in 1911 = $P_0 + P_0 r = P_0(1 + r)$
P_2 = the population in 1912 = $P_1 + P_1 r = P_0(1 + r)^2$

. . . . . .

P_{10} = the population in 1920 = $P_0(1 + r)^{10}$

Therefore:

$$92(1 + r)^{10} = 105.7$$

To solve this equation, we may use logarithms. Hence:

$$\begin{aligned} 10 \log (1 + r) &= \log 105.7 - \log 92 \\ &= 2.0241 - 1.9638 = 0.0603 \\ \log (1 + r) &= 0.00603 \\ (1 + r) &= 1.014 \\ r &= 0.014 = 1.4 \text{ per cent} \end{aligned}$$

Hence:

$$\begin{aligned} P_1 &= P_0(1 + r) = 92(1.014) = 93.3 \text{ millions in 1911} \\ P_2 &= P_1(1 + r) = 93.3(1.014) = 94.6 \text{ millions in 1912} \end{aligned}$$

If we assume the same annual rate to continue from 1920 to 1921, then the population in 1921 is given by:

$$P_{11} = P_{10}(1 + r) = 105.7(1.014) = 107.2 \text{ millions in 1921}$$

We are now ready to define *the geometric mean of N measures to be the Nth root of their product. If* $X_1, X_2, \ldots, X_N$ are the measures, then:

$$M_g = \sqrt[N]{X_1 \cdot X_2 \cdot \cdots \cdot X_N} \tag{6}$$

It is convenient to express this equation in logarithmic form, thus:

$$\log M_g = \frac{\log X_1 + \log X_2 + \cdots + \log X_N}{N}$$

$$= \frac{\Sigma \log X}{N}$$

In other words, the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the original measures.

If the data are arranged in the form of a frequency distribution — that is, if X_1 appears $f(x_1)$ times, X_2, $f(x_2)$ times, and so on — the formula becomes:

$$M_g = \sqrt[N]{X_1^{f(x_1)} \cdot X_2^{f(x_2)} \cdot \ldots \cdot X_n^{f(x_n)}} \tag{7}$$

where $$N = f(x_1) + f(x_2) + \cdots + f(x_n)$$

Suggested Exercise: Using the frequencies in formula (7) as weights, show that the logarithm of a weighted geometric mean is the weighted arithmetic mean of the logarithms of the measures; that is:

$$\log M_g = \frac{\Sigma f(x_i) \log X_i}{N}$$

Example 1. The following table gives for the years indicated the number of divorces in the United States per 1,000 marriages. Find M_g.[1]

Year	*No. of divorces per 1,000 marriages* X	*Log X*
1906	84	1.9243
1916	108	2.0334
1922	131	2.1173
1926	150	2.1761
1931	170	2.2304
		$10.4815 = \Sigma \log X$
		$2.0963 = \frac{\Sigma \log X}{5} = \log M_g$
		$124.9 = M_g$

Example 2. Find the annual rate of increase of the divorce rate for the period 1906–1916. (See Example 1 above.)

[1] Four-place tables of logarithms and anti-logarithms are found in the Appendix.

Solution. Let r_1 be the rate of increase. Following the line of reasoning that was used on page 89, we find

$$84(1 + r_1)^{10} = 108$$

Taking logarithms,

$$\log (1 + r_1) = \frac{\log 108 - \log 84}{10}$$

$$= \frac{2.0334 - 1.9243}{10} = 0.0109$$

$$1 + r_1 = 1.026$$

$$r_1 = 0.026 = 2.6\%$$

Exercise. Find the rates of increase of the divorce rate for the periods 1916–1922, 1922–1926, 1926–1931.

Problem 1. Prepare a skeleton table with the proper headings for finding the geometric mean of a frequency distribution. See formula (7).

Problem 2. A population P_0 *increases* at a constant rate r per period for n periods. Show that the population P_n at the end of n periods is given by

$$P_n = P_0(1 + r)^n$$

Problem 3. A population P_0 *decreases* at a constant rate r per period for n periods. Show that the population, P_n, at the end of n periods is given by

$$P_n = P_0(1 - r)^n$$

Problem 4. If $M_{g,X}$ is the geometric mean of N X's, and $M_{g,Y}$ is the geometric mean of N Y's, then the geometric mean M_g of the $2N$ values is given by

$$M_g^2 = M_{g,X} \cdot M_{g,Y}$$

Problem 5. Plot the curve: $Y = 100(1 + X)^4$.

Problem 6. Plot the curve: $Y = 100(1 - X)^4$.

Problem 7. Prove: $\frac{X_1 + X_2}{2} > \sqrt{X_1 X_2}$.

EXERCISES

1. Complete column 3 for the following data, and note that the ratio is approximately constant. Find the geometric mean for the number of registrations.

REGISTRATION OF MOTOR VEHICLES IN THE UNITED STATES, 1920–1929 [1]

Year	*Number X (thousands)*	*Ratio of Each Item to One above*	Log *X*
1920	9,232		
1921	10,463		
1922	12,238		
1923	15,092		
1924	17,594		
1925	19,937		
1926	22,001		
1927	23,133		
1928	24,493		
1929	26,501		

2. The United States gross imports of crude rubber increased from 252,922 long tons in 1920 to 563,812 long tons in 1929. Find the annual rate of increase during this period, assuming that the rate of growth was constant.

3. The value of a machine *decreases* at a constant rate from the cost price of $1,000 to the scrap value of $100 in ten years. Find the annual rate of decrease, and the value of the machine at the end of one, two, three, years.

4. The number of divorces per 1,000 marriages increased from 62 per 1,000 in 1890 to 166 per 1,000 in 1928. Assuming the annual rate of increase was constant, find its value. (See *Statistical Abstract of the United States*, 1930, p. 91.)

28. THE HARMONIC MEAN, M_h

Another measure of central tendency that is of value in solving certain special types of problems is the harmonic mean. Owing to its unfamiliarity and to the difficulties in interpreting it, it is probably less used than any of the other measures of central tendency, yet for certain problems its use is very desirable.

The *harmonic mean* is defined as the reciprocal of the arithmetic mean of the reciprocals of the given numbers. If the given numbers be $X_1, X_2, \ldots, X_N$, then:

[1] The data are taken from *Statistical Abstract of the United States*, 1930, p. 385.

$$M_h = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \cdots + \frac{1}{X_N}} = \frac{N}{\sum \frac{1}{X}} \tag{8}$$

The harmonic mean of 2, 4, 6 is:

$$\frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{6}} = \frac{36}{11}$$

The harmonic mean is especially useful in averaging *time rates,* in finding the average price per unit when the data give the amount of the commodity for a given price — "so much for a unit of money" — and in the development of index numbers. For example, suppose a man travels two miles, the first at the rate of 10 miles an hour and the second at the rate of 20 miles an hour, what is the average speed? The "obvious" answer of 15 miles an hour is not correct for the man traveled only *two miles* and he consumed only $(\frac{1}{10} + \frac{1}{20})$ of an hour, or $\frac{3}{20}$ of an hour. He would have traveled in $\frac{3}{20}$ of an hour at 15 miles an hour the distance $\frac{3}{20}(15) = 2\frac{1}{4}$ miles, not 2 miles. If r is his average rate in miles an hour, then:

$$(\tfrac{1}{10} + \tfrac{1}{20})r = 2$$

from which we obtain:

$$r = 13\tfrac{1}{3} \text{ miles per hour}$$

and this is *the harmonic mean of the rates.*

As a second illustration suppose a man on a journey purchases gasoline as in the following table:

TABLE 21. PURCHASE OF GASOLINE

Dealer	*Number of Gallons for* \$1 X	*Cost per Gallon (Dollars)*
1	8	$\frac{1}{8}$
2	12	$\frac{1}{12}$
3	10	$\frac{1}{10}$

We wish to find the average price per gallon and the average number of gallons for \$1.

In the preceding illustration we assumed that the average rate times the total time gave the total distance. Similarly we assume

here that the average price times the number of units purchased gives the total cost.

CASE A. Spending the *same amount* with each dealer. Suppose D dollars are spent with each dealer. We then have $(8D + 12D + 10D)$ gallons bought at a total cost of $3D$ dollars. Hence the average price per gallon is $(3D) \div (30D)$, or 10 cents per gallon.

CASE B. Buying the *same amount* from each dealer. Suppose G gallons are purchased from each dealer. We then have $3G$ gallons purchased at a total cost of $(G/8 + G/12 + G/10)$ dollars. Hence the average price per gallon is $(G/8 + G/12 + G/10) \div (3G)$, or 10 5/18 cents per gallon. The reciprocal of this quantity gives the average number of gallons for \$1 and is the harmonic mean of the given X values.

Exercise. A manufacturer of rivets purchased copper as follows: In 1918, 4 pounds for \$1; in 1921, 8 pounds for \$1; in 1925, $6\frac{2}{3}$ pounds for \$1; in 1932, 20 pounds for \$1. Find the average price per pound and the average number of pounds for \$1 on two hypotheses.

The observant student will note that the price of an article may be expressed in two ways,

(a) p units of money per unit of quantity, or
(b) q units of quantity per unit of money.

Thus, the price of sugar may be given as (a) 5 cents per pound or as (b) 20 pounds for a dollar or $\frac{1}{5}$ of a pound for a cent.

Similarly, the speed of a moving body may be expressed as

(a) d units of distance per unit time, or
(b) t units of time per unit distance.

Thus the speed of a car may be given as (a) 30 miles per hour or as (b) 2 minutes per mile.

Of course we are more familiar with prices and speeds expressed in forms (a) but we need to give attention to forms (b) since they do occur. Moreover, the correct average will depend upon how the data are stated, as the previous illustrations confirm.

The following theorems will clarify some of the apparent confusion in which we find ourselves. Note that in the hypotheses and the conclusions of these theorems *we assume that prices and speeds are expressed in the familiar forms* (a).

A. Average Speeds and Rates

If

s = speed, number of units distance per unit time, and

t = number of units time,

then we define:

$$\text{Average speed} = \frac{\text{total distance}}{\text{total time}} = \frac{\Sigma st}{\Sigma t} \tag{9}$$

Theorem I. If the time on each trip is constant, that is, if $t = c$, then

$$\text{Average speed} = \frac{\Sigma st}{\Sigma t} = \frac{\Sigma sc}{\Sigma c} = \frac{c\Sigma s}{Nc} = \frac{\Sigma s}{N} \tag{10}$$

which is the arithmetic mean of the several speeds.

Example 1. A man traveled by auto 3 days. He drove 10 hours each day. He drove

the first day 10 hours at 45 mi. per hr.,
the second day 10 hours at 40 mi. per hr., and
the third day 10 hours at 38 mi. per hr.

What was his average speed?

Solution: Here we have the case in which the time t of each trip is constant and equal to 10 hours. Hence, by (10)

$$\text{Average speed} = \frac{\Sigma s}{N} = \frac{45 + 40 + 38}{3} = 41 \text{ mi. per hr.}$$

The student may verify this by formula (9).

Theorem II. If the total distance covered each trip is constant, that is, if $st = c$, then

$$\text{Average speed} = \frac{\Sigma st}{\Sigma t} = \frac{\Sigma c}{\sum \frac{c}{s}} = \frac{Nc}{c\sum \frac{1}{s}} = \frac{N}{\sum \frac{1}{s}} \tag{11}$$

which is the harmonic mean of the speeds.

Example 2. A man traveled by auto 3 days. He covered 480 miles each day. He drove

the first day 10 hours at 48 mi. per hr.,
the second day 12 hours at 40 mi. per hr., and
the third day 15 hours at 32 mi. per hr.

What was his average speed?

Solution: Here we note that the total distance covered each trip (day) is constant and equal to 480 miles. Hence, by (11)

$$\text{Average speed} = \frac{N}{\sum \frac{1}{s}} = \frac{3}{\frac{1}{48} + \frac{1}{40} + \frac{1}{32}}$$

$$= \frac{3}{\frac{37}{480}} = 38\tfrac{24}{37} \text{ mi. per hr.}$$

The student may verify this by formula (9).

B. Average Prices

Let

p = price per unit (number of units of money per unit of quantity)

q = quantity (number of units purchased at price p)

then we define:

$$\text{Average price} = \frac{\text{total amount spent}}{\text{total quantity purchased}} = \frac{\Sigma pq}{\Sigma q} \tag{12}$$

Theorem III. If the total amount spent at each transaction is constant, that is, if $pq = c$, then

$$\text{Average price} = \frac{\Sigma pq}{\Sigma q} = \frac{\Sigma c}{\sum \frac{c}{p}} = \frac{Nc}{c\sum \frac{1}{p}}$$

$$= \frac{N}{\sum \frac{1}{p}} \tag{13}$$

which is the harmonic mean of the prices.

Example 3. Mr. Jones usually spends \$120 a year for coal. He bought during

the first year 15 tons at \$8 per ton,
the second year 12 tons at \$10 per ton, and
the third year 10 tons at \$12 per ton.

What was the average price of the coal?

Solution: Here we note that the total amount spent each year is constant and equal to \$120. Hence, employing (13) we find the

$$\text{Average price} = \frac{N}{\sum \frac{1}{p}} = \frac{3}{\frac{1}{8} + \frac{\cdot 1}{10} + \frac{1}{12}} = \frac{3}{\frac{37}{120}}$$

$$= \$9\tfrac{27}{37} = \$9.73 \text{ per ton}$$

The student may verify this by formula (12).

Theorem IV. If the same number of units is purchased at each transaction, that is, if $q = c$, then

$$\text{Average price} = \frac{\Sigma pq}{\Sigma q} = \frac{\Sigma pc}{\Sigma c} = \frac{c\Sigma p}{Nc}$$

$$= \frac{\Sigma p}{N} \tag{14}$$

which is the arithmetic mean of the prices.

Example 4. When Mr. Brown purchased gasoline, he regularly purchased 10 gallons. He purchased

at station A, 10 gallons at 14¢ per gal.,
at station B, 10 gallons at 18¢ per gal.,
at station C, 10 gallons at 15¢ per gal., and
at station D, 10 gallons at 13¢ per gal.

What was the average price per gallon?

Solution: In this case we note that the same number of units, 10 gallons, was purchased at each station. Hence, by (14) we obtain

$$\text{Average price} = \frac{\Sigma p}{N} = \frac{14 + 18 + 15 + 13}{4}$$

$$= \frac{60}{4} = 15¢ \text{ per gallon}$$

The student may verify this by formula (12).

EXERCISES

1. A young man took a trip by bicycle. He rode 8 hours each day. He traveled 32 miles the first day, 28 miles the second day, 24 miles the third day, and 20 miles the fourth day. What was his average speed?

2. A man bought four kinds of apples at the following prices:

5 bushels of the first kind at 40¢ per bu.,
5 bushels of the second kind at 50¢ per bu.,
5 bushels of the third kind at 75¢ per bu., and
5 bushels of the fourth kind at $1.00 per bu.

What was the average price per bushel?

3. William Smith purchased gasoline from three dealers. He purchased

from A, 20 gallons at 17¢ per gallon,
from B, 10 gallons at 11¢ per gallon, and
from C, 15 gallons at 15¢ per gallon.

What was the average price per gallon?

4. Three ships make the same round-trip in 20, 24, and 30 days respectively. What was the average number of days for the trip?

5. In a certain factory a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes, and by E in 12 minutes. Find (a) the average number of units per hour, (b) the average number of minutes per unit, and (c) the total number of units they will complete in 8 hours.

6. A man travels 20 miles at 40 miles per hour, 10 miles at 30 miles per hour, and 30 miles at 60 miles per hour. What was his average speed?

7. Five boys were given a page of problems with the instruction to solve as many as they could in an hour. A solved 12, B solved 10, C solved 8, D solved 6, and E solved 4. What was the average number of problems per hour and the average number of minutes per problem?

8. Given two unequal observations X_1 and X_2, prove

$$M_h < M_g < M$$

9. Given three unequal observations X_1, X_2, and X_3, prove

$$M_h < M_g < M$$

10. a. Given N unequal observations $X_1, X_2, X_3, \ldots, X_N$, prove

$$M_h < M_g < M$$

b. How do the three means compare if all the observations are equal in value?

This is a fairly tough problem. For references, see Burgess, *Mathematics of Statistics*, page 101; Chrystal, *Algebra, Part II*, page 46; Hall and Knight, *Higher Algebra*, page 211.

11. Prove that the product of the first N integers is less than $\frac{(1+N)^N}{2^N}$.

Hint. Use Exercise 10a above and Exercise 4a page 74.

12. Prove that the product of the first N odd integers is less than N^N.

Hint. Use Exercise 10a above and Exercise 4b page 74.

29. DISCUSSION AND CRITICISM OF THE MEASURES OF CENTRAL TENDENCY

Owing to the fact that many distributions tend to "pile up" near the center, we have chosen the term *central tendency* to describe this

behavior. The measures of central tendency are statistical constants that give the striking features of the central, the predominant, the typical variates. The arithmetic mean, the median, and the mode are the most widely used. The arithmetic mean is that measure the algebraical sum of the deviations from which is equal to zero. The median is that quantity such that half of the observed measures exceed it in value and half are exceeded by it. We shall see later that the median is the point from which the sum of the absolute values of the deviations of all the measures from it is a minimum. The mode is the value at which the ideal frequency curve fitted to the given distribution has a maximum.

All the measures of central tendency are called *averages*. Since the averages we have thus far considered are so different in their meanings and since we shall meet other averages in succeeding chapters, to the statistician the term *average* is quite indefinite. In the consideration of the first exercise at the end of this chapter the student will have opportunity to observe that such terms as "the average student," "the average-sized apple," et cetera, do not connote the same to all. We should therefore speak definitely as to *which* average is meant when the term is used.

A. The Arithmetic Mean. The *arithmetic mean* is probably the best understood of all the averages. To many people it is *the* average. It is easy to compute, is rigidly defined, is based on all the measures, and is well designed for algebraical manipulation. Arithmetic means of different series can be readily combined to determine the arithmetic mean of the entire group. The arithmetic mean can be determined if the total and the number of the items are known, and is useful in case a large weight is desired for the extreme measures. The arithmetic mean is especially admired for its stability or its reliability. If many samples are drawn from some parent population, the arithmetic means of the given samples will usually show less fluctuation than the other averages. We describe this property by saying that the arithmetic mean is a very reliable or a very stable average.

Situations frequently arise where the emphasis upon the extreme measurements is undesirable. For example, the great wealth of one man in a community will unduly influence the arithmetic average of the wealth of the community, and thus the arithmetic mean will give a distorted picture of the average wealth of the community.

A disproportionately large salary paid to one employee of a group may cause the average salary of the group based upon the arithmetic mean to give an unfair impression of the salaries of the group as a whole. A hundred-dollar bill in a collection plate may cause the arithmetic average of the donations to appear absurdly large.

B. The Median. The *median* is rigidly defined, is easy to determine. It is based upon all the measures, each of which has equal influence, and it is not unduly influenced by the extreme measures. It follows that the median is useful wherever extreme items are of little importance. It is useful in characterizing groups of a nonmathematical character which we cannot measure and yet can arrange them according to size.

The median is not so well understood as the arithmetic mean or the mode, and it is not designed for further mathematical treatment. It shows a greater fluctuation from sample to sample and hence is generally less reliable than the arithmetic mean.

A further objection to the median is its insensitivity. Thus we can replace certain measurements of a given group by other measurements without having any effect upon the median. Let us consider the series 1, 3, 5, 7, 9, 11, 13. For this series we have:

$$M_d = 7 \text{ and } M = 7$$

I may replace, for example, the three numbers which are larger than 7 by three other numbers which are likewise larger than 7 and this replacement will have no effect upon the median. Thus the series 1, 3, 5, 7, 16, 20, 32 has:

$$M_d = 7 \text{ and } M = 12$$

The student will discover other replacements that will in no way affect the median but may have tremendous effect upon the arithmetic mean. Exercise 19 at the end of this chapter is an illustration of the fact that shifting the positions of certain measurements of a group may have no effect whatever upon the median provided the median point is not crossed.

C. The Mode. Though the technical term may not be well known, the concept of the *mode* is well understood and easily comprehended. It is probably what the editors of our newspapers have in mind when they speak of "the average citizen." Like the median, it is not greatly influenced by the extreme variates. Though the

true mode is difficult to determine, yet the term *mode* is so important that even an approximate mode is often satisfactory. An approximate mode is not difficult to determine, and it owes its importance to the fact that it is located in the region where the frequency is most dense. It shows the most frequent measure. For a clothing merchant, the mode of a distribution of chest measurements is *the* important average.

It frequently happens that a distribution has no well-defined mode, or there may be several apparent modes. The mode therefore has no meaning unless there is a decided central tendency. The mode is also insensitive.

D. The Geometric Mean. The *geometric mean* is based on all the measures, is rigidly defined, is suited to algebraic manipulation, is not unduly influenced by extreme measures, and gives equal weight to equal rates of change. It may be appropriately used when emphasis is on the ratio between two quantities rather than on their absolute difference.

The objections to the geometric mean are that it is not well understood, is difficult to compute, and is difficult for the non-mathematical student to comprehend.

It is evident that no one measure of central tendency can be considered as the best. Each measure is useful in shedding some light upon a given problem, and the best selection can be made only by the experienced statistician for the particular purpose he has in mind. The values of the averages considered depend entirely upon the discrimination with which they are used and interpreted. The arithmetic mean is perhaps the most useful. The ease of its computation, its wide uses in later applications, and its familiarity to the general reader make it highly serviceable in statistical work.

EXERCISES

1. What average is meant in each of the following: *the average student? the average citizen? the average amount of material in a dress pattern? the average-sized apple? the average annual rainfall? the average price of wheat? the average ability in arithmetic? the average height? the average length of life? the average speed of a train between two stops? the average salary of teachers in a state? the average number of bushels of corn per acre in a nation?*

2. A college student carries 15 hours of class work per week and makes the grades listed in the following table. What is his average grade?

GRADES OF STUDENT CARRYING 15 HOURS OF CLASS WORK

Course	*Semester Hours Credit*	*Grades in Percentages*
English	2	88
Mathematics	5	96
Language	5	80
Science	3	78
Total	15	

3. A man has \$10,000 invested at 5 per cent, \$5,000 at 6 per cent, and \$3,000 at 8 per cent. What is his average rate of interest?

4. The following are the distributions of the scores of 334 Freshmen on an achievement test in English given at Bucknell University in September, 1929. In (a) the class width is 15, whereas in (b) the class width is 10. What are the class boundaries? Compute M for (a) and (b).

SCORES OF 334 STUDENTS IN ENGLISH

(a)

X	$f(x)$
47	3
62	7
77	15
92	20
107	22
122	37
137	45
152	41
167	55
182	27
197	30
212	11
227	15
242	6
Total	334

(b)

X	$f(x)$
45.5	1
55.5	3
65.5	6
75.5	12
85.5	9
95.5	14
105.5	13
115.5	22
125.5	24
135.5	28
145.5	24
155.5	34
165.5	42
175.5	25
185.5	15
195.5	20
205.5	18
215.5	3
225.5	12
235.5	5
245.5	4
Total	334

5. Compute the medians for the distributions of Exercise 4.

6. Compute the approximate modes by formula (5) for the distributions (a) and (b) of Exercise 4.

7. In the following table deaths from collisions of automobiles with railroad trains and street cars are not included.

AUTOMOBILE FATALITIES IN THE ENTIRE REGISTRATION AREA IN CONTINENTAL UNITED STATES, 1911–1930[1]

Year	*Number of Deaths*	*Year*	*Number of Deaths*
1911	1,291	1921	10,168
1912	1,758	1922	11,666
1913	2,488	1923	14,411
1914	2,826	1924	15,528
1915	3,978	1925	17,571
1916	5,193	1926	18,871
1917	6,724	1927	21,160
1918	7,525	1928	23,765
1919	7,968	1929	27,066
1920	9,103	1930	29,080

Make a chart representing these data. Find the geometric mean of the annual rates of increase. Also find the geometric mean of the number of deaths.

8. The following table gives the deaths from tuberculosis by ages. Note that the class intervals are not all equal. Using the result of the theorem in Exercise 7 on page 74, find the mean age of death from this cause.

DEATHS FROM TUBERCULOSIS BY AGES[2]

Age of Death	*Number Dying* $f(x)$	*Age of Death*	*Number Dying* $f(x)$
0– 4	1,356	30–34	8,776
5– 9	537	35–44	15,456
10–14	1,278	45–54	11,060
15–19	6,300	55–64	7,455
20–24	10,911	65–74	4,788
25–29	10,349	75–84	1,866

[1] The data are taken from *Statistical Abstract of the United States*, 1936, p. 367.

[2] The data are taken from *Mortality Statistics*, 1928, p. 150. The original data have been altered somewhat, e.g., the Bureau's final classification was "75 and over."

9. In Exercise 8 we combined the number of deaths from 0 to 4 inclusive into a single group. We felt justified in doing this because in so doing we did not conceal any outstanding facts. However, in the accompanying table such a procedure would do violence to some outstanding facts.

DEATHS FROM DIPHTHERIA BY AGES [1]

Age of Death	*Number Dying* $f(x)$	*Age of Death*	*Number Dying* $f(x)$
Under 1	602	20–24	89
1– 2	1,133	25–29	56
2– 3	1,183	30–34	67
3– 4	1,112	35–44	110
4– 5	913	45–54	60
5– 9	2,290	55–64	52
10–14	435	65–74	22
15–19	118	75–84	12

Using the result of the theorem in Exercise 8 on page 74, find the mean age of death from this cause. Also find the median age of death.

10. The total number of divorces granted in the continental United States increased from 33,461 in 1890 to 195,939 in 1928. Assuming the annual rate of increase was constant, find its value. From this result, estimate the number of divorces granted in the years 1895, 1900, 1905, and 1925. (See *Statistical Abstract of the United States*, 1930, p. 91.) The actual numbers given are: 1895, 40,387; 1900, 55,751; 1905, 67,976; 1925, 170,952.

11. For \$1 a person purchased each of the following amounts of the given articles:

butter, 3 pounds	potatoes, 40 pounds
sugar, 20 pounds	coffee, 4 pounds

Find the average number of pounds for \$1 and the average price per pound.

12. Prove that the product of the ratios of each of N measures to their geometric mean is equal to unity.

13. Prove that the geometric mean of the ratios of corresponding measures in two series of N measures each is equal to the ratio of their geometric means.

14. The following table gives the number of pounds of sugar that could be bought for \$1 in the given years:

[1] The data are taken from *Mortality Statistics*, 1928, p. 150.

POUNDS OF SUGAR FOR $1, 1918–1922

Year	*Pounds of Sugar Bought for* $1
1918	10.3
1919	8.8
1920	5.2
1921	12.5
1922	13.7

What is the average price per pound during this period? Get two answers.

15. The following distributions are of eggs from Barred Plymouth Rock pullets. The measurements of length were recorded to the nearest millimeter and those of breadth to the nearest half a millimeter. Are these tables consistent with our theory? What are the class boundaries?

LENGTHS AND BREADTHS OF RANDOM SAMPLE OF 450 EGGS FROM 450 PULLETS [1]

(a)

Length in Millimeters X	*Frequency* $f(x)$
49.5	1
50.5	1
51.5	7
52.5	22
53.5	36
54.5	71
55.5	68
56.5	77
57.5	78
58.5	35
59.5	29
60.5	10
61.5	4
62.5	3
63.5	6
64.5	1
65.5	0
66.5	0
67.5	1

(b)

Breadth in Millimeters X	*Frequency* $f(x)$
38.25	2
38.75	4
39.25	9
39.75	18
40.25	41
40.75	52
41.25	41
41.75	65
42.25	73
42.75	48
43.25	41
43.75	26
44.25	15
44.75	7
45.25	5
45.75	2
46.25	1
Total	450

Compute M for the distributions (a) and (b).

[1] The data are taken from Raymond Pearl and F. M. Surface, *A Biometrical Study of Egg Production in the Domestic Fowl*, Part III, p. 183.

16. Find the medians for (a) and (b) in Exercise 15.

17. The number of births during a year is 1/48 of the population at the beginning of the year and the number of deaths during a year is 1/60 of the population at the beginning of the year. Find the number of years for the given population to be doubled.

18. Compute column 3 for these data, and note that the ratio is approximately constant. Find the geometric mean of the expenditures.

EXPENDITURE FOR PUBLIC SCHOOLS IN THE UNITED STATES, 1909–1919

Year	*Expenditure (millions)* X	*Ratio of Each Item to One above*	Log X
1909–1910	401.4		
1910–1911	426.3		
1911–1912	446.7		
1912–1913	482.9		
1913–1914	521.5		
1914–1915	555.1		
1915–1916	605.5		
1916–1917	640.7		
1917–1918	702.2		
1918–1919	763.7		

19. Here are data for two groups of laborers. Find the median wage for each group. Find the arithmetic mean wage of each group. Which is the "better-paid" group?

WAGES OF TWO GROUPS OF LABORERS

Wages per Week (dollars)	*Frequencies*	
	Group A	*Group B*
9.00–9.49	2	2
8.50–8.99	2	2
8.00–8.49	10	10
7.50–7.99	39	39
7.00–7.49	20	1
6.50–6.99	16	16
6.00–6.49	6	6
5.50–5.99	4	4
5.00–5.49	1	20
Total	100	100

20. The population of New York State increased at a constant annual rate from 9,114,000 in 1910 to 10,385,000 in 1920. What was the annual rate of increase? Assuming the same annual rate to continue during the period 1920 to 1925, estimate the population of New York State in 1925. Compare your estimate with the count of the State Census which was 11,161,000.

21. The number of bacteria in a certain culture was found to be $4(10^6)$ at noon of one day. At noon the next day, the number was found to be $9(10^6)$. If the number increased at a constant hourly rate, how many bacteria were there at midnight?

22. The price of an automobile decreased in value at a constant annual rate from \$1,000 to \$300 in five years. What was the annual rate of decrease? What was the value of the car at the end of three years?

23.

Year	*Production (Thousands)* X
1908	65
1909	131
1910	187
1911	210
1912	378
1913	485
1914	569
1915	970
1916	1618
1917	1874

The accompanying table gives the production of motor vehicles in the United States for the years 1908 to 1917 inclusive. Find M and M_g of the production.

24. The production of Portland Cement in the United States increased from 99 million barrels in 1921 to 176 million barrels in 1928. Assuming that the production increased at a constant annual rate, find the average annual rate of increase.

25. The population of Detroit increased at a constant annual rate from 465,700 in 1910 to 993,700 in 1920. What was the average annual rate of increase? Assuming the same annual rate to continue during the period 1920 to 1930, estimate the population of Detroit in 1930. Compare your estimate with the actual census report which gave 1,568,700.

26. If in 1930 the city of Detroit built a water system sufficient to supply a population of 2,500,000, how many years may elapse before the city finds it necessary to enlarge its water system? Base your estimate upon the three census reports. [See Exercise 25.]

Hint: If a is the annual rate of increase the first decade, and b is the annual rate of increase the second decade, the average annual rate is

$$x = \sqrt{(1 + a)(1 + b)} - 1.$$

27. During five successive years a certain investment earned 5 per cent, 6 per cent, 6.5 per cent, 4 per cent, and 3.5 per cent. What was the average annual rate of increase?

28. A does a unit of work in 20 minutes, and B does a unit of work in 24 minutes. What is their average rate of working?

29. The sales record of a certain firm showed the following items: 800 articles at 10 cents; 400 articles at 25 cents; 300 articles at 50 cents. What was the average price per article?

30. A man travels two miles, the first at a miles an hour and the second at b miles an hour. Show that his average rate is

$$\frac{2ab}{a+b}$$

miles an hour. What type of average is this?

31. The annual wages earned by a group of 423 chief wage earners in families are given in the following table. (Houghteling, Leila: *The Income and Standard of Living of Unskilled Laborers in Chicago*, p. 27.) Compute M, M_d, and M_o (by fitting a parabola) for this distribution.

Class	X	$f(x)$
$800– 899		6
900– 999		11
1000–1099		40
1100–1199		50
1200–1299		63
1300–1399		63
1400–1499		81
1500–1599		45
1600–1699		24
1700–1799		20
1800–1899		6
1900–1999		7
2000–2099		2
2100–2199		4
2200–2299		0
2300–2399		1
Total		423

32. Milk is standardized according to its butterfat content. For example, ordinary legal milk is a 3 per cent milk, that is, 3 per cent of its weight is butterfat. If a farmer mixes 8 gallons of 3 per cent milk, 10 gallons of 2.9 per cent milk, 5 gallons of 3.5 per cent milk, 4 gallons of 5.3 per cent milk, what per cent butterfat is the mixture? What type of average is this?

33. Compute M_g for the data of Exercise 12, page 57.

34. Which measure of central tendency would you use to summarize the frequency distribution of the following cases, and why?

(1) Income of parents of Bucknell students.
(2) Amount spent for food by Bucknell students.
(3) Number of hours per week spent in outside preparation by Bucknell students.
(4) Height of Bucknell men.
(5) Weight of Bucknell women.

35. Exercise 12, page 74, gives three distributions of weekly wages in the clothing industry. Find M_o for each distribution.

36. The annual salaries received by a group of Senior federal employees are given in the following table. (White: *Public Administration*, page 290.) Note that the class intervals are not all equal.

Class	X	$f(x)$
\$720 and under \$840		2
840 " " 900		5
900 " " 1000		18
1000 " " 1100		123
1100 " " 1200		369
1200 " " 1320		1208
1320 " " 1440		437
1440 " " 1560		63
1560 " " 1800		74
1800 " " 2000		30
2000 " " 2500		5
Total		2334

Compute M, M_d and M_o (by fitting a parabola) for this distribution. Which average is the most appropriate?

37. A hardware company makes 7% on the invested capital the first year. The profit is added to the original capital, and 9% is made on the total investment the second year. Proceeding in this way, the profits are 10% the third year, 12% the fourth year, and 15% the fifth year. What is the average rate during the five-year period?

38. The value in millions of dollars of exports from the U.S. in the given years are shown in the following table. Compute the geometric mean of the values of the exports.

Year	*Value of Exports*	*Year*	*Value of Exports*
1885	742.2	1905	1518.6
1890	857.8	1910	1745.0
1895	808.5	1915	2768.6
1900	1394.5		

39. A man wishes to travel two miles, the first at 30 miles an hour and the second at such a speed that his average speed over the two mile course will be 60 miles an hour. At what speed must he travel the second mile?

40. (For students who are familiar with determinants.)[1] Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be the coördinates of three points on the parabola $Y = AX^2 + BX + C$. Show that the quotient, $-\frac{B}{2A}$, is given by

$$\begin{vmatrix} X_1^2 & Y_1 & 1 \\ X_2^2 & Y_2 & 1 \\ X_3^2 & Y_3 & 1 \end{vmatrix} \div 2 \begin{vmatrix} X_1 & Y_1 & 1 \\ X_2 & Y_2 & 1 \\ X_3 & Y_3 & 1 \end{vmatrix}$$

Thus, if X_2 is the crude mode and X_1 and X_3 the class marks of the adjacent classes with $X_1 < X_2 < X_3$, the above quotient gives the value of X of the approximate mode.

41. If the class interval is taken as a unit and $x_i' = \frac{X_i - h}{w}$, $i = 1, 2, 3$, show that we may obtain from Exercise 40 above the value of x' of the approximate mode to be

$$\begin{vmatrix} 1 & Y_1 & 1 \\ 0 & Y_2 & 1 \\ 1 & Y_3 & 1 \end{vmatrix} \div 2 \begin{vmatrix} -1 & Y_1 & 1 \\ 0 & Y_2 & 1 \\ 1 & Y_3 & 1 \end{vmatrix}$$

when $x_2' = 0$ is taken at the crude mode.

42. Supposing the frequencies of the X values are the terms of the expansion $(q + p)^n$ as indicated in the table, find M if $(q + p) = 1$.

X	$f(x)$
0	q^n
1	$nq^{n-1}p$
2	$\frac{n(n-1)}{2} q^{n-2}p^2$
⋮	⋮
n	p^n

43. Find the arithmetic, the geometric, and the harmonic means of the numbers: $1, 2, 4, 8, \ldots, 2^n$.

44. Show that the median of the numbers $1, 2, 3, \ldots, n$ is $\frac{n+1}{2}$.

[1] I am indebted to Dr. C. W. Bruce for suggesting this exercise.

Chapter 4

MEASUREMENT OF DISPERSION

30. THE INADEQUACY OF MEASURES OF CENTRAL TENDENCY

The preceding chapters have called attention to the necessity of inventing summary numbers to characterize masses of numerical data. Chapter 3 has dealt with certain terse expressions, single magnitudes, by means of which we may obtain an understanding of the *typical* characteristics of the group as a whole. They represent the acme of condensation. The arithmetic mean, for example, represents an average size of the measures, and is the value such that the algebraic sum of the deviations of the measures from it is zero. All must admit the value of the measures of central tendency, but we must come to realize their insufficiency. Two groups of measures may have the same mean [1] and yet differ widely. Consider the two groups below:

Group I	*Group II*
42	10
45	22
50 (the mean)	50 (the mean)
55	78
58	90

The numbers in Group I are concentrated about their mean, whereas those of Group II are widely scattered. Similarly, we may have two groups of laborers with the same mean salary and yet their distributions may differ widely. The mean salary may not be so important a characteristic as the variation of the items from the mean. To the student of social affairs, the mean income is not so vitally important as to know how this income is distributed. Are a large number receiving the mean income or are there a few with enormous incomes and millions with incomes far below the mean?

[1] In what follows, when the term *mean* is used without a qualifying adjective, the arithmetic mean is meant.

Figures 9, 10, and 11 represent frequency distributions with some of the characteristics we wish to emphasize here. The two curves in (*a*) represent two distributions with the same mean, M, but with

FIGURE 9

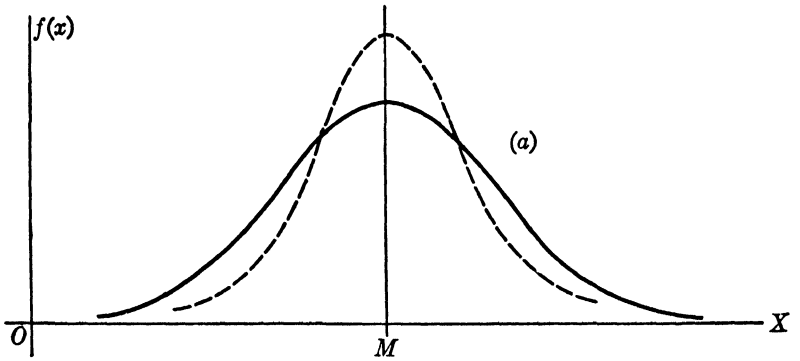


FIGURE 10

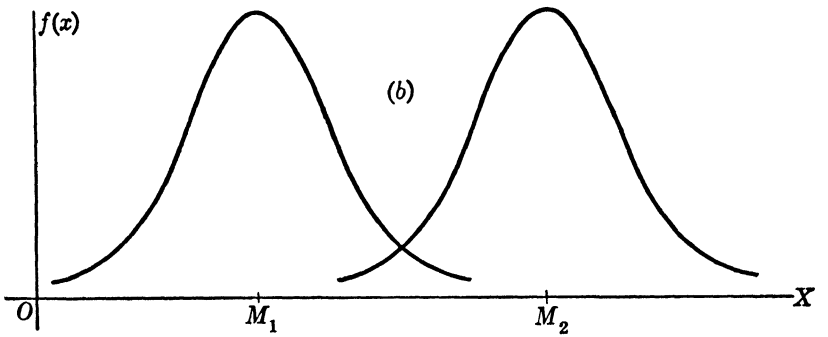
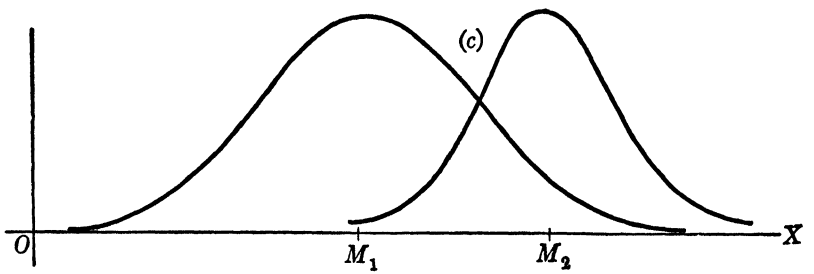


FIGURE 11



different dispersions. The two curves in (*b*) represent two distributions with the same dispersion but with unequal means, M_1 and M_2. Finally, (*c*) represents two distributions with unequal means and unequal dispersions.

The measures of central tendency are therefore insufficient. They must be supported by and supplemented with other measures. In this chapter, we shall be especially concerned with the measures of variability, or spread, or dispersion. A measure of dispersion is designed to state the extent to which the individual measures differ *on the average* from the mean.[1] In measuring dispersion we shall be interested in the *amount* of the variation or its *degree* but not in the *direction*.[2] For example, a measure of 4 inches below the mean has just as much dispersion as a measure of 4 inches above the mean. The amount of variability, or *absolute variability*, will be expressed in concrete units, the same units that are used for the original variates, while the *degree* or *relative variability*, will be expressed in abstract numbers or ratios. A measure of absolute variation is useful in describing a single frequency distribution, but if two different distributions are to be compared, difficulties are encountered.

The real significance of the statements of the paragraph above will be comprehended as we proceed further into the chapter. The computation of the measures of variation for several distributions will convince us that a measure of absolute variation is significant only in proportion to the size of the thing varying. Therefore, for the comparison of the variation in two distributions, we shall find it necessary to define certain measures of relative variability.

There are several measures of absolute variability to which we shall give attention. They are (1) the range, (2) the semi-interquartile range, (3) the mean deviation, (4) the standard deviation, and (5) the probable error.[3] As to measures of relative variability, we shall call attention to several, but we shall express our preference for the *coefficient of variation*, an invention of Professor Karl Pearson.

[1] Generally from the mean; infrequently from other measures of central tendency.

[2] The question of the *direction* of the variation will be answered in Chapter 5 in connection with the *skewness*.

[3] A more extensive treatment of probable error will be found in Chapter 12. Also, see Section 35.

EXERCISES

1. The heights of 11 men were 61, 64, 68, 69, 67, 68, 66, 70, 65, 67, and 72 inches. If the shortest man is omitted, what is the percentage change in the range?

2. The weights of 11 forty-year-old men were 148, 154, 158, 160, 161, 162, 166, 170, 182, 195, and 236 pounds. If the heaviest man is omitted, what is the percentage change in the range?

3. The range of the heights of the 11 men considered in Exercise 1 is $72 - 61 = 11$ inches and the range of the weights of the men considered in Exercise 2 is $236 - 148 = 88$ pounds. Can you determine from this information which shows the greater variation, the 11 measurements of height or the 11 measurements of weight?

4. Find the ratio of the range to the mean in Exercise 1 and Exercise 2. If these ratios are used to measure the relative variations, can you answer the question proposed in Exercise 3?

5. A sample of 1515 college men was measured as to height. Their mean height was found to be 67.9 inches. What would you consider a reasonable variation on either side of the mean for such a set of data?

6. A sample of 1515 college men was measured as to weight. Their mean weight was found to be 138.9 pounds. What would you consider a reasonable variation on either side of the mean for such a set of data?

7.

X	A $f(x)$	B $f(x)$
2.5	1	0
7.5	2	0
12.5	3	0
17.5	5	1
22.5	7	3
27.5	8	14
32.5	9	17
37.5	9	17
42.5	8	14
47.5	7	3
52.5	5	1
57.5	3	0
62.5	2	0
67.5	1	0
Total	70	70

Construct frequency polygons on the same sheet for distributions A and B. Compare their arithmetic means, their medians, and their modes. Do the measures of central tendency constitute a sufficient description of these groups?

31. THE RANGE

The simplest possible measure of the variation of a group of measures is the *range*, that is, the difference between the highest recorded

score and the lowest recorded score. Since the range is determined by only the two extreme measures, it tells us nothing of the distribution between these extremes; it tells us nothing about the concentration of the measures about the center.

Consider the distribution of heights in Exercise 1 (p. 54) and note that the *one man* in the tallest class increases the range about 10 per cent. Such an erratic measure is of little use for purposes of comparison. We need a more stable measure.

32. THE QUARTILE DEVIATION

A measure of variation superior to the range is the quartile range or half of it, the *semi-interquartile range*, sometimes called the *quartile deviation*. The quartiles are the points on the X-scale that divide the distribution into four equal parts. Obviously, there are three quartiles, the second coinciding with the median. More precisely stated, the lower quartile, Q_1, is that point on the X-scale such that one-fourth of the total frequency is less than Q_1 and three-fourths are greater than Q_1. The upper quartile, Q_3, is that point on the X-scale such that three-fourths of the total frequency are below Q_3 and one-fourth is above it. Between Q_1 and Q_3, then, are included one-half the total frequency. Since, under most circumstances, the central half of a distribution tends to be fairly typical, the quartile range Q_3–Q_1 affords a convenient measure of absolute variation. The greater the quartile range, the greater the dispersion.

It is customary to use one-half the quartile range as a measure of dispersion, and to it is given the name of semi-interquartile range. We denote it by Q, and hence:

$$Q = \frac{Q_3 - Q_1}{2} \tag{1}$$

We can determine the quartiles in a manner similar to that used in the determination of the median (see Section 25, p. 76). The class intervals in which the quartiles lie are called the *quartile classes*.

area wN represents N measures

" $\frac{wN}{4}$ " $\frac{N}{4}$ "

Let f_1 and f_3 be the frequencies of the lower and upper quartile classes.
Let b_1 and b_3 be the lower boundaries of these classes.
Let n_1 and n_3 be the accumulated frequencies of all classes below the lower and upper quartile classes respectively.
w = the class width
N = the total frequency
$z_1 = b_1Q_1$
Q_1 = the lower quartile

Then in Figure 12 we have:

$$\text{area } ABCDQ_1 = \frac{wN}{4}$$

$$ABb_1 + b_1CDQ_1 = \frac{wN}{4}$$

$$n_1w + f_1 \cdot z_1 = \frac{wN}{4}$$

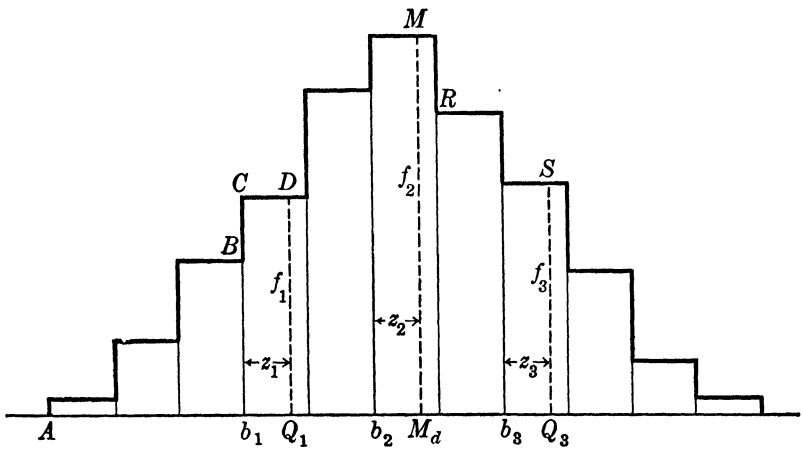
From which:

$$z_1 = \left(\frac{\frac{N}{4} - n_1}{f_1}\right)w$$

and

$$Q_1 = b_1 + z_1 = b_1 + \left(\frac{\frac{N}{4} - n_1}{f_1}\right)w \qquad (2)$$

FIGURE 12



In a similar manner, by equating area $ACMRSQ_3$ to $\frac{3wN}{4}$ we obtain:

$$Q_3 = b_3 + \left(\frac{\frac{3N}{4} - n_3}{f_3}\right)w \tag{3}$$

If the median be designated by Q_2, formula (4) of Section 25 (p. 77) may be written:

$$M_d = Q_2 = b_2 + \left(\frac{\frac{2N}{4} - n_2}{f_2}\right)w$$

and the three formulas may be written in the form:

$$Q_i = b_i + \left(\frac{\frac{iN}{4} - n_i}{f_i}\right)w, \qquad i = 1, 2, 3$$

It should be noted that the determination of Q_1 and Q_3 requires that we know the class boundaries of the classes that contain Q_1 and Q_3.

Therefore, to determine Q_1 we must first locate the class that contains Q_1, the Q_1 class. This done, we will then know $N/4$, n_1, b_1, and f_1. To locate the Q_1 class we find $N/4$, then begin at the *lower end of the scale* and add the frequencies of the successive classes until the lower boundary of the class containing Q_1 is reached. We then know n_1, b_1, and f_1, and thus can immediately find Q_1. A similar statement may be made with regard to Q_3.

The quartile points may also be found by simple analysis without using formulas just as we found the median, Q_2. The method is explained in Exercise 8 of the next appearing list of exercises.

Returning to the data of Table 8 (p. 26), for an illustrative example, we have:

$$\frac{N}{4} = \frac{125}{4} = 31.25$$

$$\frac{3N}{4} = 93.75$$

The quartile class of Q_1 is the class of 67.5–72.5, and the quartile class for Q_3 is the class 77.5–82.5. Hence:

$$b_1 = 67.5 \quad \text{and} \quad b_3 = 77.5$$
$$n_1 = 23 \quad \text{and} \quad n_3 = 84$$
$$f_1 = 24 \quad \text{and} \quad f_3 = 19$$

Then:

$$Q_1 = 67.5 + \left(\frac{31.25 - 23}{24}\right)5 = 69.22 \text{ c.u.}$$

and

$$Q_3 = 77.5 + \left(\frac{93.75 - 84}{19}\right)5 = 80.06 \text{ c.u.}$$

This gives the quartile range to be:

$$Q_3 - Q_1 = 10.84 \quad \text{and} \quad Q = 5.42 \text{ c.u.}$$

In other words, half of the scores occupy a range of 10.84 on the centigrade scale, almost equally distributed on either side of the median. For, by Section 25 (p. 78), $M_d = Q_2 = 74.60$, and therefore:

$$M_d - Q_1 = 74.60 - 69.22 = 5.38 \text{ c.u.}$$
$$Q_3 - M_d = 80.06 - 74.60 = 5.46 \text{ c.u.}$$

As previously stated, the quartiles, and hence Q, are expressed in terms of the original units, but if we divide Q by $\frac{Q_3 + Q_1}{2}$ we have a quartile coefficient of dispersion which may be used to measure relative variation. This coefficient is a ratio, a pure number less than unity in value, and hence in using it we may compare the variations in distributions of unlike units, as distributions of heights in inches with distributions of weights in kilograms. Designating the quartile coefficient of dispersion by V_q, we have:

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{4}$$

In case the distribution is symmetrical:

$$Q_2 - Q_1 = Q_3 - Q_2$$

from which

$$Q_2 = M_d = \frac{Q_3 + Q_1}{2}$$

and

$$V_q = \frac{Q_3 - Q_1}{2Q_2}$$

In this case, the distance from the median to either Q_3 or Q_1 is called the *probable deviation*, sometimes loosely called the *probable error*. In

other words, the *probable deviation* is that distance which if laid off on either side of the median of a symmetrical distribution will include 50 per cent of the measures. If the distribution is not only symmetrical but *normal*[1] (see Section 35, p. 134), this distance is properly called the *probable error*.

EXERCISES

1. Find V_q for the distribution of college algebra grades as described in Table 8 (p. 26). State a use of this result.

2. Find Q and V_q for the distributions of heights and weights as described in Exercise 1 (p. 54). Give meaning to your results.

3. The *deciles* are the points on the X-scale which divide the distribution into ten equal parts. If $b_1, b_2, \ldots, b_9$ be the lower boundaries of the decile classes; $f_1, f_2, \ldots, f_9$ be the frequencies of the decile classes, and $n_1, n_2, \ldots, n_9$ be the accumulated frequencies in all classes below the respective decile classes, and if D_i be the ith decile, show that:

$$D_i = b_i + \left(\frac{\frac{iN}{10} - n_i}{f_i}\right)w, \qquad i = 1, 2, \ldots, 9$$

4. Find the deciles for the distributions of English scores as described in Exercise 4, page 102.

5. Suggest some measures of absolute and relative variability based upon the deciles.

6. The percentiles are the points on the X-scale which divide the distribution into one hundred equal parts. If $b_1, b_2, \ldots, b_{99}$ be the lower boundaries of the percentile classes; $f_1, f_2, \ldots, f_{99}$ the frequencies of the percentile classes; and $n_1, n_2, \ldots, n_{99}$ the accumulated frequencies in all classes below the respective percentile classes, and if P_i be the ith percentile, show that:

$$P_i = b_i + \left(\frac{\frac{iN}{100} - n_i}{f_i}\right)w$$

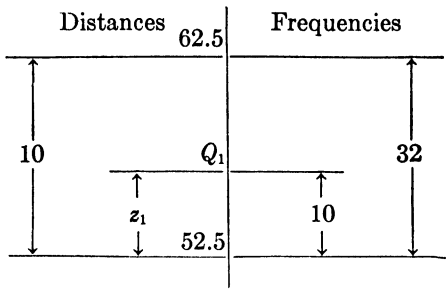
7. Find the fifth, the fifteenth, and the seventy-fifth percentiles for the distribution in Exercise 2, page 54.

8. The quartile points may be determined by simple arithmetic in a manner similar to that used in finding the median. (See p. 79.) Complete the outline on page 120.

[1] A normal distribution is one whose frequency curve is of the type $y = Ce^{-h^2x^2}$. Chapter 12 will be concerned with normal distributions.

Consider the adjacent distribution. By counting from the smaller X-values we determine $52.5 - 62.5$ to be the Q_1 class. Below this class are $4 + 11 = 15$ scores. We need to move up the scale above 52.5 a distance z_1 until we obtain 10 scores from the 32 scores of the Q_1 class, and thus have $15 + 10 = 25$, or $N/4$.

Class	X	$f(x)$
92.5–102.5	97.5	4
82.5– 92.5	87.5	9
72.5– 82.5	77.5	17
62.5– 72.5	67.5	23
52.5– 62.5	57.5	32
42.5– 52.5	47.5	11
32.5– 42.5	37.5	4
Total		100



$$\frac{z_1}{10} = \frac{10}{32} \qquad z_1 = (\ )$$

$$Q_1 = 52.5 + z_1 = (\ )$$

By the method employed here, find Q_3 of this distribution.

9. What are the limiting values of the earnings of the middle half of each distribution of Exercise 12, page 74?

10. Compute Q_1, Q_3, and Q for the distribution of head-breadths given in Exercise 2, page 54. Does $M_d \pm Q$ give values coincident with Q_3 and Q_1? Can you suggest a reason?

11. Find Q_3 and Q_1 for the distribution of Exercise 3, page 42. Since this is a distribution of discrete variates, what meaning can you give to your computed values?

12. What are the limiting values of the earnings of the middle half of the distribution given in Exercise 31, page 108?

13. Does $M_d \pm Q$ for the distribution of Exercise 12 above give values coincident with Q_3 and Q_1? Can you suggest a reason?

14. In what units are the following constants measured: M, M_d, M_o, Range, Q_1, Q_3, Q, and V_q?

15. Derive the formulas for Q_1 and Q_3 by the method of Exercise 8, above.

33. THE MEAN DEVIATION

In the previous chapter we have shown it to be a property of the arithmetic mean that the *algebraic* sum of the deviations from it is zero. The *algebraic* sum of the deviations about any other measure of central tendency will probably be small. Further, we have em-

phasized in the beginning of this chapter that in measuring variability we are interested in the *amount* and not in the *direction* of the variation. And, too, whatever constant is used to measure variability should be one that is based upon all the original measures.

These considerations lead us to define the *mean deviation* as the mean of the absolute values [1] of the deviations of the separate measures from some measure of central tendency. Although the mean deviation is a minimum when taken about the median [2] — which is a splendid argument for insisting upon its being taken about that average — yet it is more frequently taken about the mean. If X is any measure and M the mean, then:

$$M.D. \text{ about } M = \frac{\Sigma \mid X - M \mid f(x)}{N}$$

We have previously designated the deviation of any measure from the mean by x (see Figure 1, p. 73), that is:

$$x = X - M$$

hence

$$M.D. \text{ about } M = \frac{\Sigma \mid x \mid f(x)}{N}$$

Similarly, we may define the mean deviation about the median by the formula:

$$M.D. \text{ about } M_d = \frac{\Sigma \mid X - M_d \mid f(x)}{N}$$

Of course if the numbers are not arranged in a frequency distribution, then considering each frequency as unity we have:

$$M.D. \text{ about } M = \frac{\Sigma \mid x \mid}{N} = \frac{\Sigma \mid X - M \mid}{N}$$

$$M.D. \text{ about } M_d = \frac{\Sigma \mid X - M_d \mid}{N}$$

Corresponding coefficients of relative dispersion may be found by

[1] The magnitude represented by a signed number is called the *absolute value* or the *numerical value* of the number, and is indicated by placing a vertical line on either side of the number. Thus the absolute value of + 5 and of − 5 is 5; in symbols, $\mid + 5 \mid = \mid - 5 \mid = 5$.

[2] Yule and Kendall, *op. cit.*, p. 145.

dividing any mean deviation by the average about which it is taken. Thus:

$$V_{M.D.\text{ about }M} = \frac{M.D.\text{ about }M}{M}$$

For an illustrative example we shall compute the mean deviation about the mean for the distribution of the grades in college algebra. In Section 22 (p. 62) we computed the mean to be:

$$M = 74.48 \text{ c.u.}$$

We then have: $$x = X - M = X - 74.48$$

TABLE 22. COMPUTING *M.D.* ABOUT *M* FOR THE GRADES IN COLLEGE ALGEBRA

X	$f(x)$	$\lvert x\rvert = \lvert X - M\rvert$	$\lvert x\rvert \cdot f(x)$
95	4	20.52	82.08
90	6	15.52	93.12
85	12	10.52	126.24
80	19	5.52	104.88
75	37	0.52	19.24
70	24	4.48	107.52
65	11	9.48	104.28
60	6	14.48	86.88
55	4	19.48	77.92
50	2	24.48	48.96
Total	125		851.12

$$M.D.\text{ about }M = \frac{851.12}{125} = 6.81 \text{ c.u.}$$

$$V_{M.D.\text{ about }M} = \frac{6.81}{74.48} = 0.09143 = 9.1\%$$

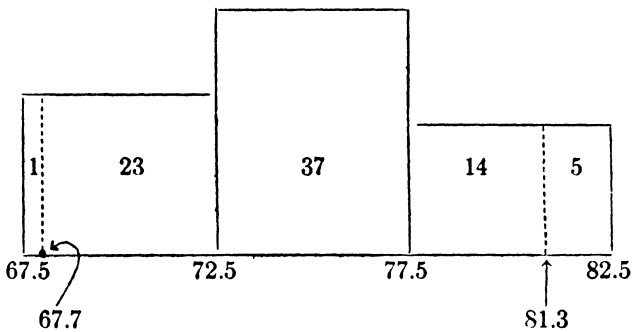
Of the three measures of absolute variability that we have thus far considered, the mean deviation is the only one which has considered the deviations of all the individual members from a given average. The range and the semi-interquartile range are distances that are not based upon the consideration of all the members of the distribution. The mean deviation, however, is based upon all the members of the group, is rigidly defined, is readily computed, and is

not difficult to comprehend. It gives due weight to the extreme items, and is an especially good measure to use with economic data. The artificial step of ignoring the signs of the deviations, of course, renders it useless in further mathematical treatment.

It is a property of approximately normal distributions that the interval

$$M \pm (M.D. \text{ about } M)$$

includes about 58 per cent of the total frequency. For this distribution of college algebra grades this interval extends from 74.5 − 6.8 to 74.5 + 6.8, that is, from 67.7 to 81.3.



By constructing a portion of the histogram of Table 8 and recalling that a class frequency is proportional to the area of the rectangle, we find that

$$\frac{72.5 - 67.7}{5}(24) + 37 + \frac{81.3 - 77.5}{5}(19)$$

or

$$23 + 37 + 14 = 74$$

scores lie in this interval. This is 59 per cent of the total frequency, 125, which checks the theory approximately.

This example illustrates an important function of a measure of dispersion when it is combined with a measure of central tendency. They give a summarized description of the distribution because they make possible the determination of intervals that include rather definite proportions of the total frequency. Thus $M_d \pm Q$ determines an interval that includes about $N/2$ variates and $M \pm M.D.$ determines an interval that includes about $3N/5$ variates.

EXERCISES

1. Find Σx, $\Sigma |x|$, and *M.D.* about M for each set of numbers:

(a)

X	x	$\lvert x\rvert$
3		
5		
13		
20		
27		
58		

(b)

X	x	$\lvert x\rvert$
62		
68		
74		
76		
88		
94		

(c)

X	x	$\lvert x\rvert$
124		
146		
162		
178		
190		
220		

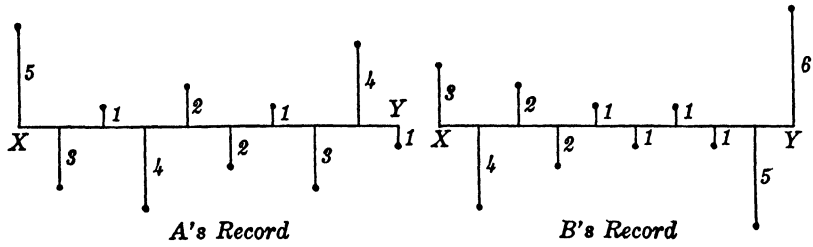
2. Statistical data of the United States Department of Agriculture show the following average yields in bushels per acre for the three specified crops. Compute *M.D.* about *M*.

Year	*Wheat*	*Rye*	*Oats*	*Year*	*Wheat*	*Rye*	*Oats*
1923	13.3	11.3	30.5	1928	15.4	11.7	32.9
1924	16.0	15.0	34.0	1929	13.0	11.4	29.3
1925	12.8	11.3	31.9	1930	14.2	12.8	32.2
1926	14.7	10.3	26.6	1931	16.3	10.4	28.1
1927	14.7	15.1	27.1	1932	13.0	12.2	30.1

3. Complete the following table. Find *M.D.* about *M*. What per cent of the total frequency is included in the interval $M \pm M.D.$?

Class	X	$f(x)$	x'	$x'f(x)$	x	$xf(x)$	$\lvert xf(x)\rvert$
92.5–102.5	97.5	4					
82.5– 92.5	87.5	11					
72.5– 82.5	77.5	32					
62.5– 72.5	67.5	25	0				
52.5– 62.5	57.5	15					
42.5– 52.5	47.5	8					
32.5– 42.5	37.5	5					
Total		100					

4. Each of two marksmen A and B fires 10 shots at a horizontal line XY. Their records are indicated by the following diagrams. Basing your conclusion upon the mean deviation, can you determine who made the better record?



34. THE STANDARD DEVIATION

Unquestionably the most universally used measure of dispersion is the *standard deviation.* It is usually denoted by σ_X (sigma),[1] and is defined as the square root of the mean of the squares of all the individual deviations measured from the arithmetic mean. Expressed as a formula, this definition becomes:

$$\sigma_X = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma (X - M_X)^2}{N}} \tag{5}$$

If the original measures are grouped in a frequency distribution, the definition becomes:

$$\sigma_X = \sqrt{\frac{\Sigma x^2 f(x)}{N}} = \sqrt{\frac{\Sigma (X - M_X)^2 f(x)}{N}} \tag{6}$$

It will be noted that the squaring of the deviations removes the objectionable feature of signs noted in the preceding section when discussing the mean deviation. Further, the squaring gives added weight to the extreme measures, a desirable feature for some types of data. It should also be noted that taking the square root of the mean of the squared deviations leaves σ *expressed in the original unit of measure.*

Formulas (5) and (6) should be learned in several forms, thus:

$$\sigma^2 = \frac{\Sigma x^2 f(x)}{N}, \quad N\sigma^2 = \Sigma x^2 f(x), \text{ etc.}$$

[1] We shall generally omit the subscripts, employing them only when necessary, as in theoretical developments and for purposes of identification. [See p. 61.]

Unless otherwise stated, the standard deviation is always computed with the deviations measured from the arithmetic mean. This is due to the theorem that the sum of the squares of the deviations about M is less than if taken at any other point. We shall soon prove this theorem.

The quantity

$$\sigma^2 = \frac{\Sigma x^2 f(x)}{N}$$

is usually spoken of as the second moment — since each deviation is squared — of the distribution about the mean expressed in (original units)^2, and is designated by ν_2 (read: nu two). Hence:

$$\sigma^2 = \nu_2$$

The quantity, σ^2, is also known as the *variance* of the distribution.

The computation of σ from the definition, or formula (5), is a decidedly simple though sometimes tedious matter. Let us consider the familiar distribution of college algebra marks as previously considered in Tables 8, 15, and 17. The arithmetic mean has been found to be 74.48. The following table shows the steps involved.

TABLE 23. COMPUTING σ FOR THE DISTRIBUTION OF GRADES IN COLLEGE ALGEBRA BY THE DEFINITION $M = 74.48$

X	$f(x)$	$x = X - M$	x^2	$x^2 f(x)$
95	4	20.52	421.0704	1,684.2816
90	6	15.52	240.8704	1,445.2224
85	12	10.52	110.6704	1,328.0448
80	19	5.52	30.4704	578.9376
75	37	0.52	0.2704	10.0048
70	24	− 4.48	20.0704	481.6896
65	11	− 9.48	89.8704	988.5744
60	6	− 14.48	209.6704	1,258.0224
55	4	− 19.48	379.4704	1,517.8816
50	2	− 24.48	599.2704	1,198.5408
Total	125			10,491.2000

$$\sigma^2 = \nu_2 = \frac{10491.2}{125} = 83.9296 \text{ (c.u.)}^2$$

$$\sigma = \sqrt{\nu_2} = 9.16 = 9.2 \text{ c.u. (approximately)}$$

It may frequently happen that the measures are not sufficiently numerous to warrant their arrangement in a frequency distribution. Thus, consider the 10 scores in centigrade units that were made on a certain test by 10 students of algebra. The scores are given in column one of the table. To apply formula (5) to these values we proceed, as the table shows, to find M and then x corresponding to each score. We have

X	$x = X - 81$	x^2
86	5	25
93	12	144
73	− 8	64
66	− 15	225
88	7	49
96	15	225
80	− 1	1
70	− 11	121
95	14	196
63	− 18	324
$\Sigma X = 810$	00	1374
$M = 81$		

$$N = 10 \quad \Sigma X = 810 \quad M = 81 \text{ c.u.}$$

$$\Sigma x^2 = 1374$$

$$\sigma = \sqrt{\frac{1374}{10}} = 11.7 \text{ c.u.}$$

EXERCISES

1. Statistical data of the United States Department of Agriculture show the following average yields in bushels per acre for the three specified crops. Compute σ for each grain.

Year	*Wheat*	*Rye*	*Oats*	*Year*	*Wheat*	*Rye*	*Oats*
1923	13.3	11.3	30.5	1928	15.4	11.7	32.9
1924	16.0	15.0	34.0	1929	13.0	11.4	29.3
1925	12.8	11.3	31.9	1930	14.2	12.8	32.2
1926	14.7	10.3	26.6	1931	16.3	10.4	28.1
1927	14.7	15.1	27.1	1932	13.0	12.2	30.1

How many of the given 10 values are included in the interval $M \pm \sigma$? Test for each grain.

2. a. Prove: $M_{X+A} = M_X + A$
State this theorem in words.
b. Prove: $M_{X-A} = M_X - A$
State this theorem in words.
c. Prove: $\sigma_{X+A} = \sigma_X$
d. Prove: $\sigma_{X-A} = \sigma_X$
e. Prove: $\Sigma[X - M]^2 = \Sigma X^2 - NM^2 = \Sigma X^2 - \frac{1}{N}[\Sigma X]^2$

3. Compute σ for the given distribution.

Class	X	$f(x)$
32.5–37.5	35	2
27.5–32.5	30	8
22.5–27.5	25	12
17.5–22.5	20	26
12.5–17.5	15	16
7.5–12.5	10	6
2.5– 7.5	5	2
Total		72

Owing to the fact that the value M usually comes out decimally, computing σ by formula (6) is usually laborious, even tedious, hence we are driven to seek other methods. We shall develop two other important methods for computing σ. The first method will express σ in terms of the original variates, X_i, and the second will express σ in terms of x'_i, deviations in *class units* of X, from the arbitrary origin.

Referring to Figure 1 (p. 73), we note that:

$$x = X - M$$

Hence:
$$\nu_2 = \sigma^2 = \frac{\Sigma x^2 f(x)}{N} = \frac{\Sigma (X - M)^2 f(x)}{N}$$

$$= \frac{\Sigma (X^2 - 2MX + M^2) f(x)}{N}$$

$$= \frac{\Sigma X^2 f(x)}{N} - \frac{2M \Sigma X f(x)}{N} + \frac{M^2 \Sigma f(x)}{N}$$

But:
$$\frac{\Sigma X f(x)}{N} = M \quad \text{and} \quad \Sigma f(x) = N$$

Therefore:

$$\sigma^2 = \nu_2 = \frac{\Sigma X^2 f(x)}{N} - 2M^2 + M^2 = \frac{\Sigma X^2 f(x)}{N} - M^2$$

from which we obtain:

$$\sigma = \sqrt{\frac{\Sigma X^2 f(x)}{N} - M^2} = \sqrt{\frac{\Sigma X^2 f(x)}{N} - \left[\frac{\Sigma X f(x)}{N}\right]^2} \qquad (7)$$

This formula gives a straightforward method for computing σ when the original values of X are not too large or when a table of squares is accessible. We shall illustrate the use of the formula for the distribution of college algebra grades. As a matter of fact this table is a continuation of Table 15 (p. 62).

TABLE 24. COMPUTING σ OF THE GRADES IN COLLEGE ALGEBRA BY (7)

X	$f(x)$	$Xf(x)$	$X^2f(x)$
95	4	380	36,100
90	6	540	48,600
85	12	1,020	86,700
80	19	1,520	121,600
75	37	2,775	208,125
70	24	1,680	117,600
65	11	715	46,475
60	6	360	21,600
55	4	220	12,100
50	2	100	5,000
Total	125	9,310	703,900

$$M = \frac{9310}{125} = 74.48 \qquad M^2 = 5547.2704$$

$$\frac{\Sigma X^2 f(x)}{N} = \frac{703900}{125} = 5631.2$$

$$\sigma = \sqrt{5631.2 - 5547.2704} = \sqrt{83.9296} = 9.16 \text{ c.u.}$$

A third, and still more useful, method for computing σ will now be established. The method is analogous to that used in deriving formula (3) of Section 24 (p. 71). From Figure 1 (p. 73) we have

$$x + wb_x = wx' \quad \text{or} \quad x = w(x' - b_x)$$

where w, x', and b_x are defined as in Section 24:

$$\sigma^2 = \nu_2 = \frac{\Sigma x^2 f(x)}{N} = \frac{\Sigma[w(x' - b_x)]^2 f(x)}{N}$$

$$\sigma^2 = w^2\left[\frac{\Sigma x'^2 f(x)}{N} - \frac{2b_x \Sigma x' f(x)}{N} + \frac{b_x^2 \Sigma f(x)}{N}\right]$$

Recalling that $\frac{\Sigma x' f(x)}{N} = b_x$ and $\Sigma f(x) = N$, we have:

$$\sigma^2 = w^2\left[\frac{\Sigma x'^2 f(x)}{N} - b_x^2\right] \tag{8}$$

$$\sigma = w\sqrt{\frac{\Sigma x'^2 f(x)}{N} - b_x^2} \tag{9}$$

Computing σ by (9), which is based upon the class interval as a unit of measure, we shall call the *short method for computing the standard*

TABLE 25. COMPUTING σ FOR THE GRADES IN COLLEGE ALGEBRA BY (9)

X	$f(x)$	$x' = \frac{X - 75}{5}$	$x'f(x)$	$x'^2f(x)$
95	4	4	16	64
90	6	3	18	54
85	12	2	24	48
80	19	1	19	19
75	37	0	0	0
70	24	− 1	− 24	24
65	11	− 2	− 22	44
60	6	− 3	− 18	54
55	4	− 4	− 16	64
50	2	− 5	− 10	50
Total	125		− 13	421

deviation. We shall illustrate its use by computing σ for the distribution of the grades in college algebra. It will be noted that Table 25 is a mere continuation of Table 17 (p. 73).

$$h = 75, \qquad w = 5, \qquad N = 125$$

$$b_x = \frac{-13}{125} = -0.104 \qquad b_x^2 = 0.010816$$

$$M = 75 + 5(-0.104) = 74.48 \text{ c.u.}$$

$$\frac{\Sigma x'^2 f(x)}{N} = \frac{421}{125} = 3.368$$

$$\sigma = 5\sqrt{3.368 - 0.010816} = 5\sqrt{3.357184} = 5(1.832) = 9.16 \text{ c.u.}$$

The observant student will note that in computing σ we have the quantities needed to compute M.

The quantity $\frac{\Sigma x'^2 f(x)}{N}$ is usually denoted in statistics by ν_2' (read:

nu two prime), and is called the *second moment* about the arbitrary origin expressed in (*class units*)^2. Hence:

$$\sigma^2 = \nu_2 = w^2(\nu_2' - \nu_1'^2)$$

If we write formula (8) in the form

$$\sigma^2 = \frac{\Sigma(wx')^2 f(x)}{N} - (wb_x)^2$$

or

$$N\sigma^2 = \Sigma(wx')^2 f(x) - N(wb_x)^2$$

a careful interpretation leads to an important theorem to which attention has previously been called. For $N\sigma^2 = \Sigma x^2 f(x)$ is the sum of the squares of the deviations of the variates about the mean; $\Sigma(wx')^2 f(x)$ is the sum of the squares of the deviations of the variates about any point, and $N(wb_x)^2$ is a positive quantity. Hence the theorem: *the sum of the squares of the deviations of the variates about the mean is less than the sum of the squares of the deviations about any other point.* [See Exercise 28 at end of chapter.]

If dispersion is to be measured by the root-mean-square deviation about some point, the above theorem recommends our taking M for that point, for it is about M that the root-mean-square deviation has a unique value.

The coefficient of relative dispersion based upon the standard deviation is known as the *coefficient of variation*, and is defined by the formula:

$$V_\sigma = \frac{\sigma}{M} \tag{10}$$

and is usually expressed as a percentage. That is, the variability is expressed as a certain per cent of the mean.

A word of comment at this point with regard to formula (10) in particular and to relative variation in general may be desirable. The arbitrary ratio of the standard deviation to the arithmetic mean as a measure of relative variation as well as the other ratios that we have used, *e.g.* formula (4), seems to be based more on psychological than on logical grounds.

Despite individual variation that we have noted among statistical phenomena, we have learned *from experience* to formulate judgments of the individual of normal size. That is, the establishment of norms

seems to be a natural process. We hear the expressions: "What a large apple!" "My, isn't she tiny?" "How emaciated he is!" "What a tremendous ear of corn!" "Wasn't that a hard rain?" "What a hot day in May!". All these expressions imply the notion of a *norm* as well as *variation from a norm.*

We have also formed judgments, which at this time may be crude and inadequate, of *relative variation* with respect to the norm. Any student, without using his statistical analysis, knows that a nose one inch longer than the average length of noses is more monstrous than a height that is one inch longer than the average of the heights. In other words, *a variation is large or small depending upon the norm with which it is associated.*

Doubtless such considerations as the above led Professor Karl Pearson to define the coefficient of variation as the ratio of the standard deviation to the arithmetic mean. The arithmetic mean is taken to be the norm, and the standard deviation measures the variation from the norm.

We should develop a *statistical alertness* to relative variation in characters that are less familiar. Thus, we have found for a distribution of weights of college men that $M = 138.9$ lbs. We shall find that $\sigma = 17.2$ lbs., and hence $V_\sigma = 17.2/138.9 = 0.124 = 12.4\%$. That is, for a group of weights of young men, the standard deviation is about 12.5 per cent, or one-eighth, of the mean. The heights of these same men will give $M = 67.9$ in. and $\sigma = 2.4$ in., and hence $V_\sigma = 2.4/67.9 = 0.035 = 3.5\%$. That is, for a group of heights of young men the standard deviation is about 3.5 per cent of the mean. A distribution of weights, then, shows much more variation than a distribution of heights.

The general literature of biometry records coefficients of variation for many characters. We present herewith a few of them.

Character	V_σ	*Character*	V_σ
Visual acuity	39.12	Pulse rate per min.	14.89
Wt. of heart (unhealthy)	32.39	Chest circumference	8.45
		Length of forearm	5.24
Grip, right hand	25.93	Length of foot	4.59
Wt. of heart (healthy)	17.71	Stature (English)	3.99

Economic data generally show a much larger variation than do biometric data. (Of course the coefficients of variation for much economic data will not remain constant but will vary from time to time.) The weekly earnings of 72,000 Illinois coal miners were analyzed. The analysis gave M = \$8.37, σ = \$2.49, and V_σ = 29.7. An analysis of the price of potatoes gave M = 54.4 cents, σ = 11.11 cents, and V_σ = 19. The variation in economic phenomena will be especially considered in Chapter 6 on Index Numbers.

EXERCISES

1. What norm was used in the development of the formula for V_q?

2. Compute σ for the earnings of each group of Exercise 12, p. 74. The earnings of which group show the greater dispersion?

3. Compute σ for the distribution of the salaries of federal employees that is given in Exercise 36, p. 109. Is it possible to apply formula (9) to this distribution?

4. Compute σ for the distribution of the annual wages of chief wage earners that is given in Exercise 31, p. 108. What is the coefficient of variation for this distribution?

5. Table A gives the I.Q.'s of 905 school children. Table B gives the weights of 1,000 school children. For each distribution find: M, M_d, M_o, σ. Does the interval $M \pm 3\sigma$ include all the variates of each distribution?

TABLE A

X	$f(x)$
60.5	3
70.5	21
80.5	78
90.5	182
100.5	305
110.5	209
120.5	81
130.5	21
140.5	5
Total	905

TABLE B

X	$f(x)$
29.5	1
33.5	14
37.5	56
41.5	172
45.5	245
49.5	263
53.5	156
57.5	67
61.5	23
65.5	3
Total	1000

6. Derive formula (9) from formula (7).

35. THE NORMAL CURVE[1]

In Section 19 (p. 51) reference was made to the normal distribution and to the general form of the equation that represents it. This curve is so important in statistical work, both theoretical and applied, that, although we discuss it rather fully in Chapter 12, we desire at this point to call attention to some of its properties. The general form of the curve is shown by curve (b) of Chart 8 and by the curves (a), (b), and (c) of Section 30 (p. 111). The normal curve is characterized by the symmetrical arrangement of all the variates with respect to a line through the central value, most of the observations lying close to the mean and very few differing from it considerably.

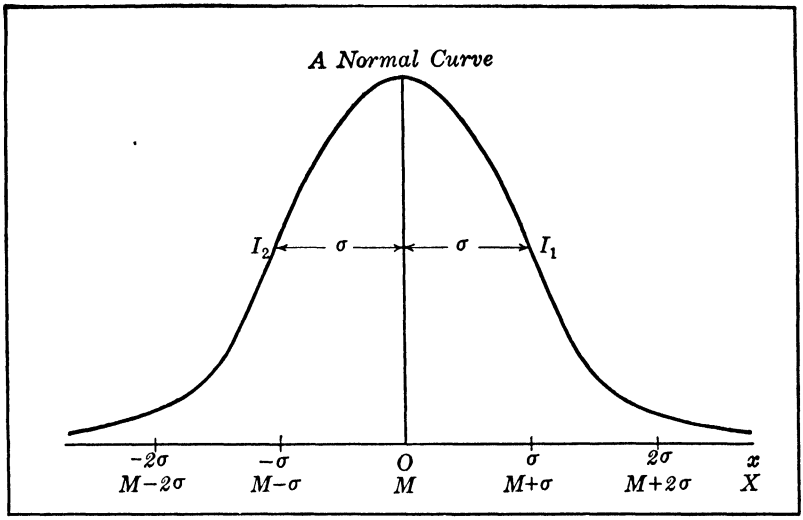
The normal curve is of importance to us just now in that its properties will assist us in making certain generalizations about distributions that do not differ too markedly from normality. And such distributions are not at all rare. Measurements of natural objects — such as the lengths of the leaves on a tree, the heights of men, the lengths of bean pods, the breadths of the heads of men, the lengths and breadths of nuts — distribute themselves with a surprising closeness to normality if large samples are taken. In an approximately normal distribution of a thousand observations we can estimate with surprising accuracy the number that differ from the mean by definite amounts, say σ, 2σ, 3σ, etc. In fact these relations are so regular with the measurements of natural objects that those which are so distributed are said to be *normal.* As has been previously noted, many data collected from the fields of psychology and education are also of this type.

Chart 9 shows a normal curve. The mean, median, and mode coincide at O. It has a maximum at the center and is symmetrical with respect to the vertical line through O. The curve crosses its tangent, that is, the curve changes from concave to convex, at points I_1 and I_2 which are at a distance σ from the vertical through O. The curve approaches the X-axis as x gets large, though we seldom extend it beyond 3σ in either direction from O because the number of such deviations outside $M \pm 3\sigma$ is relatively insignificant.

We have laid off certain multiples of σ on either side of the mean. It will be proved in Chapter 12 that:

[1] If the reader desires to know more about the normal curve, its history and importance, he should read Section 101.

CHART 9



The interval from $M - \sigma$ to $M + \sigma$
includes approximately $\frac{2}{3}N$.

The interval from $M - 2\sigma$ to $M + 2\sigma$
includes approximately 95 per cent of N.

The interval from $M - 3\sigma$ to $M + 3\sigma$
includes approximately 99 per cent of N.

Further, it will be shown that:

The range equals 6σ approximately.
Q equals $\frac{2}{3}\sigma$ approximately.
$M.D.$ from M equals $\frac{4}{5}\sigma$ approximately.

Of course as an observed distribution departs from normality, the approximations are less close.

The number of units of σ that must be laid off on either side of M of a normal distribution to include the total frequency, N, varies with N. If N is very large, more than $\pm 3\sigma$ is necessary whereas if N is small less than $\pm 3\sigma$ is needed. The following table gives the interval that includes N for a normal distribution.

N	*Interval*	N	*Interval*
10	$M \pm 1.65\sigma$	200	$M \pm 2.81\sigma$
20	$M \pm 1.96\sigma$	500	$M \pm 3.0\sigma$
30	$M \pm 2.13\sigma$	1,000	$M \pm 3.3\sigma$
50	$M \pm 2.33\sigma$	10,000	$M \pm 3.9\sigma$
100	$M \pm 2.58\sigma$	100,000	$M \pm 4.4\sigma$

For the distribution of college algebra grades we have found:

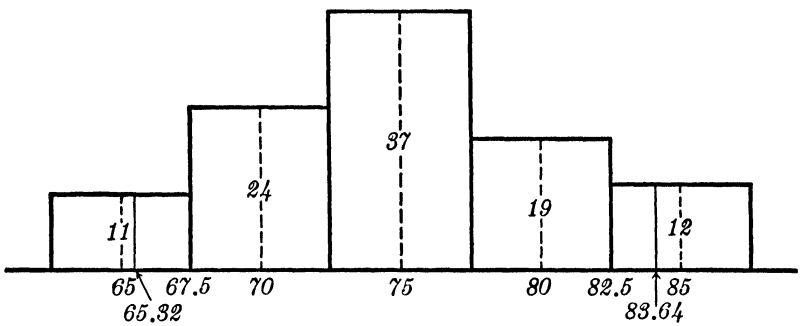
$M = 74.48$ c.u. $\qquad M - \sigma = 65.32$ c.u.

$\sigma = 9.16$ c.u. $\qquad M + \sigma = 83.64$ c.u.

How many of the 125 grades of the sample lie in this interval from 65.32 to 83.64?

To assist us in answering the question let us construct the histogram for the central portion of Table 8 (p. 26).

FIGURE 13



The interval evidently includes the total frequencies of the three central groups (24 + 37 + 19 = 80), and an undetermined part of the classes designated by the class marks 65 and 85. From 65.32 to 67.5 is 2.18, and since the variates are uniformly distributed over the interval we must include (2.18/5)11 = 4.79. Similarly, from 82.5 to 83.64 is 1.14, and hence we must include (1.14/5)12 = 2.73. Hence the interval from 65.32 to 83.64 includes 80 + 4.79 + 2.73 = 87.52, or about 70 per cent of the 125 variates. The result here is more than $\frac{2}{3}N$ for the reason that our distribution is loaded at 75.

When dealing with a distribution of discrete variates, interpolation is usually not necessary. For example consider the distribution given in Exercise 3 (p. 42). We have previously computed for this distribution:

$$\begin{aligned} M &= 53.67 \text{ c.u.} \\ \sigma &= 2.16 \text{ “} \\ M - \sigma &= 51.51 \text{ “} \\ M + \sigma &= 55.83 \text{ “} \end{aligned}$$

The interval from 51.51 to 55.83 includes the frequencies with class marks at 52, 53, 54, and 55, that is, a total of 468 (= 96 + 134 + 127 + 111) or 66.57 per cent of N.

36. THE PROBABLE ERROR

A measure of dispersion that particularly relates to the normal curve is the *probable error*,[1] E_X. It is a distance which, when laid off on either side of the mean of a normal curve, defines an interval that includes one half the total area under the curve. Stated somewhat differently, the probable error of a distribution of variates normally distributed is that deviation on either side of the mean within which half the variates lie. Then, since half the total frequency lies within the interval $M_X - E_X$ to $M_X + E_X$, it is an even chance that a variate selected at random falls within this interval.

The following figure may assist in clarifying the probable error concept. This figure shows the per cent of the total frequency that is included by the indicated probable error units.

The probable error is closely related to the standard deviation. The relationship is indicated by the equation

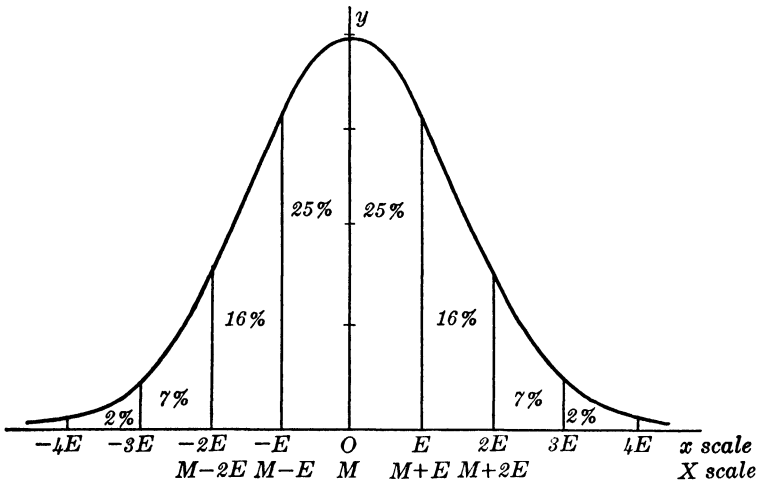
$$E_X = 0.6745\sigma_X \tag{11}$$

Approximately then, E_X is about $\frac{2}{3}\sigma_X$ and σ_X is about $\frac{3}{2}E_X$. *If a distribution is not normal, it is customary to define the probable error by (11).*

Since any multiple of σ can be expressed in terms of E, and vice versa, it is natural to inquire why we have both and what are the

[1] Generally we shall omit the subscript, employing it only for identifications.

FIGURE 14



advantages of E. The standard deviation σ, or a synonym for it, is the older measure. The early nineteenth-century astronomers, particularly Bessel in 1815 and Gauss in 1816 — who were among the first men to work with statistical analysis — desired an interval within and outside of which it is *equally probable* that a random measurement of a normal distribution will occur. This interval on the X-scale is from $M - E$ to $M + E$, or from $M - 0.6745\sigma$ to $M + 0.6745\sigma$. Bessel first used the term probable error, Gauss and the contemporary writers liked the term, and so tradition has kept it in use to this day.

Of course there is a facility of language when using the probable error that may account for its popularity. For example, it is an "even chance," a "fifty-fifty chance," or a "one-to-one shot" that a measure selected at random from a normal distribution falls within or without $M \pm E$. In other words it is as "likely as not" that a measure selected at random from a group of normally distributed variates will fall within the interval $M \pm E$. Equally simple language does not obtain when using the standard deviation. Thus, assume a distribution of the heights of 1,500 men with $M = 67.5$ in., $\sigma = 2.5$ in., and $E = 0.6745\ (2.5) = 1.7$ in. Then if a measure is selected at random from this group *it is as likely as not* to fall within the interval 67.5 ± 1.7 inches.

EXERCISES

1. The data of the following tables are taken from Bulletin No. 620 of the U.S. Department of Labor, "Wages, Hours, and Working Conditions in the Folding-Paper-Box Industry, 1933, 1934, 1935." They present the hourly earnings of employees in the U.S. in the paper-box industry.

Compute M, σ, and E for each table.

X (*cents*)	*May, 1933* $f(x)$	*August, 1934* $f(x)$	*August, 1935* $f(x)$
10 a.u. 15	57	1	0
15 a.u. 20	168	2	8
20 a u. 25	538	9	19
25 a.u. 30	507	20	96
30 a.u. 35	622	231	286
35 a.u. 40	541	1371	1332
40 a.u. 45	485	1834	1670
45 a.u. 50	357	919	969
50 a.u. 55	328	729	806
55 a.u. 60	172	484	563
60 a.u. 70	327	739	758
70 a.u. 80	211	420	457
80 a.u. 100	174	533	555
100 a.u. 120	42	224	264
120 a.u. 150	17	85	82
Total	4546	7601	7865

2. Let the five numbers 3, 4, 5, 6, 7 be a universe. Select different samples of three from these five numbers, 10 samples in all, and compute their means. Thus

$$M_1 = \frac{3+4+5}{3}, \quad M_2 = \frac{3+4+6}{3}, \quad M_3 = \frac{3+4+7}{3}, \ldots, \quad M_{10} = \frac{5+6+7}{3}$$

a. Find the mean of the 10 sample means. How does it compare with the mean of the universe?

b. Find the standard deviation of the 10 sample means. How does it compare with the standard deviation of the universe?

3. Consider the universe of numbers 5, 10, 15, 20, 25. Treat these numbers as you did those of Exercise 2.

4. The problem of sampling has been called by Karl Pearson the fundamental problem in statistics. Often our only statistical knowledge of the parent population is obtained from a study of samples drawn from it.

Stated in rather general terms, the question is: how well does the sample describe the parent population? More precisely, the problem is: how do M and σ computed from the sample compare with M and σ computed from the parent population?

Table 26(a) presents 10 sample distributions of 100 each. The parent population consists of 1,000 individuals. (By assigning the sample distributions to various members of the class the computational labor will be greatly lightened.)

(a) Find M and σ for each sample and for the total.
(b) Find the standard deviation of the 10 means of the samples.
(c) How many of the means are included in the interval (Mean of means) $\pm$ (standard deviation of means)?
(d) Compare the mean of the 10 means with the mean of the total.

TABLE 26(a). DISTRIBUTION OF THE WEIGHTS OF 1,000 MALE STUDENTS
(Measurements to nearest $\frac{1}{10}$ pound)

Class Mark (Pounds)	*Frequencies*										
	1st 100	*2nd 100*	*3rd 100*	*4th 100*	*5th 100*	*6th 100*	*7th 100*	*8th 100*	*9th 100*	*10th 100*	*Total*
95.25			1				1				2
105.25	2	5	2	4		1	3	3		1	21
115.25	11	11	13	9	10	9	9	14	9	9	104
125.25	14	24	15	21	19	25	20	14	21	23	196
135.25	26	16	31	19	31	21	30	22	27	25	248
145.25	23	17	16	26	21	19	17	21	18	19	197
155.25	11	16	11	11	13	16	12	17	11	15	133
165.25	2	5	3	5	4	5	5	5	9	4	47
175.25	5	1	5	0	1	1	2	4	2	4	25
185.25	3	3	1	4	0	1	1		1		14
195.25	1	1	1	1	0	2			1		7
205.25	0	1	1		1				1		4
215.25	0										0
225.25	0										0
235.25	1										1
245.25	1										1
Total	100	100	100	100	100	100	100	100	100	100	1000
M											
σ											

5. Treat the data in Table 26(b) as you did those of Table 26(a).

TABLE 26(b). DISTRIBUTION OF THE HEIGHTS OF 1000 MALE STUDENTS
(Measurements to nearest $\frac{1}{10}$ inch)

Class Mark (Inches)	*Frequencies*										
	1st 100	*2nd 100*	*3rd 100*	*4th 100*	*5th 100*	*6th 100*	*7th 100*	*8th 100*	*9th 100*	*10th 100*	*Total*
59.45		1									1
60.45	1	0									1
61.45	1	0		2			2	2			7
62.45	3	3	1	2	1	5	2	1			18
63.45	6	6	4	5	3	3	0	2	3	1	33
64.45	6	4	9	3	6	11	3	6	7	8	63
65.45	11	7	13	14	10	9	10	6	9	8	97
66.45	12	12	11	13	17	11	19	12	18	12	137
67.45	13	23	17	22	13	10	14	14	13	16	155
68.45	12	15	20	15	20	16	22	26	17	17	180
69.45	14	8	4	9	11	13	12	15	14	22	122
70.45	9	8	5	4	11	5	4	6	10	6	68
71.45	7	8	8	6	4	7	9	7	3	6	65
72.45	2	2	3	4	2	5	1	2	4	3	28
73.45	1	1	2	0	2	3	1	1	2	1	14
74.45	1	1	3	0		1	0				6
75.45	0	0		0		1	1				2
76.45	1	1		1							3
Total	100	100	100	100	100	100	100	100	100	100	1000
M											
σ											

6. Find the standard deviation of the ten standard deviations of Table 26(a), Exercise 4. Which shows the greater dispersion, the sample means or the sample standard deviations?

7. Find the standard deviation of the ten standard deviations of Table 26(b), Exercise 5. Which shows the greater dispersion, the sample means or the sample standard deviations?

37. THE SIGNIFICANCE OF THE MEAN AND THE STANDARD DEVIATION

Thus far our statistical analysis of a given group has enabled us to abstract certain qualities of the group. The most important of these qualities are: (1) the central or typical condition of the group, and

(2) the degree of variability of the members of the group. The central condition can be obtained from the appropriate measure of central tendency, and the degree of variability from the appropriate measure of dispersion, preferably the standard deviation and the coefficient of variation. For example, important facts of the distribution of college algebra marks are:

$$M = 74.48 \text{ c.u.} \qquad \sigma = 9.16 \text{ c.u.}$$

These summarizing constants contain the kernel of the distribution. They give a fairly complete numerical description of the sample.

Now what statistical judgments concerning the parent population — which consists of all the grades in college algebra that are recorded at Bucknell University — can one form from the examination of the sample? [1] We do not desire to answer this question completely at this point, as Chapter 13 is devoted entirely to the problem that we raise here, but we may appropriately state certain facts, which must at this time be accepted without proof.

If the sample that we have been considering was a random one — that is, if any mark in college algebra had the same chance of being selected a member of our sample as any other mark — we may expect that if another sample were selected its mean and its standard deviation would differ but slightly from those we have computed. Furthermore, we may expect the true mean and the true standard deviation of the parent population to differ but little from those of the sample.

It has become customary, therefore, for statisticians who desire to make *statistical estimates of the parent population from an analysis of a sample* to record the results in such a manner that a definite range of variation about the *estimated* measure is determined. The limits of the definite range of variation about an estimated value are established in such a way that we can state the probability that the known value (mean, standard deviation, etc.) found from the sample does not differ more than *a determinate amount* from the unknown and generally unknowable true values of similar constants of the parent population or universe. This *determinate amount* is generally a standard deviation or a probable error. The computed value de-

[1] As a matter of fact 125 measurements constitute far too small a sample for purposes of generalization. We use it here merely for illustrative purposes.

rived from the sample is known, whereas the estimated value belonging to the universe is generally not known. The computed value is the basis of our estimate and our task is to measure the reliability of the estimate. This measure of reliability is expressed in terms of chance or probability. Our method is to find the determinate amount and state the probability that the known value diverges this amount from the true value.

As an example, what can we say about the mean of the universe, M_u, of college algebra marks from our analysis of the sample? We must first compute the determinate amount, the probable error of the mean, E_M, then interpret the result, where

$$E_M = \frac{.6745 \text{ (Standard deviation of sample)}}{\sqrt{\text{Number in the sample}}}$$

or, in brief,

$$E_M = .6745 \frac{\sigma}{\sqrt{N}}$$

For the problem under discussion

$$E_M = .6745 \frac{(9.16)}{\sqrt{125}} = 0.55 \text{ c.u.}$$

and, as is customary, we write

$$M_u = \text{Mean of universe} = 74.48 \pm 0.55 \text{ c.u.}$$

which, translated into English, reads "74.48 with a probable error of 0.55." This means that the chances are even that the sample mean, 74.48 c.u., does not differ more than 0.55 c.u. from the true mean, M_u. Note that we do not say, "The chances are even that the true mean M_u does not differ more than 0.55 from the sample mean 74.48." M_u is a fixed value, it is not a variable as the quotation implies. The sample means, however, are variable. This is an important distinction.

Doubtless, the two preceding paragraphs look formidable. Suppose we now try to make understandable what we have said. Our first sample of 125 scores chosen at random gave a mean 74.48 c.u. and a standard deviation of 9.16 c.u. Another sample of 125 scores chosen in a similar manner would probably yield slightly different results. In other words, *these so-called statistical constants show variation as we move from sample to sample.*

We continue this sampling process until we have a large number of sample means, sample standard deviations, etc. This large number of sample means may be formed into a *distribution of means.* The distribution of means has its mean, M_M; its standard deviation, the standard deviation [1] of the means, σ_M; and its probable error, the probable error of the means, E_M.

This distribution of means has some remarkable properties:

1. It is a normal distribution.
2. Its mean, M_M, is equal to the mean of the universe, i.e. $M_M = M_u$.
3. Its standard deviation is given by $\sigma_M = \frac{\sigma}{\sqrt{N}}$.
4. Its probable error is given by $E_M = .6745\sigma_M = .6745\frac{\sigma}{\sqrt{N}}$.
5. Two thirds of the sample means are included in the interval $M_M \pm \sigma_M$ or in $M_u \pm \sigma_M$.
6. One half of the sample means are included in the interval $M_M \pm E_M$ or in $M_u \pm E_M$. Thus, *the probable error of the mean,* E_M, is a value such that the chances are even that a sample mean lies within the interval, $M_u \pm E_M$, or outside the interval.
7. Practically all the sample means are included in the interval $M_M \pm 3\sigma_M$ or $M_u \pm 3\sigma_M$.

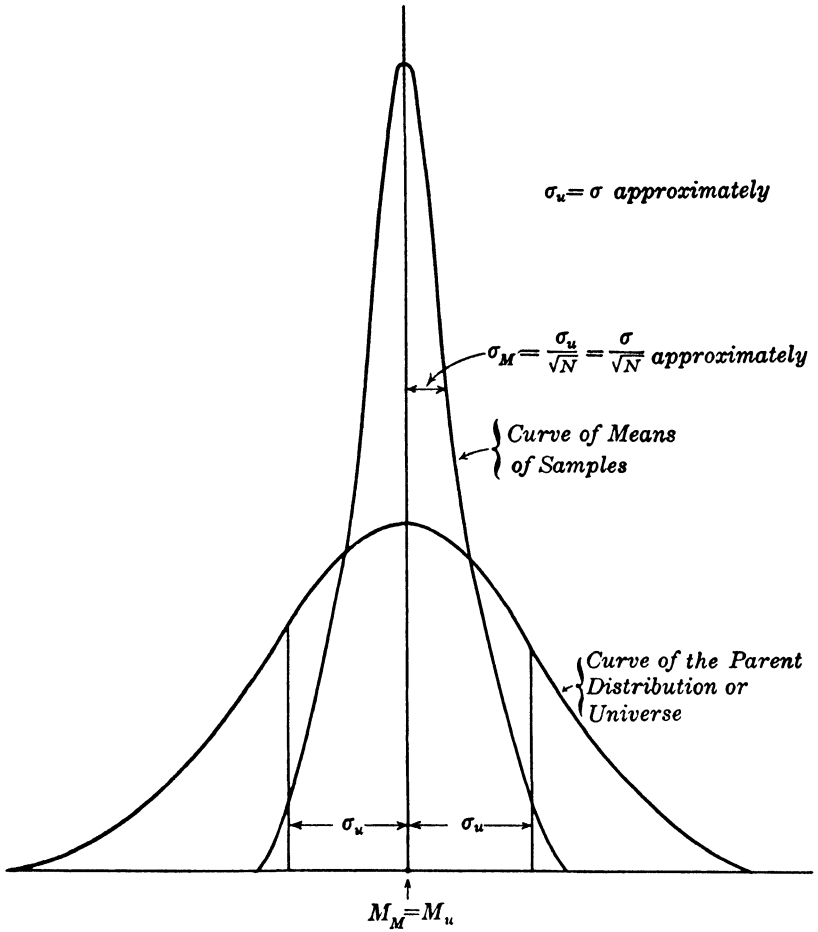
The student will observe from the third property that, even for a reasonably large sample, the distribution of means is rather concentrated. Thus for $N = 100$, $\sigma_M = \sigma/10$, and $\pm 3\sigma_M = \pm 3\sigma/10$, which is a relatively small range. So if N is large, a sample mean M is an excellent estimate of M_u. The little variation in the distribution of means shows that the mean is a stable measure of central tendency. Its stability is illustrated by the rather narrow normal curve of Figure 15.

Similarly, the sample standard deviations may be formed into a distribution of standard deviations. While this distribution is not exactly normal for large values of N, if the samples are taken from a normal universe it does not differ a great deal from normality. It has its mean, M_σ, its standard [2] deviation σ_σ, and its probable error, E_σ.

[1] The standard deviation of the mean is frequently called the *standard error of the mean.*

[2] The standard deviation of σ is frequently called the *standard error* of σ.

FIGURE 15



The sample standard deviations are distributed almost symmetrically about M_σ, which is approximately equal to the standard deviation of the universe σ_u, and with a measurable variation. We can measure the variation of the sample σ's by σ_σ or by E_σ. Formulas for evaluating them are the following:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} = \frac{1}{\sqrt{2}} \cdot \frac{\sigma}{\sqrt{N}} = .707\sigma_M$$

$$E_\sigma = .6745\sigma_\sigma = .6745\frac{\sigma}{\sqrt{2N}} = .707E_M = .4769\sigma_M$$

The *probable error of the standard deviation*, E_σ, is a value such that the chances are even that a sample σ will lie within the interval $\sigma_u \pm E_\sigma$, or outside the interval. As an example illustrating its use, what can we say about the standard deviation of the universe, σ_u, of college algebra marks from our analysis of the sample? We find

$$E_\sigma = .6745\frac{(9.16)}{\sqrt{250}} = .39 \text{ c.u.}$$

or from

$$E_\sigma = .707E_M = .707(.55) = .39 \text{ c.u.}$$

and, as is customary, we write

$$\sigma_u = 9.16 \pm 0.39 \text{ c.u.}$$

and which we read "9.16 with a probable error of 0.39." This means that the chances are even that the sample σ, 9.16 c.u., does not differ more than 0.39 c.u. from the true standard deviation, σ_u.

If the student would prefer to use a "two-to-one-chance" language, he may do so by using as the measures of variation σ_M and σ_σ. This is quite a matter of taste and about tastes we do not wish to argue.

As an illustrative example, again we consider the distribution of college algebra marks. We have

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{9.16}{\sqrt{125}} = 0.82 \text{ c.u.}$$

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} = .707\frac{\sigma}{\sqrt{N}} = (.707)(.82) = 0.57 \text{ c.u.}$$

Thus, the odds are two to one that a sample mean will not differ more than 0.82 c.u. from the mean of the universe, M_u. Or, about two thirds of all sample means are included in the interval $M_u \pm \sigma_M$.

Similarly, the odds are two to one that a sample standard deviation will not differ more than 0.57 c.u. from the standard deviation of the universe, σ_u. That is, about two thirds of the sample standard deviations are included in the interval $\sigma_u \pm \sigma_\sigma$.

EXERCISES

1. Compute σ and $M.D.$ from M for the distribution of Exercise 2, page 54.

2. Compute σ for the distribution of Exercise 1, page 41. How many measures are included by the interval $M \pm \sigma$? Does $M \pm 3\sigma$ include the entire group?

3. Find σ for the theoretical distribution of Exercise 2, page 42. How many measures are included by the interval $M \pm 2\sigma$?

4. Consider the distributions of heights and weights given in Exercise 1, page 54. Which distribution has the greater dispersion?

5. Compute σ and V_σ for the distributions (a) and (b) of scores in English found in Exercise 4, page 102.

6. Find σ and V_σ, for the distributions of the measurements of eggs found in Exercise 15, page 105. Compare these results with those obtained in Exercise 5.

7. a. Show that the standard deviation of the first N integers is given by the equation:

$$\sigma^2 = \frac{1}{12}(N^2 - 1)$$

b. Find σ for the first N odd integers.

8. If N_1, M_1, and σ_1 are the frequency, mean, and standard deviation for one group of measures and N_2, M_2, and σ_2 for a second group, show that the standard deviation of the group formed by combining the two groups is given by:

$$\sigma^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N} + \frac{N_1 N_2}{N^2}(M_1 - M_2)^2$$

where

$N = N_1 + N_2$ = the total frequency of the combined groups, and
σ = the standard deviation of the combined groups

Hint: See Exercise 7 on page 74.

9. Apply the result of the preceding exercise to find the σ for the distribution given in Exercise 8, page 103.

10. At a university 1,000 students were given an objective test. The distribution of marks was closely normal. The analysis gave $M = 72$, $\sigma = 8$. What were the approximate values of Q, Q_1, Q_3, $M.D.$ from M, M_o? Find E_X, E_M, E_σ, and interpret them.

11. Compute E_M and E_σ for the distributions of Exercise 1, page 54. Interpret them.

12. Compute E_M and E_σ for the distribution of Exercise 2, page 54 Interpret these values.

13. In a paper, "Experiment and Statistics in the Selection of Employees," in the *Journal of American Statistical Association*, March 1923, p. 605, Mr. Harry A. Wembridge has presented data that show the points scored on a mental test by 290 prospective employees and the per cent of standard production attained by these same 290 persons after being employed.

The results are:

Scores on Test	*Per cent Production*
$N = 290$	$N = 290$
$M_1 = 42.33$	$M_2 = 92.02$
$\sigma_1 = 9.25$	$\sigma_2 = 24.47$

Compare the variability in mental ability with that of productive ability.

14. Find E_M for the data in Number 13 above and interpret the results.

15. The analysis of an approximately normal distribution of weekly salaries of 300 men gave: $M = \$60.00$ and $\sigma = \$10$.

(1) About how many received salaries between \$50 and \$70?
(2) About how many received salaries between \$40 and \$80?
(3) Approximately, what were the largest and the smallest salaries received?

16. The analysis of two approximately normal distributions of the weekly salaries of 300 men each gave:

1st distribution	*2nd distribution*
$M_1 = \$35.00$	$M_2 = \$60.00$
$M_d = \$34.00$	$M_d = \$58.00$
$\sigma_1 = \$\ 7.00$	$\sigma_2 = \$10.00$

Relatively, which distribution shows the greater dispersion?

17. Distributions of the heights and weights of 1,500 college men were analyzed with the following results:

Heights	*Weights*
$N = 1500$	$N = 1500$
$M_1 = 67.5$ inches	$M_2 = 135.4$ pounds
$\sigma_1 = 2.5$ inches	$\sigma_2 = 15.2$ pounds

Which distribution shows the greater dispersion?

18. From the statistical summaries given in Number 17, assuming the distributions were approximately normal, what are some conclusions that may safely be drawn?

19. Prove: $M_{AX} = AM_X$. Illustrate.

20. Prove: $M_{AX+B} = AM_X + B$. Illustrate.

21. Prove: $\sigma_{AX} = A\sigma_X$. Illustrate.

22. Prove: $\sigma_{AX+B} = A\sigma_X$. Illustrate.

23. Prove: $\Sigma(y - ax) = 0$.

24. Prove: $\Sigma(y - ax)^2 = N(a^2\sigma_X^2 + \sigma_Y^2) - 2a\Sigma xy$.

25. Prove: $\Sigma X \cdot \Sigma Y$ does not equal ΣXY.

26. Prove: $\Sigma(Y - mX - b) = N(M_Y - mM_X - b)$.

27. Supposing that the frequencies of the X-values are the terms of the binomial expansion $(q + p)^n$ as indicated in the table, find σ_X if $(q + p) = 1$. Hint: complete the table as shown and recall that $\Sigma Xf(x) = np$. [See Exercise 42, p. 110.]

X	$f(x)$	$X(X - 1)f(x)$
0	q^n	
1	$nq^{n-1}p$	
2	$\frac{n(n-1)}{2}q^{n-2}p^2$	
⋮	⋮	
n	p^n	

28. On the X-axis are N fixed points $X_1, X_2, \ldots, X_N$ and an unknown point X. Find X so that $\sum_{i=1}^{N}(X_i - X)^2$ is a minimum. Compare with theorem on page 131.

29. In the result of Exercise 27 above, substitute $n = 10$, $p = q = 1/2$, and find σ. Compare your result with that found in Exercise 3 of this set.

Chapter 5

SKEWNESS: EXCESS: MOMENTS

38. INTRODUCTION

The two preceding chapters have been concerned with characterizing masses of numerical data by means of certain summarizing numbers. These summarizing numbers have in general been well-defined statistical constants that were designed to measure central tendency and dispersion. With the computation of these constants the distribution has been partially characterized and described.

When we say of a distribution of heights, for example, that it shows a mean of 67.5 inches and a standard deviation of 2.5 inches, we know that approximately two-thirds of the total frequency is found within the interval 65 to 70; that it is extremely unlikely that any member of the distribution will be found without the limits $67.5 \pm 3(2.5)$; that the total range is about 6(2.5) inches. If other summarizing numbers such as Q_1, Q_3, M_o, etc. are given, then our knowledge of the distribution is considerably enlarged. The main purpose of these summarizing numbers is to assist us in comprehending the important features of a distribution though the distribution may not be present before us.

39. THE MEANING OF SKEWNESS

Our confidence in the conclusions mentioned in the preceding section is especially strengthened by the knowledge that the distributions of heights of men chosen at random are fairly symmetrical. However, we can conceive of a city police force constituted of men at least 65 inches in height, that the symmetry of such a selected group would be greatly disturbed by the selectivity, and that the range of values greater than the mean would be longer than the range of values less than the mean. This characteristic feature of lack of symmetry in distributions is usually called *skewness* or *asymmetry*.

In the preceding chapter emphasis was placed upon the fact that

dispersion is concerned with the *amount of the variation* rather than with its *direction.* We feel the need for a statistical constant which will summarize the direction of the variation or the departure from symmetry. And just as we found it advisable to measure dispersion for purposes of comparison by measures of *relative variability,* so for purposes of comparison we must invent measures of *relative skewness.* Owing to the fact that skewness is dependent upon the amount of dispersion, the coefficients of relative skewness are obtained by dividing the absolute skewness by some measure of absolute dispersion. This method will result in ratios or abstract numbers which are independent of the units in which the original variates are measured.

40. THE MEASUREMENT OF SKEWNESS

It is an obvious fact that in unimodal symmetrical distributions the mean, the median, and the mode coincide. Also in symmetrical distributions the numerical distances from the median to the lower and upper quartiles are equal, and certain pairs of deciles are equidistant from the median. As the distribution departs from symmetry there is a separation of the three measures of central tendency, the difference between the mean and the mode being greatest. Also skewness is indicated when the distances from the median to the quartiles become unequal, and when pairs of deciles are not equidistant from the median. Evidently any of these differences can be made the bases for measurements of skewness.

Since the mean and the median are pulled away from the mode in the direction of the skew, or the tail of the curve representing the extreme measures, an evident measure of absolute skewness could be taken to be $M - M_o$. Professor Karl Pearson has used this as the basis for his formula for relative skewness, namely:

$$Sk = \frac{M - M_o}{\sigma} \tag{1}$$

If the mean is to the right of the mode,[1] that is if $M > M_o$, as in curve A, the skewness is positive, whereas if the mean is to the left of the mode, that is, $M < M_o$, as in curve C, the skewness is negative.

[1] See Figures 16 and 17.

If the mean and the mode coincide, as in curves B and D, the skewness is zero.

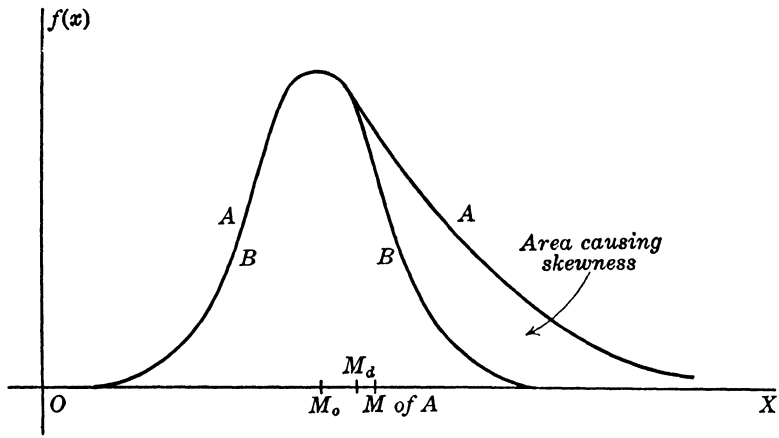
The formula (1), known as Pearson's formula, is open to the objection that in many distributions there is no well-defined mode. Since in many distributions the approximate relation

$$M - M_o = 3(M - M_d)$$

has been found to obtain, this relation suggests the use of the alternative Pearson form:

$$Sk = \frac{3(M - M_d)}{\sigma} \tag{2}$$

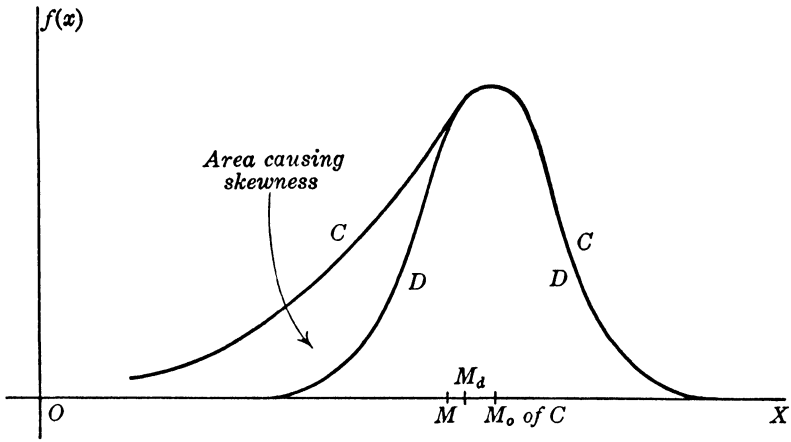
FIGURE 16



Curve A shows positive skewness while curve B is symmetrical.

Since in measuring skewness we are interested in the *degree* of asymmetry a coefficient of skewness is always an index that may be used to compare the unsymmetrical distribution with a symmetrical one that we superimpose. Thus, in Figure 16 we may consider A as the frequency curve for a given distribution and B as the symmetrical curve that is drawn to display the skewness in A. We indicate on the figure that the area bounded by the curves A and B and the X-axis causes the skewness in A. We can make a similar statement about the curves shown in Figure 17.

FIGURE 17



Curve C shows negative skewness while curve D is symmetrical.

Let us be more specific and consider the four distributions of Table 27. The student should verify the statistical constants pertaining to each distribution. We have drawn histograms of these distributions (see page 155) and on them we have located the *points* which mark the position of the arithmetic mean and the median, and the *distance* which indicates the value of the standard deviation. The coefficients of skewness, since they are pure numbers or indexes, cannot of course be shown on the graphs.

The reader should be warned that coefficients of skewness like all relative numbers may not mean much until he has had a considerable experience with many and varied distributions. Only by drawing the histograms, marking on them the points for M and M_d, and the distance for σ, then computing Sk by any of our formulas and comparing the results for several distributions — the more the better — will these values take on a real meaning.

It has been shown [1] that $(M - M_d)/\sigma$ lies between -1 and $+1$, and thus skewness computed by formula (2) is always between -3 and $+3$. This measure of skewness is obviously quite sensitive. While it is dangerous to set limits on such indexes, we may say, as

[1] Harold Hotelling and Leonard M. Solomons, *Annals of Mathematical Statistics*, May 1932, pp. 141–142.

TABLE 27

Class	*X*	*A* $f(x)$	*B* $f(x)$	*C* $f(x)$	*D* $f(x)$
87.5–92.5	90	0	4	0	4
82.5–87.5	85	12	4	4	8
77.5–82.5	80	24	20	40	20
72.5–77.5	75	28	44	24	24
67.5–72.5	70	24	20	20	40
62.5–67.5	65	12	4	8	4
57.5–62.5	60	0	4	4	0
N		100	100	100	100
M		75	75	75	75
M_d		75	75	76.25	73.75
σ		6	6	6	6
Sk by (2)		0	0	− 0.625	+ 0.625
α_3		0	0	− 1.2	+ 1.2

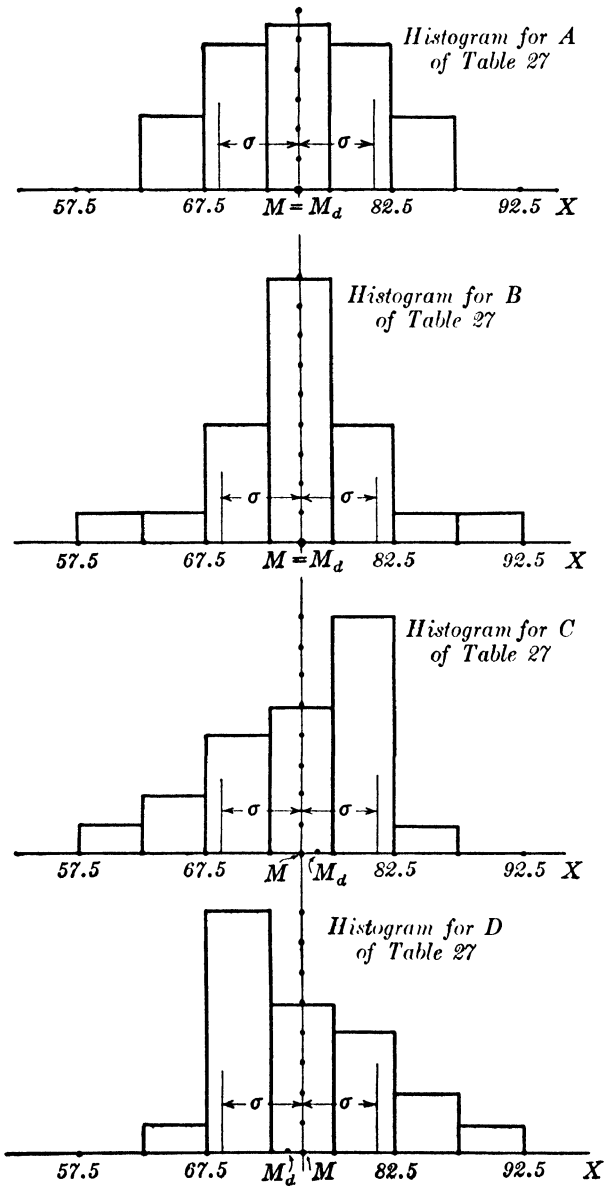
a rough measuring stick, numerical values of skewness computed by (2) less than 0.25 may be considered small, numerical values between 0.25 and 0.5 as moderate, and numerical values greater than 0.5 as large. Numerical values as large as 1 are unusual.

Exercise. a. For each of the following distributions compute M, M_d, σ, and S_k.

b. Draw the histogram and indicate upon it the points M, M_d, and the distance σ.

X	*A* $f(x)$	*B* $f(x)$	*C* $f(x)$	*D* $f(x)$
35	4	16	2	1
30	12	48	4	2
25	20	12	8	3
20	28	10	10	4
15	20	8	12	15
10	12	4	48	25
5	4	2	16	50
Total	100	100	100	100

FIGURE 18


Histogram for A of Table 27
σ
σ
57.5
67.5
M = M_d
82.5
92.5
X
Histogram for B of Table 27
σ
σ
57.5
67.5
M = M_d
82.5
92.5
X
Histogram for C of Table 27
σ
σ
57.5
67.5
M
M_d
82.5
92.5
X
Histogram for D of Table 27
σ
σ
57.5
67.5
M_d
M
82.5
92.5
X

A third measure of skewness that has become well known is that due to Bowley. It is based upon the fact previously mentioned that in an asymmetrical distribution the numerical distances from the median to the lower and the upper quartiles are unequal. If q_1 and q_2 are defined by the equations

$$q_1 = M_d - Q_1 \text{ and } q_2 = Q_3 - M_d$$

then:

$$Sk = \frac{q_2 - q_1}{q_2 + q_1} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \tag{3}$$

In regard to formula (3), Bowley says:

> If the curve is symmetrical, $q_2 = q_1$, and $Sk = 0$; if $q_2 > q_1$, Sk is positive, and if $q_2 < q_1$, Sk is negative. Sk becomes $+1$ if $q_1 = 0$, that is, if the median and lower quartile coincide; and Sk becomes -1 if $q_2 = 0$. Sk is therefore a measurement which never exceeds 1 numerically, and has a definite significance at zero and at its extreme values. . . . The significance of the various values can only be obtained by experience, but it may be suggested that 0.1 is a moderate degree of skewness, and 0.3 a considerable degree.[1]

The quartile measure of skewness is rigidly defined, is simple to compute, and is easily understood. It is a pure number, and the restriction of its value to the small interval from -1 to $+1$ leaves it sufficiently sensitive for many needs. A just criticism is that it fails to take into consideration the size of the extreme variations. Since the main question in skewness is the determination of how much more the items deviate on one side of the mean than on the other, the ideal measure of skewness should give due emphasis to the extreme variations.

Many of the objections to the previously mentioned methods for measuring skewness may be met by returning to a consideration of the deviations of the variates from their mean. Since we are interested in *how* the variates are situated with respect to the mean and since we wish to give emphasis to the extreme measures, we require some function of the form

$$\Sigma x^n f(x)$$

for some value of n. Now if n is even, we obtain the amount and not the direction of the variation. In order to secure the direction of

[1] Bowley, *op. cit.*, p. 116.

the variation, we are compelled to use *odd* numbers for n. If $n = 1$, $\Sigma x^n = 0$. If $n = 3$, we obtain $\Sigma x^3 f(x)$, a basic factor in our next measure for skewness known as α_3 (read: alpha three). α_3 is defined as the third moment of the distribution about the mean divided by the cube of the standard deviation, or by the equation:

$$\alpha_3 = \frac{\frac{\Sigma x^3 f(x)}{N}}{\sigma^3} = \frac{\text{the third moment about } M}{\text{cube of the standard deviation}} \tag{4}$$

that is

$$\alpha_3 = \frac{\nu_3}{\sigma^3} = \frac{\text{nu three}}{\text{sigma cube}}$$

In what follows we shall consider α_3 as the preferable measure of skewness.[1]

As an illustrative problem we shall compute α_3 for the distribution of grades in college algebra. The table will be a continuation of Table 23, page 126.

TABLE 28. COMPUTING α_3 FOR THE DISTRIBUTION OF GRADES IN COLLEGE ALGEBRA BY THE DEFINITION $M = 74.48$

X	$f(x)$	x	$x^2f(x)$	$x^3f(x)$
95	4	20.52	1,684.2816	34,561.458432
90	6	15.52	1,445.2224	22,429.851648
85	12	10.52	1,328.0448	13,971.031296
80	19	5.52	578.9376	3,195.735552
75	37	0.52	10.0048	5.202496
70	24	− 4.48	481.6896	− 2,157.969408
65	11	− 9.48	988.5744	− 9,371.685312
60	6	− 14.48	1,258.0224	− 18,216.164352
55	4	− 19.48	1,517.8816	− 29,568.333568
50	2	− 24.48	1,198.5408	− 29,340.278784
Total	125		10,491.2000	− 14,491.152000

$$\sigma^2 = \nu_2 = \frac{10491.2}{125} = 83.9296 \text{ (c.u.)}^2$$

$$\sigma = 9.16 \text{ c.u.}$$

$$\sigma^3 = 768.795136 \text{ (c.u.)}^3$$

[1] α_3 is zero for the normal curve. See page 405.

$$\nu_3 = \frac{\Sigma x^3 f(x)}{N} = \frac{-\ 14491.152}{125} = -\ 115.929216 \quad \text{(c.u.)}^3$$

$$\alpha_3 = \frac{\nu_3}{\sigma^3} = \frac{-\ 115.929216}{768.795136} = -\ 0.1507$$

α_3 is a very refined measure of skewness. The process of cubing maintains the proper signs for the deviations and also gives emphasis to the extreme variates. Further, the division by σ^3 reduces the measure to an abstract number. Hence it is a coefficient of relative skewness and is independent of the unit of measure. Since it is not restricted in its range, it is a very sensitive measure, the sensitiveness being emphasized by the cubing of the deviations. Its chief disadvantage is the apparent labor of computing it. We shall greatly overcome this apparent trouble in Section 44 (p. 164) by developing a "short method."

41. EXCESS OR KURTOSIS

In elementary statistics a distribution is usually satisfactorily characterized by the measures of central tendency, the measures of dispersion, and the measures of skewness, or more briefly, by M, σ, and α_3. We may add one other important constant to the summarized description by considering the relative number of the variates in the immediate neighborhood of the mean or mode. This measure of relative flatness (or peakedness) of a curve fitted to the distribution as compared with that of the normal curve fitted to the same distribution is called a measure of *excess* or *kurtosis*.

The excess or kurtosis is measured by:

$$K = \alpha_4 - 3 = \frac{\dfrac{\Sigma x^4 f(x)}{N}}{\sigma^4} - 3 = \frac{\nu_4}{\sigma^4} - 3 \qquad (5)$$

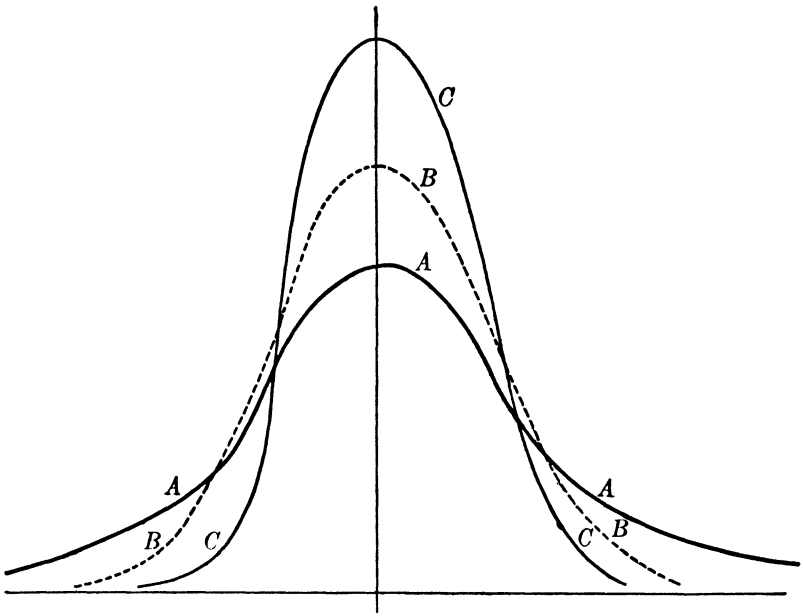
Again the normal curve is our standard for comparison. Since $\alpha_4 = 3$ for the normal curve (see page 405) the excess for any other curve is merely a comparison of its α_4 with that of the normal curve which has the same standard deviation.

If the excess is positive (leptokurtic), the number of variates near the mean is greater than in a normal distribution. If the excess is negative (platykurtic), the curve is more flat-topped than the

corresponding normal frequency curve. The normal curve, in which $\alpha_4 = 3$, is said to be mesokurtic.

Figure 19, exhibiting three curves with the same mean and the same standard deviation, illustrates graphically the meaning of excess.

FIGURE 19



Curve A is platykurtic and $\alpha_4 - 3 < 0$; curve B is mesokurtic (normal), and $\alpha_4 - 3 = 0$; and curve C is leptokurtic and $\alpha_4 - 3 > 0$.

42. THE UNADJUSTED MOMENTS OF A DISTRIBUTION

In the preceding chapters we have several times mentioned the term *moment*. It is a concept so important in statistical analysis that we cannot longer defer its more complete consideration. We shall soon learn that a statistical distribution — which we have characterized by its mean, its standard deviation, its skewness, its excess — is, in brief, characterized by its moments.

Further, the notion of moments serves as a guide in curve-fitting. It was remarked in Section 16 (p. 38) that the total area under a

frequency curve should equal the area of the histogram, which is another way of saying that the total frequency should be unchanged. As the total frequency is the zeroth moment, this is equivalent to requiring that the zeroth moment of the frequency curve equal the zeroth moment of the given distribution. In like manner, we may require that a sufficient number of successive moments of higher orders of the frequency curve be equal to the corresponding higher moments of the given distribution. This is the so-called *principle of moments*[1] for the determination of the parameters in a curve which is to be selected to represent a given distribution. We shall have an opportunity to observe an application of the principle of moments in Chapters 7, 12, and 13.

The moments of a distribution can be computed about any point at pleasure. They can be expressed in various units. The most significant moments are referred to the mean, M, and are usually expressed either in the given or the class unit. They may then be reduced to abstract numbers by dividing by the appropriate powers of σ as was done, for example, in defining α_3 and α_4. As illustrations, we have learned that if x equals the deviation of any frequency from M expressed in the given unit, then:

$$\left.\begin{aligned}
\nu_1 &= \frac{\Sigma x f(x)}{N} = \text{the 1st moment of the distribution about } M \text{ expressed in the given unit} = 0 \\
\nu_2 &= \frac{\Sigma x^2 f(x)}{N} = \text{the 2nd moment of the distribution about } M \text{ expressed in the (given unit)}^2 = \sigma^2 \\
\nu_3 &= \frac{\Sigma x^3 f(x)}{N} = \text{the 3rd moment of the distribution about } M \text{ expressed in the (given unit)}^3 \\
\nu_4 &= \frac{\Sigma x^4 f(x)}{N} = \text{the 4th moment of the distribution about } M \text{ expressed in the (given unit)}^4
\end{aligned}\right\} \quad (6)$$

etc.

Hence in general we define:

$$\nu_n = \frac{\Sigma x^n f(x)}{N} = \text{the } n\text{th moment of the distribution about } M \text{ expressed in the (given unit)}^n$$

If $n = 0$, we have:

$$\nu_0 = \frac{\Sigma x^0 f(x)}{N} = 1$$

[1] Rietz and others, *op. cit.*, p. 68.

In our previous discussion, not only have we encountered moments about M but about other points as well. Formula (7), page 9, involves, for example:

$$\frac{\Sigma Xf(x)}{N} = \text{the 1st moment about zero} = M$$

and

$$\frac{\Sigma X^2 f(x)}{N} = \text{the 2nd moment about zero.}$$

The higher moments about zero are similarly defined.

In computing the standard deviation we noted that the arithmetical computations were simplified by referring the variates to some point near the mean and expressing them in class units (see Table 25, p. 130). We shall soon discover that this transformation to class units is especially useful when computing the higher moments. In computing α_3 (Table 28, p. 157) we felt a need for some short method.

On pages 72 and 130 we have noted that if x' equals the deviation of any frequency from the assumed origin $O'(h, 0)$ expressed in the *class unit*, then:

$$\left.\begin{aligned} \nu_1' &= \frac{\Sigma x' f(x)}{N} = \text{the 1st moment of the distribution about } O' \text{ expressed in the class unit} = b_x \\ \nu_2' &= \frac{\Sigma x'^2 f(x)}{N} = \text{the 2nd moment of the distribution about } O' \text{ expressed in the (class unit)}^2 \end{aligned}\right\} \quad (7)$$

Hence in general we may define:

$$\left.\nu_n' = \frac{\Sigma x'^n f(x)}{N} = \text{the } n\text{th moment of the distribution about } O' \text{ expressed in the (class unit)}^n\right.$$

If w = the class width, x' class units = wx' given units, and hence:

$$\nu_n' = \frac{\Sigma (wx')^n f(x)}{N} = \frac{w^n \Sigma x'^n f(x)}{N} = \text{the } n\text{th moment about } O' \text{ expressed in the (given unit)}^n$$

Therefore we have the theorem: *The nth moment of a given distribution about any point O′ in the nth power of the given unit equals w^n times its nth moment about O′ in the nth power of the class unit.* Or in short:

$$\nu_n' \text{ in the (given unit)}^n = w^n \nu_n' \text{ in the (class unit)}^n \quad (8)$$

While the most significant moments are those computed about the mean, yet their computation directly from the definition, is very tedious, owing to the fact that M usually involves several decimals, and hence

$$x = X - M$$

also involves several decimals. Raising these decimals to the third, fourth, and higher powers is laborious, even with the aid of a calculating machine. However, just as we were able to avoid this tedium with the short method for computing σ (Table 25, p. 130), so we shall avoid it in computing the higher moments.

From Figure 1 (p. 73) we have noted that:

$x = wx' - wb_x = w(x' - b_x)$ expressed in the given unit

Hence:

$$\left.\begin{aligned} \nu_1 &= \frac{\Sigma x f(x)}{N} = \frac{\Sigma w(x' - b_x) f(x)}{N} = \frac{w\Sigma x' f(x)}{N} - \frac{wb_x \Sigma f(x)}{N} = 0 \\ \nu_2 &= \frac{\Sigma x^2 f(x)}{N} = \frac{\Sigma w^2 (x' - b_x)^2 f(x)}{N} = \frac{\Sigma w^2 [x'^2 - 2b_x x' + b_x^2] f(x)}{N} \\ &= w^2 \left[\frac{\Sigma x'^2 f(x)}{N} - b_x^2\right] = w^2(\nu_2' - b_x^2) \end{aligned}\right\} \quad (9)$$

In like manner, it follows that:

$$\left.\begin{aligned} \nu_3 &= w^3(\nu_3' - 3\nu_2' b_x + 2b_x^3) \\ \nu_4 &= w^4(\nu_4' - 4\nu_3' b_x + 6\nu_2' b_x^2 - 3b_x^4) \end{aligned}\right\}$$

etc.

These moments described in (9) of course express the ν's in *given units*. If the class interval is taken as the unit, which is usually the case, $w = 1$, and then the moments are expressed in class units. If $h = 0$, ν_n' becomes the nth moment about zero as origin.

We have found it desirable to express M and σ in terms of the given unit. However the third, fourth, and higher moments are usually expressed as ratios in such a manner that they are independent of the unit of measure. This was accomplished in defining α_3 and α_4 by dividing ν_3 and ν_4 by σ^3 and σ^4 respectively.[1] Thus:

[1] $\alpha_n = \frac{\nu_n}{\sigma^n}$ is the nth moment about M expressed in (*standard* units)^n. A variate is expressed in *standard* units by dividing its deviation from M by σ. It is usually indicated by t. Thus, $t_i = \frac{X_i - M}{\sigma} = \frac{x_i}{\sigma}$.

$$\alpha_3 = \frac{\nu_3}{\sigma^3} = \frac{\Sigma t^3 f(x)}{N}, \quad \alpha_4 = \frac{\nu_4}{\sigma^4} = \frac{\Sigma t^4 f(x)}{N}$$

and in general

$$\alpha_n = \frac{\nu_n}{\sigma^n} = \frac{\Sigma t^n f(x)}{N}$$

In particular we note that:

$$\alpha_1 = 0 \quad \text{and} \quad \alpha_2 = 1$$

The moments

$$\nu_n = \frac{\Sigma x^n f(x)}{N} \quad \text{and} \quad \nu'_n = \frac{\Sigma x'^n f(x)}{N}$$

are frequently called the *crude* or *unadjusted moments* about the mean and assumed point respectively. The standard deviation, the skewness, and the excess, if based upon them, are called the *unadjusted standard deviation*, the *unadjusted skewness*, etc.

43. THE ADJUSTED MOMENTS: SHEPPARD'S CORRECTIONS

In arranging our data in a frequency distribution we have assumed that the items in a given class were concentrated at its mid-point. This procedure introduces a slight error, which we call a *grouping error*. By a process too abstruse for consideration here, certain corrections — known as Sheppard's Corrections — have been devised to assist in correcting the errors in the moments due to grouping. When applied to the crude moments they give the *adjusted moments* of the distribution. It is quite customary to denote the adjusted moments by μ_n (read: mu enn), $n = 1, 2, 3$, etc. They find their widest application in fitting a frequency function to a distribution of observed measurements by what is known as the *method of moments*.

The adjusted moments are not generally recommended for use in unrefined statistical analysis. Especially is this true if the original data were not taken with sufficient accuracy to warrant our using the niceties of analysis that are implied in the corrections, for certainly we should not adopt methods in computation that are inconsistent with the data at hand. A more potent reason for our failure to recommend their employment generally is due to the fact that an

intelligent use of them requires a knowledge of their development.[1] Finished statisticians use them with care and discrimination. We do not wish to discourage their use when the data warrant it and when they can be employed with safety and confidence, but we do insist that they should be used with understanding. We mention them here to add completeness to our text, to illustrate the method of computing them to the student who may continue the study of statistical analysis beyond this introductory text, and to caution the reader against their indiscriminate use.

The adjusted moments involving Sheppard's Corrections are given by the following equations:

$$\left.\begin{aligned} \mu_2 &= \nu_2 - \frac{w^2}{12} \\ \mu_3 &= \nu_3 \\ \mu_4 &= \nu_4 - \frac{\nu_2 w^2}{2} + \frac{7w^4}{240} \end{aligned}\right\} \quad \text{Sheppard's Corrections if moments are expressed in the given unit} \qquad (10)$$

where w is the class interval.

If the moments are expressed in the class unit, $w = 1$, and the simplifications are evident.

The refined or adjusted formulas for the standard deviation, the skewness, and the excess are given by:

$$\sigma = \sqrt{\mu_2}, \qquad \alpha_3 = \frac{\mu_3}{\sigma^3}, \qquad \alpha_4 - 3 = \frac{\mu_4}{\sigma^4} - 3$$

No corrections are applied to the moments of theoretical distributions and curves. In such cases we indicate the nth moment about M by μ_n and about any other point by μ'_n.

44. COMPUTATION OF THE MOMENTS

The order of procedure when computing the moments should be:

1. Choose a convenient arbitrary origin, and compute ν'_1, ν'_2, ν'_3, ν'_4.
2. Transfer the moments to the mean by means of equations (9), and thus compute ν_1, ν_2, ν_3, ν_4. See that the proper units are included.
3. If Sheppard's Corrections are to be applied, use equations (10) and compute μ_1, μ_2, μ_3, μ_4.

We shall illustrate this procedure by computing the moments of the following distribution:

[1] Rietz and others, *op. cit.*, pp. 92 *et seq.*

TABLE 29. FREQUENCY DISTRIBUTION OF PULSE BEATS PER MINUTE IN ENGLISH CONVICTS [1]

X	$f(x)$	x'	$x'f(x)$	$x'^2f(x)$	$x'^3f(x)$	$x'^4f(x)$
46.5	2	− 8	− 16	128	− 1,024	8,192
50.5	5	− 7	− 35	245	− 1,715	12,005
54.5	17	− 6	− 102	612	− 3,672	22,032
58.5	57	− 5	− 285	1,425	− 7,125	35,625
62.5	90	− 4	− 360	1,440	− 5,760	23,040
66.5	150	− 3	− 450	1,350	− 4,050	12,150
70.5	120	− 2	− 240	480	− 960	1,920
74.5	131	− 1	− 131	131	− 131	131
78.5	109	0	000	000	000	000
82.5	86	1	86	86	86	86
86.5	62	2	124	248	496	992
90.5	42	3	126	378	1,134	3,402
94.5	15	4	60	240	960	3,840
98.5	18	5	90	450	2,250	11,250
102.5	9	6	54	324	1,944	11,664
106.5	5	7	35	245	1,715	12,005
110.5	3	8	24	192	1,536	12,288
114.5	3	9	27	243	2,187	19,683
Total	924		− 993	8,217	− 12,129	190,305

Choosing $h = 78.5$, we have for the ν''s:

$$\nu'_1 = \frac{\Sigma x'f(x)}{N} = b_x = \frac{-993}{924} = -1.0746753$$

$$\nu'_2 = \frac{\Sigma x'^2f(x)}{N} = \frac{8217}{924} = 8.8928571$$

$$\nu'_3 = \frac{\Sigma x'^3f(x)}{N} = \frac{-12129}{924} = -13.12662338$$

$$\nu'_4 = \frac{\Sigma x'^4f(x)}{N} = \frac{190305}{924} = 205.9577923$$

Using equations (9) we shall now express the ν's in the given unit. We have, noting that $w = 4$:

$$\nu_1 = 0$$

$$\nu_2 = 16[8.8928571 - (-1.0746753)^2]$$
$$= 16(8.8928571 - 1.1549270) = 16(7.73793014)$$

[1] The data are taken from *Biometrika*, Vol. 11.

$$\nu_3 = 64[-13.12662338 - 3(8.8928571)(-1.0746753) + 2(-1.0746753)^3]$$
$$= 64(-13.12662338 + 28.67080162 - 2.48234304)$$
$$= 64(13.06183520)$$
$$\nu_4 = 256[205.9577923 - 4(-13.12662338)(-1.0746753) + 6(8.8928571)(-1.0746573)^2 - 3(-1.0746753)^4]$$
$$= 256(205.9577923 - 56.42753004 + 61.62360463 - 4.0015692)$$
$$= 256(207.1522977)$$

Assuming that Sheppard's Corrections may be applied, we find the μ's:

$$\mu_1 = \nu_1 = 0$$
$$\mu_2 = 16(7.73793014) - 16(.08333333) = 16(7.65459681)$$
$$\mu_3 = \nu_3 = 64(13.06183519)$$
$$\mu_4 = 256[207.152298 - \frac{256(7.73793014)}{2} + 256(.029167)]$$
$$= 256(203.312500)$$

Hence we have:

the unadjusted constants

$$M = 78.5 + 4(-1.074675) = 74.2 \text{ p.b.}$$
$$\sigma = \sqrt{\nu_2} = 4(2.7817135) = 11.12685 \text{ p.b.}$$
$$\alpha_3 = \frac{\nu_3}{\sigma^3} = \frac{64(13.06183520)}{64(21.5247046)} = 0.6068$$
$$\alpha_4 = \frac{\nu_4}{\sigma^4} = \frac{256(207.1522977)}{256(59.875562)} = 3.4597$$
$$K = \alpha^4 - 3 = 0.4597$$

and the adjusted constants

$$M = 74.2 \text{ p.b.}$$
$$\sigma = \sqrt{\mu_2} = 4(2.7667) = 11.0668 \text{ p.b.}$$
$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{\nu_3}{\sigma^3} = 0.6168$$
$$\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{256(203.312500)}{256(58.592852)}$$
$$= 3.46992$$
$$K = \alpha_4 - 3 = 0.46992$$

The student will note that the application of Sheppard's Corrections here has affected the constants slightly.

Assuming that the parent population is normal, the values of $\alpha_{3,u}$ and $\alpha_{4,u}$ of the universe are usually written:

$$\alpha_{3,u} = (\text{the computed } \alpha_3) \pm 0.6745\sqrt{\frac{6}{N}}$$

$$\alpha_{4,u} = (\text{the computed } \alpha_4) \pm 0.6745\sqrt{\frac{24}{N}}$$

These statements mean that the chances are even, or it is equally likely, that the computed values of α_3 and α_4 do not differ numerically more than the specified amounts from the true values, $\alpha_{3,u}$ and $\alpha_{4,u}$.

For the illustrative problem we are considering we have:

$$\left.\begin{aligned} M &= 74.2 \pm 0.2456 \\ \sigma &= 11.0668 \pm 0.1736 \\ \alpha_3 &= 0.6168 \pm 0.0544 \\ \alpha_4 &= 3.46992 \pm 0.1088 \end{aligned}\right\} \text{(See Section 37, p. 142.)}$$

EXERCISES

1. Using Bowley's coefficient, formula (3), find the skewness for the distribution of grades in college algebra as given in Table 8 (p. 26).

2. Using Pearson's formula (2), find the skewness for the distributions of heights and weights described in Exercise 1, page 54.

3. Using Bowley's coefficient, find the skewness for the distributions of heights and weights described in Exercise 1, page 54.

4. Find σ, α_3, and α_4 for the distributions of Exercise 4, page 102.

5. Continue the analysis of Exercise 6 on page 147 by finding α_3 and α_4 for the distributions described in Exercise 15, page 105.

6. If the class interval is taken as a unit, i.e., if $w = 1$, show that:

$$\begin{aligned} \nu_2 &= \nu_2' - b_x^2, \\ \nu_3 &= \nu_3' - 3\nu_2 b_x - b_x^3, \\ \nu_4 &= \nu_4' - 4\nu_3 b_x - 6\nu_2 b_x^2 - b_x^4 \end{aligned}$$

7. Compute M, σ, α_3, and α_4 for the distribution in Exercise 2, page 54.

8. Compute α_3 and α_4 for the data of Exercise 19 at the end of this chapter.

9. Compute α_3 and α_4 for the data of Exercise 24 at the end of this chapter.

10.

Compute M, M_d, M_o, σ, α_3, and α_4 for this table of chest measurements.

THE CHEST MEASUREMENTS OF 10,000 MEN

(Original measurements to the nearest inch) [1]

X	$f(x)$
33	6
34	35
35	125
36	338
37	740
38	1,303
39	1,810
40	1,940
41	1,640
42	1,120
43	600
44	222
45	84
46	30
47	5
48	2
Total	10,000

11.

Compute M, M_d, M_o, σ, α_3, and α_4 for this table of heights.

DISTRIBUTION OF HEIGHTS 6,441 COLORED SOLDIERS

(Original measurements to the nearest centimeter) [2]

X	$f(x)$
148.5	2
150.5	9
152.5	13
154.5	23
156.5	56
158.5	88
160.5	162
162.5	318
164.5	468
166.5	564
168.5	665
170.5	708
172.5	749
174.5	747
176.5	586
178.5	469
180.5	314
182.5	207
184.5	133
186.5	70
188.5	38
190.5	22
192.5	15
194.5	10
196.5	3
198.5	2
Total	6,441

[1] The data are taken from E. T. Whittaker and George Robinson, *The Calculus of Observations*, 1924, p. 189.

[2] The data are taken from *Annual Report of the Surgeon General*, Medical Department of the United States Army, Vol. XV, Pt. I, p. 522.

45. RETROSPECT AND PROSPECT

We have now come to the end of our first important statistical problem, the elementary analysis of a simple frequency distribution. This analysis has been accomplished by computing certain statistical constants and making simple and concise statements about them. If a distribution is fairly symmetrical, the arithmetic mean and the standard deviation are usually sufficient to give a numerical description. If it is skew, then a coefficient of skewness is included. If further refinement is desired, a coefficient of kurtosis is computed. Each computed parameter adds to our information about the distribution in question.

To proceed further into the analysis of a frequency distribution would take us into the study of frequency curves which, as we have previously stated, is beyond the scope of this text. While in a later chapter, we do consider the normal frequency curve (see Chapter 12), the study of skew frequency curves would take us too far afield. This is a topic to which the student trained in the calculus and elementary statistical analysis may look forward.

The second problem that we shall consider is the important problem of correlation. However, before we approach it, we shall deviate somewhat from our course and give a brief consideration to the application of averages to Index Numbers.

MISCELLANEOUS QUESTIONS FOR REVIEW

1. Find the sums:

(1) $\sum_{X=1}^{100} (2X + 5)$ (2) $\sum_{X=10}^{50} (2X^2 - 3X + 6)$

2. What is meant by the statistical analysis of a group of data?

3. What are the purposes of a graphical presentation of a set of statistical data?

4. What is a histogram? A frequency polygon? Give directions for constructing each.

5. What is meant by: "The central tendency of a distribution"? The "dispersion of a distribution"? The "skewness of a distribution"?

6. Define three measures of central tendency; three measures of dispersion; three measures of skewness.

7. From the formula defining the arithmetic mean, derive another formula for M.

8. If $a = 127 \pm 0.2$ and $b = 2.2 \pm 0.3$, find the extreme values of

(1) $a + b$ (2) $a - b$ (3) $a \cdot b$ (4) $\frac{a}{b}$

9. In what type of distributions are Q_1 and Q_3 equally distant from M_d?

10. From the formula defining σ, derive two other formulas for computing σ.

11. To compute σ, is it necessary to compute M?

12. A measurement of the length of a room is recorded 22.6 ± 0.05 feet. What does this statement mean?

13. The arithmetic mean of a sample distribution of 100 grades is written 70 ± 0.6. What does this statement mean? What is the standard deviation of the sample?

14. The standard deviation of a sample distribution of 100 grades is written 8 ± 0.38. What does this statement mean? What is the standard deviation of the sample?

15. Why is the standard deviation a good measure of dispersion?

16. Criticize the following statements:

(1) The range is the most perfect measure of variability because it includes all the measurements.

(2) In constructing a frequency distribution the selection of the class interval is arbitrary.

(3) If the probable error of the mean is attached to the computed mean, the true mean is then exactly known.

(4) If the sum of the frequencies is equal to the count of the original scores, the tabulation is correct.

(5) A score recorded as 80 means that the measure extends from 80 to 81.

(6) If a class is designated "85–89," the correct midpoint would always be 87.5.

(7) $M \pm \sigma$ establishes an interval that always includes about $(\frac{2}{3})N$.

17. The analysis of an approximately normal distribution of the weekly salaries of 600 men gave $M = \$30$ and $\sigma = \$5$.

(1) About how many received salaries between \$25 and \$35?

(2) Assuming that Range $= 6\sigma$, about what was the maximum salary? The minimum salary?

18. The heights and weights of 1,515 men gave two approximately normal distributions with the following statistical constants:

Heights	*Weights*
$N = 1515$	$N = 1515$
$M = 67.92$ inches	$M = 138.88$ pounds
$M_d = 68.02$ inches	$M_d = 137.62$ pounds
$\sigma = 2.43$ inches	$\sigma = 17.2$ pounds

(1) Which distribution shows the greater dispersion? Why?

(2) Which distribution shows the greater skewness? Why?

19.

X	$f(x)$
3.85	3
4.05	41
4.25	127
4.45	303
4.65	524
4.85	852
5.05	1033
5.25	1106
5.45	1137
5.65	983
5.85	799
6.05	532
6.25	281
6.45	177
6.65	80
6.85	37
7.05	16
7.25	3
7.45	3
Total	8037

The accompanying distribution gives the percentage fat content of milk as shown by 8,037 milking records. The data were taken from Bulletin 245 of the University of Illinois Agricultural Experiment Station, p. 603.

Compute: M, M_d, M_o by fitting a parabola, Q_1, Q_3, σ, and Sk.

Find E_M and interpret it. Find E_σ and interpret it.

20.

X	$f(x)$
51	4
52	23
53	59
54	108
55	224
56	257
57	230
58	110
59	38
60	16
61	2
Total	1071

The data in the accompanying table give the head circumference (centimeters) of 1,071 boys. The data were taken from: "The Evaluation of Anthropometric Data," by Winfield S. Hall, *Journal of American Medical Association*, Vol. 37, p. 1646.

Find M, σ, α_3 and α_4 for this distribution.

21. What are two points of view that may be adopted with regard to the statistical analysis of a set of data?

22. Does σ meet Yule's requirements of a good average?

23. A class was given two tests with the following results: $M_1 = 76$, $\sigma_1 = 11$; $M_2 = 59$, $\sigma_2 = 14$. A student made 92 on the first test and 82 on the second test. On which test did he do better?

24. The following distribution presenting the life experience of wooden

telephone poles was adopted from Robley Winfrey and Edwin B. Kurtz: *Life Characteristics of Physical Property*, Bulletin 103, Iowa Engineering Experiment Station, p. 57. Compute M, σ, E_M and E_σ.

Life in Years X	*Number of Poles Replaced* $f(x)$	*Life in Years* X	*Number of Poles Replaced* $f(x)$
1	4	12	95
2	7	13	91
3	15	14	73
4	32	15	64
5	30	16	38
6	57	17	30
7	61	18	18
8	73	19	5
9	96	20	1
10	104	21	1
11	103	22	2
Total			1000

25. Criticize the following statements:

(1) The number 2.340 has four significant figures.
(2) The relative error in a measurement is the ratio of the absolute error to the true value of the quantity measured.
(3) The population of a city was recorded as 300,000 ± 3,000. The percentage error was 3 per cent.
(4) The length of a line was measured twenty times. The arithmetic mean of the measurements gives the true length.
(5) In our notation X indicates class frequency.
(6) The guessed mean, h, should be chosen at the midpoint of a class interval.
(7) Ordinarily the number of class intervals should be more than ten and less than thirty.
(8) It would be possible for three people to get three different frequency distributions from the same data and all be right.
(9) If the sum of the frequencies agrees with the count of the original measurements, the tabulation of the frequency distribution is correct.
(10) The quartile points are used to measure both dispersion and skewness.
(11) No matter what value of h is chosen, the same result will be obtained for σ if the computation is correct.
(12) In symmetrical distributions the first and third quartile points are equidistant from M_d.

(13) The standard deviation is a point, not a distance.
(14) The range of a mound-shaped distribution equals 3σ approximately
(15) The probable error of the mean shows what mistake was probably made in computing M.
(16) The statement $M = 75 \pm 3$ means that the true value of M lies between 72 and 78.
(17) When M is greater than M_d, the skewness is positive. The skewness is also positive if q_2 is greater than q_1.
(18) If the probable error is attached to a statistical constant, the results are then exact.
(19) A distance of 3σ laid off on both sides of M establishes an interval that includes about 99 per cent of the total frequency of a mound-shaped distribution.
(20) For a manufacturer of hats, the mode is a more important measure of central tendency than the arithmetic mean.
(21) M_h of a group of numbers is the reciprocal of M of the group.
(22) M_h of the numbers 2, 3, and 6 is greater than their M.

26. The data of the following tables are taken from Bulletin No. 623 of the U.S. Department of Labor, "Wages, Hours, and Working Conditions in the Bread-Baking Industry, 1934." They present the hourly earnings in December, 1934 of employees distributed as to sex.

Compute M, M_d, σ, and Sk for each distribution.

Class (cents)	*Males* $f(x)$	*Females* $f(x)$	*Males and Females* $f(x)$
0 a.u. 12.5	1	0	1
12.5 a.u. 17.5	6	1	7
17.5 a.u. 22.5	14	3	17
22.5 a.u. 27.5	148	165	313
27.5 a.u. 32.5	509	635	1144
32.5 a.u. 37.5	1517	746	2263
37.5 a.u. 42.5	2615	545	3160
42.5 a.u. 47.5	2325	205	2530
47.5 a.u. 52.5	1853	138	1991
52.5 a.u. 57.5	1698	69	1767
57.5 a.u. 62.5	1387	34	1421
62.5 a.u. 67.5	1418	32	1450
67.5 a.u. 72.5	1169	10	1179
72.5 a.u. 77.5	1052	9	1061
77.5 a.u. 85.0	876	6	882
85.0 a.u. 100	1148	13	1161
100 a.u. 120	465	3	468
120 a.u. 150	147	0	147
Total	18348	2614	20962

Chapter 6

INDEX NUMBERS[1]

46. INTRODUCTION

In the preceding chapters we have devoted no little attention to *variation* as a characteristic of statistical phenomena. In characterizing a frequency distribution, we devoted an entire chapter to the measurement of dispersion, a measurement of the extent to which the individual items *vary* on the average from the arithmetic mean. From one point of view, simple correlation is a study of the *variation* that occurs on the average in one variable when a linearly related variable changes by a given amount. In the study of the normal curve, we must have been impressed with the fact that the equation defining this curve describes a very particular kind of *variation* of a group of measurements from their arithmetic mean. Our formulas for estimating reliability are efforts to define a range of *variation* about a statistical constant within which *fluctuations*, due to pure chance, may be expected to occur according to definite probabilities. Each of these important statistical concepts emphasizes, therefore, a particular kind of variation. Speaking rather broadly, we may say that statistical analysis is largely a study of *variation* in statistical phenomena.

In this chapter we shall still be concerned with variation as a characteristic of our data, but we shall regard the variation in a different manner than we have done previously. Stated in rather general terms, our present objective is the reduction of series of data, more or less complex, to numbers *purely relative* which will facilitate comparison. Thus, we shall be interested primarily in measuring *relative variations* in the magnitudes of statistical groups. The statistical devices by which we do this are called *index numbers.*

47. RELATIVES

In their simplest forms, index numbers are ratios, generally expressed as percentages, of one quantity to another quantity of the

[1] This chapter may be omitted without destroying the continuity.

same kind called the *base*. Index numbers have been most widely employed in the study of price changes, but they also may be employed in the study of variation in unemployment, in production, in building, in manufacturing, — in short, wherever group movements are to be measured.

TABLE 30. PRODUCTION OF MOTOR VEHICLES IN THE UNITED STATES, 1920–1929 [1]

Year (1)	*Number (in thousands)* (2)	*Relatives to 1920* (3)	*Link Relatives* (4)
1920	2227	100	...
1921	1682	76	76
1922	2646	119	157
1923	4180	188	158
1924	3738	168	89
1925	4428	199	118
1926	4506	202	102
1927	3580	161	79
1928	4601	207	129
1929	5622	252	122

Consider the data of Table 30. Column (2) gives the total production (*aggregates*) of motor vehicles produced in the United States in the years 1920–1929. It is readily observed from column (3) that a comparison of the values for different dates with the value at some fixed base, or a study of the variation in production relative to some fixed base, is greatly facilitated by reducing the several aggregates to a series of percentages (*relatives*). If the production in 1920 is taken as the date production and is represented by 100, the production relative for any other year merely expresses the production of that year as a percentage of the production for the base year. That is,

$$\text{Relative for a given year} = \frac{\text{Production for given year}}{\text{Production for base year}} \times 100$$

Thus, each item in column (3) is the ratio of the corresponding item in column (2) to the 1920 production, expressed as a percentage.

[1] The data are taken from *Statistical Abstract of the United States*, 1930, p. 385.

If it is desired to compare the values for each year with those of the preceding year, a *link relative* may be employed. The link relative for any year is constructed by dividing the value in that year by the value in the preceding year, and expressing the result as a percentage. That is,

$$\text{Link relative for a given year} = \frac{\text{Value for given year}}{\text{Value for preceding year}} \times 100$$

Thus, in Table 30, the link relative for 1922 is $\frac{2646}{1682} \times 100 = 157$. In order to distinguish them, the relatives shown in column (3) are called *fixed-base relatives.*

The link relatives thus establish a chain of relatives, each year being tied to the preceding year, and from the link relatives we may obtain a further set of relatives called *chain relatives.* We assign 100 as the chain relative for the first year and define the chain relative for any other year to be the product of the link relative for that year and the chain relative for the preceding year, the product to be divided by 100. It should be evident from the definitions that, when a *single* commodity is involved, the chain relatives are equal to the fixed-base relatives.

Simple relatives may be employed to compare the fluctuations in two or more variables, and to permit the computation of an average price relative. To facilitate the comparison of the fluctuations in the prices of corn and hogs in the United States for the decade 1920–1929 — see Table 31 — we have computed their fixed-base relatives, shown in columns (4) and (5), with the prices in 1920 as the base prices. It can now be seen at a glance how one set of relative prices changes as compared with the other. To explain the behavior of the fluctuations recorded in the table would require other data that are not included here. The numbers in column (6), which are the arithmetic means of the numbers in columns (4) and (5), give the average price relatives based upon the two given commodities. Thus, the general average price of these two commodities was 3 per cent higher in 1924 than in 1920, and was 19 per cent lower in 1923 than in 1920.

The price relatives in Table 31 have been based upon the prices of 1920. Of course the prices for any other year could have been chosen as the bases. The averages of the decade prices, 70.3 cents

TABLE 31. PRICES OF CORN AND HOGS IN THE UNITED STATES FOR THE YEARS 1920–1929, AND THEIR RELATIVES [1]

Year	*Corn (cents per bushel)*	*Hogs (dollars per 100 pounds)*	*Relatives* (1920 = 100)		*Average price relative*
			Price of corn	*Price of hogs*	
(1)	(2)	(3)	(4)	(5)	(6)
1920	67.2	13.91	100	100	100
1921	42.3	8.51	63	61	62
1922	65.8	9.22	98	66	82
1923	72.6	7.55	108	54	81
1924	98.2	8.11	147	58	103
1925	67.4	11.81	101	85	93
1926	64.2	12.34	96	89	93
1927	72.3	9.95	108	72	90
1928	75.2	9.22	112	66	89
1929	78.1	10.16	117	73	95
Totals	703.1	100.78	1050	724	888
Means	70.3	10.08	105	72	89

per bushel for corn and 10.08 dollars per hundred pounds for hogs, would have been more satisfactory bases since they are representative and are less affected by chance variations.

EXERCISES

1. Compute the fixed-base relatives (1909 = 100) for the data of Table 11, page 45.

2. Compute the fixed-base and the link relatives (1909–1910 = 100) for the data of Exercise 18, page 106.

3. With the arithmetic mean of the production as base, compute the fixed-base relatives for the data of Exercise 12, page 57.

4. Using the arithmetic means of columns (2) and (3) as bases, compute the average price relatives for the data of Table 31.

48. DEFINITIONS AND NOTATION

We have defined index numbers to be devices which summarize the relative fluctuations in a *group* of variables. Inasmuch as the essential purpose of an index number is to measure the variation in a

[1] The data are taken from *Statistical Abstract of the United States*, 1930, p. 682 and p. 661.

group of variables, it is probably better practice to employ the terms "relative numbers" and "relatives" when referring to *single* series in terms of a fixed base, and to reserve the term "index number" to describe the variation in a group of variables in combination. The numbers in column (3) of Table 30 may properly be called "relatives" whereas those of column (6) of Table 31 may properly be called "index numbers." While index numbers are sometimes expressed as mere aggregates, yet more generally they are expressed as percentages of the values in an arbitrarily chosen base period.[1]

Many methods may be employed in the construction of index numbers, and there are differences of opinion as to which is the best method. In our treatment, we shall devote the emphasis to the best known methods of construction and attempt to avoid controversial questions. We shall make use of the following symbols:

p_0' = price of the first commodity at time "0" (the base period)
p_i' = price of the first commodity at time "i"
$p_i^{(n)}$ = price of the nth commodity at time "i"
q_0' = quantity of the first commodity at time "0"
q_i' = quantity of the first commodity at time "i"
$q_i^{(n)}$ = quantity of the nth commodity at time "i"
$\frac{p_i'}{p_0'}$ = a price relative (ratio of price of a given commodity at time "i" to the price of the same commodity at time "0," expressed as a percentage)
$\frac{q_i'}{q_0'}$ = a quantity relative
$\Sigma p_i q_i = p_i'q_i' + p_i''q_i'' + \cdots + p_i^{(n)}q_i^{(n)}$
$\Sigma p_i q_0 = p_i'q_0' + p_i''q_0'' + \cdots + p_i^{(n)}q_{0i}^{(n)}$
${}_0P_i$ = the price index for the time "i"
${}_0Q_i$ = the quantity index for the time "i"

49. UNWEIGHTED INDEX NUMBERS

In the construction of unweighted (or *simple*) index numbers, the individual members of the group are all regarded as of equal importance. The influence of no member of the group is to be weighted

[1] In this book we shall assume that relatives and index numbers are expressed as percentages.

by multiplying the member by some quantity or weight. If some members of the group are to be considered as more important than others, we shall apply to the important members *weights* that are expected to reflect their relative importance. Unweighted indices will be considered in this section; weighted, in the next.

A. Simple Aggregative Relatives. An aggregative index number is based upon the sums (aggregates) of the items for the several years. The *aggregative relative* is found by comparing the results thus secured for different dates. If prices are in question, the aggregative relative is given by

$$ {}_0P_i = \frac{\Sigma p_i}{\Sigma p_0} \tag{1} $$

To illustrate the method of computing aggregative relatives, let us consider the data of Table 32, which gives the farm prices in cents

TABLE 32. FARM PRICES IN CENTS PER BUSHEL OF GRAINS IN THE UNITED STATES [1]

Computing the aggregative relatives

Grain	*1921*	*1923*	*1925*	*1927*	*1929*
Corn	42.3	72.6	67.4	72.3	78.1
Wheat	92.6	92.3	141.6	111.5	104.3
Oats	30.2	41.4	38.0	45.0	43.5
Rye	69.7	65.0	78.2	85.3	87.1
Barley	41.9	54.1	58.8	67.8	55.0
Buckwheat	81.2	93.3	88.8	83.5	97.7
Rice	95.2	110.2	153.8	92.9	97.8
Σp_i	453.1	528.9	626.6	558.3	563.5
${}_0P_i = \frac{\Sigma p_i}{\Sigma p_0}$	100	117	138	123	124

per bushel of seven important grains. We find the aggregates Σp_i, of the prices for each of the several years. Choosing 1921 as the base year (where $i = 0$), we find the aggregative relatives $\Sigma p_i / \Sigma p_0$ for the other years and express our results as percentages. We note that the aggregative relative for 1925 is 138. This may be interpreted

[1] The data are taken from *Statistical Abstract of the United States*, 1930, pp. 682–683.

to mean that the farm prices of these grains for 1925 were, on the average, 38 per cent higher than for 1921.

It is evident that the computation of the aggregative relative requires that all items be reduced to the same unit, otherwise we would be combining non-homogeneous things and the sums would have no meaning.[1] To illustrate, consider the following prices of several commodities in 1925:

Anthracite coal	$5.30	per ton (2000 pounds)
Cotton	0.182	per pound
Potatoes	2.10	per bag (100 pounds)
Wheat	1.60	per bushel (60 pounds)

We may reduce these prices to the same unit, and quote them as follows:

Anthracite coal	$00.265	per 100 pounds
Cotton	18.20	per 100 pounds
Potatoes	2.10	per 100 pounds
Wheat	2.67	per 100 pounds

The well-known *Bradstreet's* index is based upon the simple aggregative method, the items being reduced to prices per pound. The aggregates Σp_i, themselves, are the indexes; however, they may be converted into a series of percentages upon any chosen base. It should be noted that the conversion of all prices into *prices per pound* affects a concealed weighting for which there is no logical basis. Thus, in 1925 in an aggregate of per pound prices, a pound of cotton was worth 9 times as much as a pound of potatoes and 69 times as much as a pound of coal. This illogical emphasis given to high-priced articles is somewhat neutralized in *Bradstreet's* index by the introduction of a logical element in that more than one quotation is given for some of the more important commodities and only one for the less important articles.

B. Simple Average of Relatives. Another method of constructing index numbers is that of finding some simple average of the relatives for the given items, the relative for a given commodity at a given time being referred to the same commodity at a certain basic date. We may use the arithmetic mean, the geometric mean, the median,

[1] It should not be assumed that an aggregative relative based upon such a reduction will necessarily present a logical index.

the mode, and the harmonic mean of the relatives. Assuming that a table of actual amounts has been prepared — such, for example, as the prices of Table 32 — the steps involved in the process are:

1. Reduce each item, — price, quantity, value, et cetera, — in the time "i" for which the index is desired to a percentage (relative) of the item for the same commodity in the base period. That is, if prices are in question, find p_i/p_0 for each commodity; if quantities are in question, find q_i/q_0 for each commodity; if values are in question, find v_i/v_0, and *express all the relatives as percentages.*
2. Compute the averages of the relatives found.

The arithmetic mean of the price relatives at time "i" is given by

$$_0P_i = \frac{1}{N}\sum\frac{p_i}{p_0} \tag{2}$$

where N is the number of prices.

The geometric mean of the N price relatives at time "i" is given by

$$_0P_i = \sqrt[N]{\frac{p_i'}{p_0'} \times \frac{p_i''}{p_0''} \times \cdots \times \frac{p_i^{(N)}}{p_0^{(N)}}} = \sqrt[N]{\Pi\frac{p_i}{p_0}} \tag{3}$$

where Π means "the product of such terms as," and is computed with the aid of logarithms.

The median of the relatives at time "i" is, of course, found by arranging the relatives at time "i" in the order of their magnitude. If N is odd, the middle term is the median. If N is even, we define the median to be one half the sum of the two middle terms.

The harmonic mean of the relatives at time "i" is given by the formula

$$_0P_i = \frac{N}{\frac{p_0'}{p_i'} + \frac{p_0''}{p_i''} + \cdots + \frac{p_0^{(N)}}{p_i^{(N)}}} = \frac{N}{\sum\frac{p_0}{p_i}} \tag{4}$$

In column (6) of Table 31 we have shown the arithmetic means of the relatives for the prices of two commodities, corn and hogs, for the years 1921 to 1929 with the year 1920 as the base. When several commodities are being investigated, it is better to arrange the table with the list of commodities in the stub and the "times" in the box heads as was done in Table 32.

As an illustrated problem, consider Table 33 which gives, in the

TABLE 33. PRICE RELATIVES OF GRAINS IN THE UNITED STATES, BASED UPON TABLE 32

(1921 = 100)

Computing simple averages of relatives

Grain	*1921*	*1923*	*1925*	*1927*	*1929*
Corn	100	172	159	171	187
Wheat	100	100	153	120	113
Oats	100	137	126	149	144
Rye	100	93	112	122	125
Barley	100	129	140	162	131
Buckwheat	100	115	109	103	120
Rice	100	116	162	98	103
Totals	700	862	961	925	923
Arithmetic Mean of relatives	100	123	137	132	132
Median of relatives	100	116	140	122	125
Geometric Mean of relatives	100	121	136	130	130

body of the table, the relatives of the farm prices of seven important grains in the United States. These data were derived from Table 32 by methods previously explained. Thus, the

$$\text{price relative for corn in 1923} = \frac{72.6}{42.3} \times 100 = 172$$

$$\text{price relative for buckwheat in 1929} = \frac{97.7}{81.2} \times 100 = 120$$

We continue this process until the table of relatives is complete. We then compute the averages for the several years. For example, the geometric mean of the relatives for 1923 is given by

$${}_0P_i = \sqrt[7]{172 \cdot 100 \cdot 137 \cdot 93 \cdot 129 \cdot 115 \cdot 116}$$

or by

$$\log {}_0P_i = \tfrac{1}{7}[\log 172 + \log 100 + \log 137 + \log 93 + \log 129 + \log 115 + \log 116]$$

Numbers	*Logarithms*
172	2.23 553
100	2.00 000
137	2.13 672
93	1.96 848
129	2.11 059
115	2.06 070
116	2.06 446
	7 14.57 648
$\log {}_0P_i =$	2.08 235
${}_0P_i =$	120.9

EXERCISES

1. Find the aggregative relatives for the production data given in Table 34 using 1921 as the base year.

TABLE 34. PRODUCTION, IN MILLIONS OF BUSHELS, OF GRAINS IN THE UNITED STATES [1]

Grain	*1921*	*1923*	*1925*	*1927*	*1929*
Corn	3069	3054	2916	2763	2622
Wheat	815	797	677	878	807
Oats	1078	1306	1488	1183	1239
Rye	62	63	46	58	41
Barley	155	198	214	266	307
Buckwheat	14	14	14	16	12
Rice	38	34	33	45	40
Totals	5231	5466	5388	5209	5068
Aggregative Relative					

2. Compute the harmonic mean of the relatives for the data of Table 32.

3. Compare the five index numbers that have been computed for the data of Table 32.

4. Verify the production relatives given in Table 35. For this table, compute (1) the arithmetic means, (2) the medians, (3) the geometric means of the relatives for the years 1923, 1925, 1927, and 1929.

[1] The data are taken from *Statistical Abstract of the United States*. 1930, pp. 682–683.

TABLE 35. PRODUCTION RELATIVES OF GRAINS IN THE UNITED STATES, BASED UPON TABLE 34

(1921 = 100)

Grain	*1921*	*1923*	*1925*	*1927*	*1929*
Corn	100	100	95	90	85
Wheat	100	98	83	108	99
Oats	100	121	138	110	115
Rye	100	102	74	94	66
Barley	100	128	138	172	198
Buckwheat	100	100	100	114	86
Rice	100	89	87	118	105

50. WEIGHTING

In our previous discussion we attempted to regard all items as of equal importance, although a concealed, "unconscious" weighting was conceded. We admitted the existence of weights inherent in the data themselves and called attention to the fact that doing nothing about them may lead to illogical results. We shall now consider how the illogical results may be somewhat eliminated by the process of weighting. *Weighting* is the term used to describe the conscious effort to assign to each commodity an influence that, in the final result, is proportionate to its relative importance. The index number that results from conscious weighting is called a *weighted index number.* When no such conscious endeavor is made and each item is permitted to exercise an influence upon the result presumably equal to that of every other item, the index is said to be *unweighted* or *simple.*

The weights are usually determined upon some rational basis such as the quantities produced or consumed in a representative period, an average of the quantities produced or consumed over several periods, or some other criterion. As multipliers, it is obvious that the weights may be abstract numbers, and thus that the weights may be numbers *proportional to* the quantities produced or consumed. The fact that actual quantity figures of production and consumption have become increasingly available within recent decades has tended to encourage their use as weights in index number construction. Two methods

of weighting by quantity figures are widely used: the first is called "weighting by base period quantities," and the second is called "weighting by given period quantities." A third method, that of weighting by an average of the base period quantities and the given period quantities, is growing in favor.

There are two reasons for weighting by base period quantities. In the first place, despite the increasing availability of quantity figures, they are not easily obtained for many commodities for the given period. In the second place, the relative variations in the quantities from period to period are frequently not sufficiently large to result in significant errors in the indexes when the quantities are assumed constant for a few successive periods.

51. WEIGHTED AGGREGATES

If we employ the quantities produced in the base period q_0 as weights, the weighted aggregative relative index number of prices at time "i" is given by

$$ {}_0P_i = \frac{\Sigma p_i q_0}{\Sigma p_0 q_0} \tag{5a} $$

which is merely the ratio of the weighted aggregate at time "i" to the total value in the base period. This is possibly our most widely used index number.

We shall illustrate the use of this formula in Table 36 by constructing the index number based upon the weighted aggregate of actual prices in cents per bushel of grains in the United States for the year 1925 with the year 1921 as the base year. The data are taken from Tables 32 and 34.

If we employ the quantities produced in the given period q_i as weights, the weighted aggregative relative index number of prices at time "i" is given by

$$ {}_0P_i = \frac{\Sigma p_i q_i}{\Sigma p_0 q_i} \tag{5b} $$

We shall find that the weighted aggregative relatives, (5a) and (5b), are basic formulas for the "Ideal" index given in Section 55.

To illustrate the construction of an index of weighted aggregates based upon formula (5b), we request the student to complete Table 37. The data are taken from Tables 32 and 34.

TABLE 36. INDEX NUMBER OF GRAIN PRICES IN THE UNITED STATES FOR 1925

(1921 = base year)

Weighted aggregative method

Grain	*Price 1921* p_0	*Weight* q_0	*Price 1921 times Weight* p_0q_0	*Price 1925* p_i	*Price 1925 times Weight* p_iq_0
Corn	42.3	3069	129 818.7	67.4	206 850.6
Wheat	92.6	815	75 469.0	141.6	115 404.0
Oats	30.2	1078	32 555.6	38.0	40 964.0
Rye	69.7	62	4 321.4	78.2	4 848.4
Barley	41.9	155	6 494.5	58.8	9 114.0
Buckwheat	81.2	14	1 136.8	88.8	1 243.2
Rice	95.2	38	3 617.6	153.8	5 844.4
Totals			253 413.6		384 268.6

Price 1921 is in *cents per bushel.*
Price 1925 is in *cents per bushel.*
Weights are quantities produced in 1921 in *millions of bushels.*

$$\Sigma p_0q_0 = 253\ 423.6 \qquad \Sigma p_iq_0 = 384\ 268.6$$

$$_0P_i = \frac{\Sigma p_iq_0}{\Sigma p_0q_0} = \frac{384\ 268.6}{253\ 413.6} = 151.6$$

TABLE 37. PRODUCTION AND PRICE OF GRAINS IN THE UNITED STATES IN 1921 AND 1925

Grain	*Price (cents)*		*Production (millions of bushels)*			
	1921 p_0	*1925* p_i	*1921* q_0	*1925* q_i	p_0q_i	p_iq_i
Corn	42.3	67.4	3069	2916		
Wheat	92.6	141.6	815	677		
Oats	30.2	38.0	1078	1488		
Rye	69.7	78.2	62	46		
Barley	41.9	58.8	155	214		
Buckwheat	81.2	88.8	14	14		
Rice	95.2	153.8	38	33		
Totals						

EXERCISES

1. (Lovitt and Holtzclaw.) The following table gives the average price and weights (quantity used per year by the average workingman's family) of several important items of food. Using 1913 as the base year, compute

(a) the simple aggregative relative of prices for 1915, 1918, 1920 and 1922,

(b) the simple arithmetic mean of relatives for 1915, 1918, 1920 and 1922,

(c) the weighted aggregative relative index for the years 1920 and 1922.

Commodity	*Unit*	*Weights*	*Prices*				
			1913	*1915*	*1918*	*1920*	*1922*
Sirloin Steak	lb.	15	$0.254	$0.257	$0.389	$0.437	$0.374
Round Steak	lb.	40	.223	.230	.369	.395	.323
Bacon	lb.	13	.270	.269	.529	.523	.398
Eggs	doz.	70	.345	.341	.569	.681	.444
Butter	lb.	76	.383	.358	.577	.701	.479
Milk	qt.	424	.089	.088	.139	.167	.131
Flour	$\frac{1}{8}$bbl.	8	.809	1.029	1.642	1.985	1.250
Potatoes	peck	50	.255	.225	.480	.945	.420
Sugar	lb.	145	.055	.066	.097	.194	.073

2. The following were the retail prices of some foods during 1926 and 1934:

Commodity	*Round Steak*	*Potatoes*	*Beans*	*Butter*	*Coffee*	*Flour*
Unit	*lb.*	*lb.*	*lb.*	*lb.*	*lb.*	*lb.*
Price 1926	$0.36	$0.05	$0.09	$0.58	$0.51	$0.06
Price 1934	.28	.02	.06	.31	.29	.05

Compute the simple indexes to fill in the blanks below, and criticize them.

	1926 = Base Year		1934 = Base Year	
	1926	1934	1926	1934
A.M. of relatives	100	——	——	100
G.M. of relatives	100	——	——	100
Aggre. relative	100	——	——	100

52. WEIGHTED AVERAGES OF RELATIVES

Weighted index numbers may also be computed from weighted averages of relatives. The averages most widely used are the arithmetic mean and the geometric mean. The formulas for the weighted arithmetic and the weighted geometric means are derived immediately from those given on pages 61 and 90 by simply considering the frequencies as weights. Thus, if X_i is any value and w_i its weight, we have

$$M = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\Sigma wX}{\Sigma w} = \frac{\Sigma wX}{N} \tag{6}$$

for the weighted arithmetic mean, and

$$M_g = \sqrt[N]{X_1^{w_1}X_2^{w_2} \cdots X_n^{w_n}} = \sqrt[N]{\Pi X_i^{w_i}} \tag{7}$$

for the weighted geometric mean, where

$$N = w_1 + w_2 + \cdots + w_n$$

Written logarithmically, formula (7) becomes

$$\log M_g = \frac{w_1 \log X_1 + w_2 \log X_2 + \cdots + w_n \log X_n}{w_1 + w_2 + \cdots + w_n}$$

$$= \frac{\Sigma w \log X}{\Sigma w} = \frac{\Sigma w \log X}{N} \tag{8}$$

The weighted harmonic mean is given by

$$M_h = \frac{w_1 + w_2 + \cdots + w_n}{\frac{w_1}{X_1} + \frac{w_2}{X_2} + \cdots + \frac{w_n}{X_n}} = \frac{\Sigma w}{\sum \frac{w}{X}} = \frac{N}{\sum \frac{w}{X}} \tag{9}$$

It should be emphasized that "in weighting individual price relatives, quantities will not serve. The abstract relatives must be weighted by *values*, if the resulting products are to be comparable. For values are in terms of a common dollar unit, while quantities may be expressed in a variety of units."[1]

A. The Weighted Arithmetic Mean of Relatives. An index of this type may be obtained in several ways. We may weight each relative by base-period values, by given period values, or by an

[1] Frederick C. Mills, *Statistical Methods, Revised*, 1938, p. 195.

average of the base period values and the given period values. Weighting by base period values is the method most widely used.

To compute a weighted arithmetic mean of relatives for time "i," weighting by base period values, we multiply each relative p_i/p_0 by the value p_0q_0 of the corresponding commodity in the base period, and express the sum of the products as a relative of the total value in the base period.

We shall illustrate the computation of this type of index in Table 38 by constructing the index of the prices of grains in the United States in the year 1925 with the year 1921 as the base year. The data are taken from Tables 33 and 36.

TABLE 38. INDEX NUMBER OF GRAIN PRICES IN THE UNITED STATES FOR 1925

(1921 = base year)

Weighted arithmetic mean of relatives

Grain (1)	*Relative Price 1921* (2)	*Relative Price 1925* (3)	*Weight* p_0q_0 (4)	*Relative Price 1925 times Weight* (3) × (4) (5)
Corn	100	159	129 818.7	20 641 173.3
Wheat	100	153	75 469.0	11 546 757.0
Oats	100	126	32 565.6	4 103 265.6
Rye	100	112	4 321.4	483 996.8
Barley	100	140	6 494.5	909 230.0
Buckwheat	100	109	1 136.8	123 911.2
Rice	100	162	3 617.6	586 051.2
Totals			253 423.6	38 394 385.1

The relative prices 1921 and 1925 were taken from Table 33.

The weights, values of the respective grains in 1921, were taken from Table 36.

$$\Sigma \text{ (price 1925} \times \text{weight)} = 38\ 394\ 385.1$$
$$\Sigma \text{ weight} = 253\ 423.6$$

$${}_0P_i = \frac{38\ 394\ 385.1}{253\ 423.6} = 151.6$$

which is the same index as that secured from the computations of Table 36.

The equality of values for the indexes secured by the two methods illustrated in Tables 36 and 38 is not a coincidence for the weighted arithmetic mean of relative prices, *weighted by the values in the base year*, is always equal to the relative of aggregates weighted by base year quantities. For

$$\frac{\frac{p_i'}{p_0'} \times p_0'q_0' + \frac{p_i''}{p_0''} \times p_0''q_0'' + \cdots + \frac{p_i^{(n)}}{p_0^{(n)}} \times p_0^{(n)}q_0^{(n)}}{p_0'q_0' + p_0''q_0'' + \cdots + p_0^{(n)}q_0^{(n)}} = \frac{\Sigma p_iq_0}{\Sigma p_0q_0}$$

or, more briefly,

$$\frac{\Sigma \frac{p_i}{p_0} \times p_0q_0}{\Sigma p_0q_0} = \frac{\Sigma p_iq_0}{\Sigma p_0q_0}$$

which was given in (5).

The arithmetical computations of Table 38 could have been considerably reduced by replacing the weights p_0q_0 in column (4) by the numbers 130, 75, 33, 4, 6, 1, 4 to which the weights are approximately proportional (see Theorem II, p. 9). We shall leave it as an exercise for the student to show that the result is

$$_0P_i = 38\,348/253 = 151.6$$

In the construction of Table 38, we made use of relatives and values that previously had been computed from the original data and recorded in Tables 33 and 36. Generally, one is called upon to construct the index from the original data, and we suggest the following form for the work-sheet when the weights are the values of the respective commodities for the base year. Of course if the weights are numbers proportional to the values, columns (7) and (8) can be changed accordingly. For the greatest simplicity in computation, the weights should be expressed as percentages of Σp_0q_0. This will mean that the sum of the weights is 100, and the consequent division can be performed mentally.

It is evident that the numbers in column (8), which are derived by multiplying columns (6) and (7), will give the actual values p_iq_0 only when the relatives given in column (6) are accurate. Since the weights may be numbers proportional to p_0q_0, column (8) should always be found by multiplying (6) and (7) and not by multiplying p_i and q_0.

FORM FOR COMPUTING INDEX NUMBERS

Weighted arithmetic mean of relatives method

Weights: Base year values

Commodity	*Unit*	*Price Base Year*	*Price Given Year*	*Quantity Base Year*	*Relative Price Given Year*	*Weight*	*Product of Weight and Relative Price*
		p_0	p_i	q_0	p_i/p_0	p_0q_0	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1st Commodity		p'_0	p'_i	q'_0	p'_i/p'_0	$p'_0q'_0$	$p'_iq'_0$
2nd Commodity		p''_0	p''_i	q''_0	p''_i/p''_0	$p''_0q''_0$	$p''_iq''_0$
....		...	...	...	.....	.....	.....
nth Commodity		$p^{(n)}_0$	$p^{(n)}_i$	$q^{(n)}_0$	$p^{(n)}_i/p^{(n)}_0$	$p^{(n)}_0q^{(n)}_0$	$p^{(n)}_iq^{(n)}_0$
Totals						Σp_0q_0	Σp_iq_0

Description of data:

Prices base year are in —— units.
Prices given year are in —— units.

$$\text{Index number} = {}_0P_i = \frac{\Sigma p_iq_0}{\Sigma p_0q_0}$$

B. Weighted Geometric Mean of Relatives. A verbal interpretation of formula (8) will point out the steps to be taken in constructing the weighted geometric mean of relatives. The steps are as follows:

1. Compute the relatives for the period "i" for which the index is being constructed.
2. Find the logarithm of each relative.
3. Multiply each logarithm by the given weight.
4. Add the results obtained in Step 3.
5. Divide the total obtained in Step 4 by the sum of the weights. This gives log M_g.
6. Find the antilogarithm of the quantity obtained in Step 5. This is the weighted geometric mean of the relatives.

We shall illustrate the computation of this type of index in Table 39 by constructing the index of the prices of grains in the United States in the year 1925 with the year 1921 as the base year. The relatives have been computed in Table 33. We shall use as weights

the numbers 130, 75, 33, 4, 6, 1, 4 which are proportional to the actual values, given in Table 36, of the commodities in the base year.

TABLE 39. INDEX NUMBER OF GRAIN PRICES IN THE UNITED STATES FOR 1925

(1921 = base year)

Grain (1)	*Relative Price 1925* (2)	*Logarithm of the Relative Price* (3)	*Weight* (4)	*Logarithm times Weight* (3) × (4) (5)
Corn.....	159	2.20 140	130	286.18 200
Wheat.....	153	2.18 469	75	163.85 175
Oats.....	126	2.10 037	33	69.31 221
Rye.....	112	2.04 922	4	8.19 688
Barley	140	2.14 613	6	12.87 678
Buckwheat.....	109	2.03 743	1	2.03 743
Rice	162	2.20 952	4	8.83 808
Totals			253	551.29 513

Relative prices for 1925 were taken from Table 33. The weights are numbers proportional to actual values of the commodities produced in the base year. They were taken from Table 36.

$$\log M_g = \frac{551.29\ 513}{253} = 2.17\ 903$$

$$M_g = 151.2$$

The student will note that the three index numbers we have computed for these data on grains show a slight variation. The methods used in Tables 36 and 38 result in an index of 151.6, whereas Table 37 gives 150.1 and Table 39 gives 151.2. To judge the relative merits of these indexes, we shall consider certain tests in Sections 54 and 55.

In the construction of Table 39, we made use of computations that previously had been made upon the original data. Generally, one is required to construct an index from the original data, and we suggest the following arrangement for the worksheet when computing an index from original data by means of the geometric mean of the relatives.

Form for Computing Index Numbers

Weighted Geometric Mean of Relatives Method

Weights: Numbers Proportional to Base Year Values

Commodity	*Unit*	*Price Base Year*	*Price Given Year*	*Quantity Base Year*	*Value Base Year*	*Relative Price Given Year*	*Weight*	*Logarithm of Relative Price Given Year*	*Logarithm of Relative Price Given Year times Weight*
		p_0	p_1	q_0	p_0q_0	p_1/p_0	w	$\log p_1/p_0$	$w \log p_1/p_0$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1st Commodity		p_0'	p_1'	q_0'	$p_0'q_0'$	p_1'/p_0'	w_1	$\log p_1'/p_0'$	$w_1 \log p_1'/p_0'$
2nd Commodity		p_0''	p_1''	q_0''	$p_0''q_0''$	p_1''/p_0''	w_2	$\log p_1''/p_0''$	$w_2 \log p_1''/p_0''$
.....		....	....	...	.....	.....	....	.....	.....
nth Commodity		$p_0^{(n)}$	$p_1^{(n)}$	$q_0^{(n)}$	$p_0^{(n)}q_0^{(n)}$	$p_1^{(n)}/p_0^{(n)}$	w_n	$\log p_1^{(n)}/p_0^{(n)}$	$w_n \log p_1^{(n)}/p_0^{(n)}$
Totals							Σw		$\Sigma w \log p_1/p_0$

Description of data:

Prices base year are in —— units.
Prices given year are in —— units.

$$\text{Log index number} = \log {}_0P_1 = \frac{\Sigma w \log p_1/p_0}{\Sigma w}$$

$${}_0P_1 =$$

53. SUMMARY AND EXTENSION

In our treatment of the index numbers of prices, we have given consideration to the following types, to some of which we have devoted considerable attention. In addition to the median of the simple relatives, we have considered the following:

1. Simple aggregative relative: $\dfrac{\Sigma p_i}{\Sigma p_0}$
2. Simple arithmetic mean of relatives: $\dfrac{1}{N}\sum\dfrac{p_i}{p_0}$
3. Simple geometric mean of relatives: $\sqrt[N]{\Pi\dfrac{p_i}{p_0}}$
4. Simple harmonic mean of relatives: $\dfrac{N}{\sum\dfrac{p_0}{p_i}}$
5. Weighted aggregative relative: (weights = base period *quantities*) $\dfrac{\Sigma p_i q_0}{\Sigma p_0 q_0}$
6. Weighted aggregative relative: (weights = given period *quantities*) $\dfrac{\Sigma p_i q_i}{\Sigma p_0 q_i}$
7. Weighted arithmetic mean of relatives: (weights = base period *values*) $\dfrac{\Sigma p_i q_0}{\Sigma p_0 q_0}$
8. Weighted geometric mean of relatives: (weights = base period *values*) $\log M_g = \dfrac{\Sigma(p_0 q_0)\log\dfrac{p_i}{p_0}}{\Sigma p_0 q_0}$
9. Weighted harmonic mean of relatives: (weights = base period *values*) $\dfrac{\Sigma p_0 q_0}{\Sigma(p_0 q_0)\dfrac{p_0}{p_i}}$

Other useful types may be developed by devising different systems of weights. Suppose we weight the base period prices p_0 by base period quantities q_0, and the given period prices p_i by the given period quantities q_i. The ratio of the aggregate value in the given period to the aggregate value in the base period gives the value index ${}_0V_i$. We thus have

10. Weighted aggregative relative. Value index: ${}_0V_i = \dfrac{\Sigma p_i q_i}{\Sigma p_0 q_0}$
(weights base period = base period *quantities*)
(weights given period = given period *quantities*)

Other aggregative relatives may be obtained by choosing as weights averages of the base period quantities q_0 and the given period quantities q_i. Thus we may choose as weights

$$\frac{q_0 + q_i}{2}, \quad \sqrt{q_0 q_i}, \quad \frac{2q_0 q_i}{q_0 + q_i}$$

which are respectively the arithmetic mean, the geometric mean, and the harmonic mean of q_0 and q_i. Employing these weights, we have the additional aggregative indexes.

11. Weighted aggregative relative: (weights = $(q_0 + q_i)/2$)
$$\frac{\Sigma p_i \frac{(q_0 + q_i)}{2}}{\Sigma p_0 \frac{(q_0 + q_i)}{2}} = \frac{\Sigma p_i (q_0 + q_i)}{\Sigma p_0 (q_0 + q_i)}$$

12. Weighted aggregative relative: (weights = $\sqrt{q_0 q_i}$)
$$\frac{\Sigma p_i \sqrt{q_0 q_i}}{\Sigma p_0 \sqrt{q_0 q_i}}$$

13. Weighted aggregative relative: (weights = $2q_0 q_i/(q_0 + q_i)$)
$$\frac{\Sigma p_i \frac{2q_0 q_i}{q_0 + q_i}}{\Sigma p_0 \frac{2q_0 q_i}{q_0 + q_i}} = \frac{\Sigma p_i \frac{q_0 q_i}{q_0 + q_i}}{\Sigma p_0 \frac{q_0 q_i}{q_0 + q_i}}$$

The formulas listed in 10, 11, 12 and 13 above take into account not only the varying prices but the varying quantities as well. They have the disadvantage of requiring the quantities q_i at time "i", which are not always available. The formula listed in 11, namely,

$${}_0P_i = \frac{\Sigma p_i (q_0 + q_i)}{\Sigma p_0 (q_0 + q_i)}$$

is the Fisher's 2153 which has met wide approval.[1] Due to its simplicity and the facility of its computation, Professor Fisher has proposed its use as a substitute for his "Ideal" index (see page 198).

In a similar manner we may construct other index numbers that are weighted averages of relatives by devising various systems of weights. In Section 52, we recommended the use of *values* as weights for the abstract relatives. Professor Fisher has outlined the following methods of weighting by values.[2]

[1] Irving Fisher, *The Making of Index Numbers*, 1927, p. 284.
[2] Irving Fisher, *op. cit.*, p. 54.

I. Each weight = base period price × base period quantity: p_0q_0
II. Each weight = base period price × given period quantity: p_0q_i
III. Each weight = given period price × base period quantity: p_iq_0
IV. Each weight = given period price × given period quantity: p_iq_i

We have previously used p_0q_0 as weights for the relatives p_i/p_0 in deriving the arithmetic mean of relatives given by 7, the geometric mean of relatives given by 8, and the harmonic mean of relatives given by 9. Let us now use the values p_iq_i of the given period as weights. We have

14. Weighted arithmetic mean of relatives: $\dfrac{\Sigma p_iq_i \dfrac{p_i}{p_0}}{\Sigma p_iq_i}$
(weights = given period *values* p_iq_i)

15. Weighted geometric mean of relatives: $\log M_g = \dfrac{\Sigma p_iq_i \log \dfrac{p_i}{p_0}}{\Sigma p_iq_i}$
(weights = given period *values* p_iq_i)

16. Weighted harmonic mean of relatives: $M_h = \dfrac{\Sigma p_iq_i}{\Sigma p_iq_i \dfrac{p_0}{p_i}} = \dfrac{\Sigma p_iq_i}{\Sigma p_0q_i}$
(weights = given period *values* p_iq_i)

EXERCISES

1. Compute the value index for grains — formula 10, Section 53 — for the year 1925. The data are given in Table 37.

2. Compute the weighted aggregative relative for the prices of grains by formula 11, Section 53. The data are given in Table 37.

3. TABLE 40. PRODUCTION AND FARM PRICE OF THE PRINCIPAL FARM CROPS IN THE UNITED STATES

Crop	*Unit*	*Production (millions)*			*Unit Price (dollars)*		
		1913	*1921*	*1929*	*1913*	*1921*	*1929*
Corn	bu.	2447	3069	2622	0.69	0.42	0.78
Wheat	bu.	763	815	807	0.80	0.93	1.04
Oats	bu.	1121	1078	1239	0.39	0.30	0.44
Barley	bu.	178	155	307	0.54	0.42	0.55
Rice	bu.	26	38	40	0.86	0.95	0.98
Potatoes	bu.	332	362	357	0.69	0.92	1.31
Apples	bu.	145	99	143	0.98	1.68	1.32
Sweet Potatoes	bu.	59	99	85	0.73	0.88	0.95
Cotton	bale 500 lbs.	14	8	15	61.00	81.00	82.00
Tobacco	lb.	954	1070	1501	0.13	0.20	0.10
Hay	ton	64	82	102	12.43	12.10	12.23

Using 1913 as the base year, compute, for Table 40 on the preceding page, indexes for the years 1921 and 1929:

(1) by formula 5, Section 53
(2) by formula 6, Section 53
(3) by formula 11, Section 53
(4) by formula 10, Section 53
(5) by formula 8, Section 53
(6) by formula 14, Section 53

54. BIAS

It is a well-known theorem of algebra that if A is the arithmetic mean, G the geometric mean, and H the harmonic mean of a set of numbers, then $H < G < A$.[1] That is, the simple harmonic mean is less than the simple geometric mean which, in turn, is less than the simple arithmetic mean. In averaging a group of simple relatives, the arithmetic mean tends to give a value too large and the harmonic mean a value too small to be a fair representation of the facts. In more technical language, the arithmetic mean is said to have an *upward bias* and the harmonic mean a *downward bias*. In contrast with the *weight bias*, to be considered later, the bias arising from the form of average used is called *type bias*.

The existence of bias in the simple arithmetic and the simple harmonic means can be explained in another manner, namely, through the use of the *time reversal test*. The time reversal test requires that the product of the index for any given period on the base period and the index for the base period on the given period should equal unity. In symbols, the time reversal test requires that

$$ {}_0P_i \cdot {}_iP_0 = 1 $$

It is very easy to show that the simple geometric mean of relatives satisfies this test and that the simple arithmetic and the simple harmonic means of relatives do not satisfy it. The simple geometric mean is thus without type bias. It has been observed by the makers of index numbers that, when the simple arithmetic mean and the simple harmonic mean are *crossed* (averaged) geometrically, the bias is considerably reduced. The fact that the simple geometric mean is

[1] For a proof, see Robert W. Burgess, *Introduction to the Mathematics of Statistics*, 1927, p. 101.

without type bias means that the index number obtained as a simple geometric mean of relatives is independent of the period taken as a base. These facts give the geometric mean remarkable merit in index number construction.

When weights are applied in the construction of index numbers, another kind of bias — *weight bias* — appears. Each system of weighting has its bias. Weighting the relatives by base period values p_0q_0 produces downward bias while weighting the relatives by given period values p_iq_i produces upward bias. The weighted arithmetic mean and the weighted harmonic mean of relatives may have both type bias and weight bias. If the base period values are employed in the construction of the weighted arithmetic mean of relatives, the net bias will likely be small. Similarly, if given period values are employed as weights in the construction of the weighted harmonic mean of relatives, the net bias will likely be small. Further, as the net bias of the arithmetic mean of relatives, weighted by base period values, has been observed to be in the opposite direction to the net bias of the harmonic mean of relatives, weighted by given period values, crossing these two indexes geometrically should produce an index practically free from bias.

55. FISHER'S IDEAL INDEX

Let

A = weighted arithmetic mean of relatives: $\dfrac{\Sigma p_iq_0}{\Sigma p_0q_0}$
(weights = base period *values* p_0q_0)

and

H = weighted harmonic mean of relatives: $\dfrac{\Sigma p_iq_i}{\Sigma p_iq_i \dfrac{p_0}{p_i}}$
(weights = given period *values*)

the geometric mean of A and H, $\sqrt{AH}$,

$$ {}_0P_i = \sqrt{\frac{\Sigma p_iq_0}{\Sigma p_0q_0} \cdot \frac{\Sigma p_iq_i}{\Sigma p_0q_i}} $$

is known as Fisher's Ideal Index Number.[1] This index is not only the geometric mean of a weighted arithmetic mean and a weighted harmonic mean of relatives; it is clearly a geometric mean of two

[1] Irving Fisher, *op. cit.*, p. 220.

aggregative relatives. The formula requires both price and quantity data for each period to which the index applies. Since the data for quantities are frequently difficult to secure, the practical usefulness of the Ideal index is to some extent limited.

Interchanging 0 and i throughout the formula, we have

$$ {}_iP_0 = \sqrt{\frac{\Sigma p_0 q_i}{\Sigma p_i q_i} \cdot \frac{\Sigma p_0 q_0}{\Sigma p_i q_0}} $$

Evidently ${}_0P_i \cdot {}_iP_0$ equals unity, and hence the Ideal formula satisfies the requirements of the time reversal test.

A second test of validity — a test strongly recommended by Professor Fisher — is the *factor reversal test.* This test, states Professor Fisher, "ought to permit interchanging prices and quantities without giving inconsistent results — *i.e.*, the two results multiplied together should give the true value ratio." [1]

If in the Ideal formula we replace every p by a q and every q by a p, we have

$$ {}_0Q_i = \sqrt{\frac{\Sigma q_i p_0}{\Sigma q_0 p_0} \cdot \frac{\Sigma q_i p_i}{\Sigma q_0 p_i}} $$

Multiplying ${}_0P_i$ and ${}_0Q_i$ together, we have

$$ {}_0P_i \cdot {}_0Q_i = \frac{\Sigma p_i q_i}{\Sigma p_0 q_0} $$

which is the true value index. Consequently, the Ideal formula meets completely the factor reversal test. This means, of course, that the formula serves equally well for constructing indexes of quantities as for constructing indexes of prices, the quantity index being derived by interchanging p and q in the Ideal formula for ${}_0P_i$.

None of the simple or weighted forms of the elementary indexes — arithmetic mean, harmonic mean, geometric mean — fulfill the requirements of the factor reversal test. It is thus obvious that the strong restrictions imposed by the factor reversal test compel its being ignored in the construction of many highly reputable index numbers.

[1] Irving Fisher, *op. cit.*, p. 72.

CONCLUSION

In our treatment of index numbers, we have not attempted to do more than touch upon the important phases of the subject. No attempt has been made to make the treatment exhaustive. We have consciously tried to avoid controversial issues. The student who desires a comprehensive treatment of the subject should read the following treatises:

Irving Fisher, *The Making of Index Numbers*, Houghton, Mifflin Company, 1927.

Wilford I. King, *Index Numbers Elucidated*, Longmans, Green and Company, 1930.

Wesley C. Mitchell, *Index Numbers of Wholesale Prices in the United States and Foreign Countries*, Bulletin Number 284 of the United States Bureau of Labor Statistics, 1921.

C. M. Walsh, *The Problem of Estimation*, King and Son, London, 1921.

EXERCISES

1. If A is the simple arithmetic mean and H is the simple harmonic mean of a group of relatives, show that their geometric mean, $\sqrt{AH}$, is an index that fulfills the time reversal test. Evaluate this index for eliminating type bias.

2. If A is the simple arithmetic mean and H is the simple harmonic mean of a group of relatives, show that their arithmetic mean, $\frac{A+H}{2}$, and their harmonic mean, $\frac{2AH}{A+H}$, do not satisfy the time reversal test.

3. Using $\sqrt{p_0 p_i}$ as the *base* prices, the simple arithmetic mean of relatives for the base year and for the given year are respectively given by

$$A_0 = \frac{1}{N}\sum \frac{p_0}{\sqrt{p_0 p_i}} \quad \text{and} \quad A_i = \frac{1}{N}\sum \frac{p_i}{\sqrt{p_0 p_i}}.$$

Show that the index $I = A_i/A_0$ fulfills the time reversal test.

4. Show that the simple geometric mean of relatives fulfills the time reversal test.

5. Show that the simple arithmetic mean of relatives and the simple harmonic mean of relatives do not satisfy the time reversal test.

6. Using the results of the computations of Tables 36 and 37, find Fisher's Ideal index for grain prices in the United States in 1925.

7. Using the results of Exercises 3 (1) and 3 (2), page 196, find Fisher's

Ideal index for the prices of the principal farm crops in the United States in 1921. Do the same for 1929.

8. TABLE 41. PRODUCTION AND WHOLESALE PRICE OF MINERAL PRODUCTS IN THE UNITED STATES FOR THE YEARS 1919, 1921, AND 1923

Product	*Unit*	*Production (millions)*			*Unit Price (dollars)*		
		1919	*1921*	*1923*	*1919*	*1921*	*1923*
Pig Iron.....	long ton	30.6	16.6	40.0	28.97	22.58	26.29
Copper..	lb.	1286.0	575.0	1667.0	0.19	0.13	0.15
Anthracite Coal	short ton	88.1	90.5	95.5	8.27	10.53	10.98
Bituminous Coal	short ton	465.9	415.9	564.2	2.34	2.19	2.27
Coke.....	short ton	44.2	25.3	55.5	4.58	3.45	5.34
Petroleum	bbl.	378.4	469.6	732.4	2.28	1.70	1.44

With 1919 as the base year, compute indexes for the years 1921 and 1923:

(1) by formula 5, Section 53
(2) by formula 6, Section 53
(3) by the formula for Fisher's Ideal using the results of the two preceding indexes
(4) by formula 10, Section 53
(5) by formula 11, Section 53

9. (Davies and Crowder.) Compute the weighted aggregative relative index number of the prices of farm products in Iowa in 1925 on a 1910–1914 base.

Commodity	*Weights* q	*Prices 1910–1914* p_0	*1925* p_1	p_0q	p_1q
Hogs	5.17 cwt.	$7.30	$11.08		
Cattle	3.85 cwt.	6.39	8.43		
Sheep.....	0.21 cwt.	4.51	7.48		
Corn	24.98 bu.	0.53	0.86		
Oats.....	19.12 bu.	0.35	0.39		
Wheat	1.03 bu.	0.85	1.44		
Hay.....	0.10 ton	9.82	11.23		
Butter	40.62 lb.	0.25	0.41		
Eggs	19.56 doz.	0.17	0.27		
Poultry ...	14.58 lb.	0.10	0.18		
. *Total*					

10. Show that the simple aggregative relative fulfills the time reversal test.

11. Show that the weighted aggregative relative (weights the base period quantities) does not fulfill the time reversal test nor the factor reversal test.

12. Show that the weighted aggregative relative (weights the given period quantities) does not fulfill the time reversal test.

13. The following data are taken from the *Statistical Abstract of the United States*, 1936, p. 632.

Compute the price indexes for these data by using (1) formula 5(a) and (2) formula 5(b).

Grain	*Price (cents)*		*Production (millions of bushels)*					
	1930 p_0	*1934* p_1	*1930* q_0	*1934* q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
Corn	59.6	81.5	2080	1478				
Wheat....	67.1	84.8	886	526				
Oats	32.2	48.0	1275	542				
Rye.....	44.5	71.8	45	17				
Barley....	40.5	68.6	300	117				
Buckwheat	78.8	58.6	7	9				
Rice	78.4	79.0	45	39				
Total								

Chapter 7

LINEAR TRENDS

56. INTRODUCTION

The foregoing chapters have been devoted mainly to the problem of securing a brief numerical description of the simple frequency distribution. We have been enabled to describe the characteristic properties of a distribution — the central tendency, the variability, the skewness, and the excess — by means of a few statistical constants. More briefly, we may say that we have been able to compress the relevant information into four measures:

$$M, \quad \sigma, \quad \alpha_3, \quad \alpha_4 - 3,$$

that are essentially the first four moments of the distribution. Additional information could be secured by fitting an appropriate frequency function to the observed data. Inasmuch as the general problem of describing frequency distributions by means of equations is beyond the scope of this text, no such refinements will be generally attempted.[1]

Certain types of data, however, admit descriptions by means of simple equations, and it is to them that we now turn our attention.

It should be kept in mind that our problem here is inverse to a kindred problem in elementary algebra. There we were given the equation that expressed the relationship between X and Y. We found sets of values of X and Y, plotted them, and drew the graph which was a pictorial representation of the relationship. Here, we have the pairs of values that have come from observation; we plot them. They seem to lie upon or nearly upon a regular curve; that is, there is an apparent mathematical relationship between the variables. What is the equation that expresses exactly or approximately this relationship? Is there a summarizing constant that can be used to measure the degree of this relationship?

[1] Distributions that may be appropriately represented by the normal curve are considered in Chapter 12.

In this chapter we shall be concerned with data that, we assume, obey the simplest mathematical law, the linear or straight-line law. Before we proceed to the real problem of the chapter, it is advisable that we devote some attention to some of the analytical properties of a straight line.

57. SOME CHARACTERISTIC PROPERTIES OF A STRAIGHT LINE

If two values of a variable, X, are given, we denote their difference by ΔX (read: delta ex). This does not mean Δ multiplied by X. It is merely a short way of writing, "the difference between the two values of X." Thus, if the values of X are 5 and 9, then:

$$\Delta X = 9 - 5 = 4$$

In general, if X_1 and X_2 are two values of X:

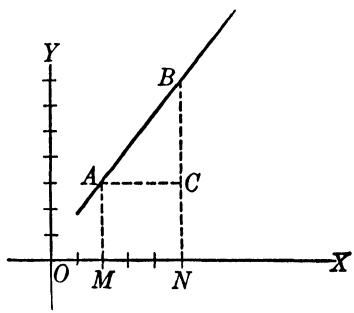
$$\Delta X = X_2 - X_1$$

(Unless otherwise specified, a difference designated by Δ will be taken in the order *second minus first.*) Similarly, ΔY means "the difference between the two values of Y." Thus, if the values of Y are -2 and 4:

$$\Delta Y = 4 - (-2) = 6$$

Consider the line AB of Figure 20 with the two points A (2, 3) and B (5, 7) upon it.

FIGURE 20



$$\Delta X = 5 - 2 = 3 = AC = MN$$
$$\Delta Y = 7 - 3 = 4 = CB$$

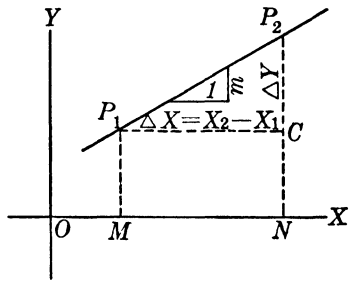
For the more general case, we have for the two points $P_1(X_1, Y_1)$ and $P_2(X_2, Y_2)$. Here:

$$\Delta X = X_2 - X_1 = P_1C = MN, \text{ and}$$
$$\Delta Y = Y_2 - Y_1 = CP_2$$

For any two points on a straight line, the ratio $\frac{\Delta Y}{\Delta X}$ gives the *slope* of the line (see Figure 21). It is usually designated by m. Hence:

$$m = \text{slope of } P_1P_2 = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{Y_1 - Y_2}{X_1 - X_2} \tag{1}$$

FIGURE 21



Thus the slope of a line between two points is equal to the difference of the Y-coördinates of the points divided by the difference of their X-coördinates, *subtracted in the same order*. It also means the change in Y due to a unit change in X.

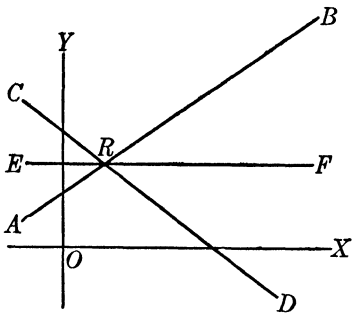
EXERCISES

Draw the lines determined by the following pairs of points, and find their slopes:

1. (3, 2) and (5, 7)
2. (− 2, − 3) and (3, 2)
3. (3, 2) and (5, − 7)
4. (− 2, 3) and (5, − 7)
5. (− 2, 3) and (2, 3)
6. (3, − 4) and (− 2, − 4)
7. Construct a line through (0, 0) with the slope equal to 2.
8. Construct a line through (0, 3) with the slope equal to 2.
9. Assuming that $Y = 3X + 4$ is the equation of a straight line, find its slope. (Hint: Find two points upon the line.)
10. Assuming that $Y = -2X + 4$ is the equation of a straight line, find its slope.
11. Prove by means of slopes that (1, − 3), (2, 3), and (3, 9) lie on the same straight line.

It was probably observed in the exercises on page 205 that the slope of a line may be positive, negative, or zero. If the line rises as we proceed from left to right, ΔY and ΔX have the same sign, and the slope is positive. If the line falls as we proceed from left to right, ΔY and ΔX have opposite signs, and the slope is negative. If the line is horizontal as we proceed from left to right, $Y_2 = Y_1$, and hence the slope equals zero.

FIGURE 22



Thus in the figure we have three lines through the point R. The slope of AB is positive; the slope of CD is negative; the slope of EF is zero.

In solving Exercise 11 on page 205, the student probably assumed that if two segments P_1P_2 and P_2P_3 have a point P_2 in common, and the same slope, the three points P_1, P_2, and P_3 are in the same straight line. This theorem and its converse are *characteristic properties of a straight line.*

58. THE EQUATION OF A STRAIGHT LINE

In elementary algebra the student has drawn graphs of certain given equations. Our problem now is *to find the equation when the graph is given.* That is, we must express in some algebraic way the relation between X and Y of any point on the line.

For example, if a point is anywhere on the X-axis, the Y-coördinate is always zero. We express this simply by the equation:

$$Y = 0$$

This equation is therefore the equation of the X-axis.

Similarly, the equation of the Y-axis is:

$$X = 0$$

What is the equation of a line parallel to the X-axis and two units above it?

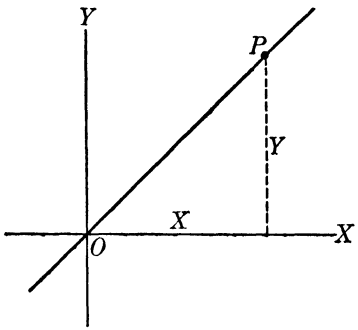
What is the equation of a line parallel to the Y-axis and two units to the right of it?

Again, if a line bisects the first and third quadrants, evidently

$$Y = X$$

for any point P on the line. That is, $Y = X$ is the equation of the line which bisects the first and third quadrants (see Figure 23).

FIGURE 23



Let us now find the equation of the line given in Exercise 8 on page 205. We have $B = (0, 3)$, and $m = 2$.

Let $P(X,Y)$ be any point on the line (see Figure 24).

By definition:

$$\text{the slope} = \frac{Y - 3}{X - 0} = 2$$

or

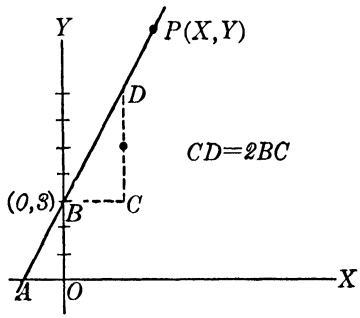
$$Y = 2X + 3$$

Note that if the equation is solved for Y, the slope is the coefficient of X.

The distance OB cut off on the Y-axis is called the *Y-intercept.* The Y-intercept in the equation above is the constant term, 3.

What is the X-intercept, OA?

FIGURE 24



We shall now turn to the problem of finding the equation of the line through $(0,b)$ with the slope equal to m, that is, the line whose Y-intercept is b and whose slope is m.

Let $P(X,Y)$ be any point on the line.

By definition:

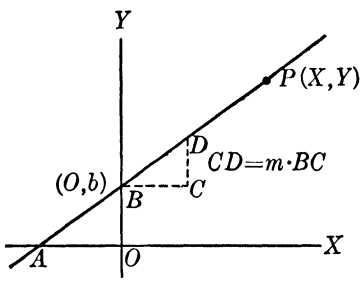
$$\text{the slope} = \frac{Y - b}{X - 0} = m$$

or

$$Y = mX + b \qquad (2)$$

Equation (2) is known as the *slope-intercept equation of the straight line.*

FIGURE 25



If the fixed point is not on the Y-axis the equation takes a different form. Suppose we wish to find the equation of the line through the point (2, 1) with the slope equal to 3 (see Figure 26).

We let $P(X, Y)$ be any point on the line.

FIGURE 26

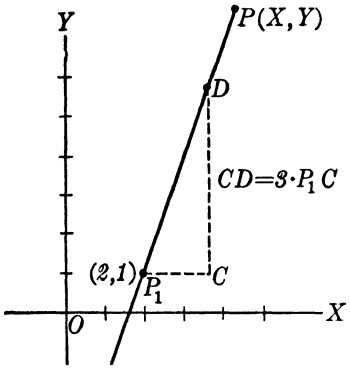


FIGURE 27

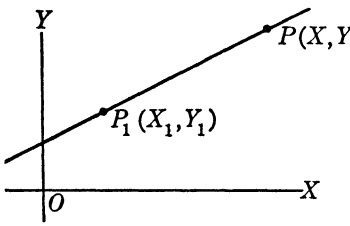
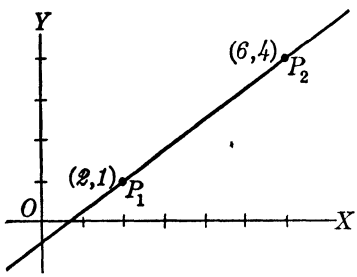


FIGURE 28



By definition:

$$\text{the slope} = \frac{Y-1}{X-2} = 3$$

or $$Y - 1 = 3(X - 2)$$

and finally

$$Y = 3X - 5$$

What is the X-intercept of the line? the Y-intercept?

In general, let $P_1(X_1, Y_1)$ be the fixed point, and m the given slope.

As before, let $P(X, Y)$ be any other point on the line. Then, by definition:

$$\text{the slope} = \frac{Y - Y_1}{X - X_1} = m$$

or $$Y - Y_1 = m(X - X_1) \qquad (3)$$

The equation (3) is called the *point-slope equation of the straight line.* Of course (2) is a special case of (3).

The point-slope form is very useful in finding the equation of a line when two points on the line are given. We can determine the slope by equation (1), then we may use equation (3) with either of the given points as the point $P_1(X_1, Y_1)$.

Thus, let us find the equation of the line through the points (2, 1) and (6, 4).

Here we have:

$$\text{the slope} = m = \frac{4-1}{6-2} = \frac{3}{4}$$

Now using equation (3), we have either:

a. $Y - 1 = \frac{3}{4}(X - 2)$

or

b. $Y - 4 = \frac{3}{4}(X - 6)$

In either case, we obtain:

$$3X - 4Y = 2$$

What are the X- and Y-intercepts of this line?

EXERCISES

1. Construct the line through (0, 2) with $m = 3$, and find its equation.
2. Construct the line through $(0, -2)$ with $m = 3$, and find its equation.
3. Construct the line through $(0, 2)$ with $m = -3$, and find its equation.
4. Construct the line through $(0, -2)$ with $m = -3$, and find its equation.
5. Determine the type form of each of the following equations. Name the two conditions given. Use that knowledge in drawing the graph.

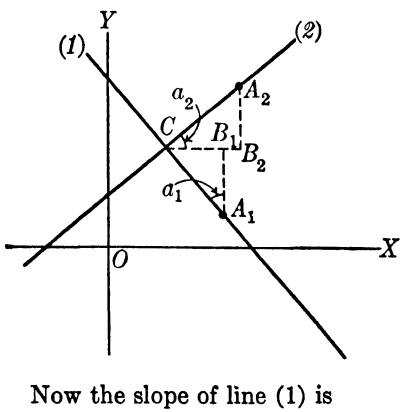
a. $Y = 3X - 4$
b. $Y - 3 = 2(X - 5)$
c. $Y = 2X$
d. $Y = \frac{3}{2}X$
e. $Y - 2 = 3(X + 4)$
f. $Y = X + 5$

6. State the equations of the straight lines:

a. Through (2, 3) with slope 5
b. Through (0, 5) with slope $\frac{3}{2}$
c. Through (6, 2) with slope -1

7. Show that $AX + BY + C = 0$, $(B \neq 0)$, is the equation of a straight line.
8. A straight line passes through the points (3, 5) and (8, 12). Find its equation and its X- and Y-intercepts.
9. Prove that if two nonvertical lines are parallel, their slopes are equal. State and prove the converse.
10. Prove that if two lines are perpendicular to each other, the product of their slopes is -1. State and prove the converse.

FIGURE 29



Now the slope of line (1) is

$$m_1 = \frac{B_1A_1}{CB_1}$$

Let the two lines intersect at C.

Lay off $CA_1 = CA_2$, and draw the parallels to the axes as shown in the figure. Then:

$$\text{angle } a_1 = \text{angle } a_2$$

(the sides being perpendicular each to each).

Hence the triangles CA_1B_1 and CA_2B_2 are congruent (why?) and

$$CB_1 = B_2A_2$$
$$CB_2 = -B_1A_1.$$

(For CB_2 is positive and B_1A_1 is negative.)

and the slope of line (2) is:

$$m_2 = \frac{B_2A_2}{CB_2}$$

Therefore:

$$m_1m_2 = \left(\frac{B_1A_1}{CB_1}\right)\left(\frac{B_2A_2}{CB_2}\right) = \left(\frac{B_1A_1}{CB_1}\right)\left(\frac{CB_1}{-B_1A_1}\right) = -1$$

The proof of the converse is left as an exercise for the student.

11. Show that $Y = 2X - 2$, $Y = 2X$, and $Y = 2X + 4$ are parallel lines.

12. Show that the lines $3X + 2Y = 6$ and $-2X + 3Y = 6$ are perpendicular.

13. Find the equation of a line through (2, 5) parallel to $Y = 3X + 2$.

14. Find the equation of a line through (2, 3) and perpendicular to $3X - 4Y = 8$.

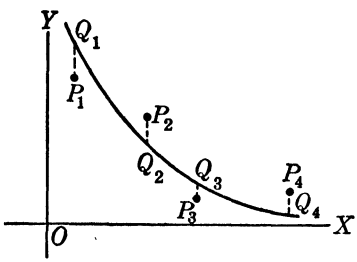
15. Are the points (2, 7), (5, 13), (9, 21), (15, 33) on a straight line?

16. Are the points (1, 5), (3, 10), (5, 13), (7, 16) on a straight line?

59. FITTING A STRAIGHT LINE TO OBSERVED DATA

A. The Method of Least Squares. Many observed data, when plotted, give a set of points that seem to lie somewhat closely upon a curve. (As an illustration, see the data of automobile fatalities on page 103.) This suggests to us that the data may approximately follow some simple mathematical law. It is not necessary that any of the points lie upon the curve selected to describe the data, but they will likely be distributed above and below the curve as the figure indicates.

FIGURE 30



Let P_1, P_2, P_3, P_4, etc. be several points determined by the data. The curve indicating their general trend is called an *empirical curve.* The difference between the ordinate of a given point and the ordinate of the corresponding point on an empirical curve is called the *Y-residual* of that point. That is:

ρ_n (read: rho enn) =

$$\text{any } Y\text{-residual} = \left[\begin{matrix}\text{ordinate of}\\ \text{given point}\end{matrix}\right] - \left[\begin{matrix}\text{ordinate of correspond-}\\ \text{ing point of curve}\end{matrix}\right]$$

Thus the Y-residuals of the points P_1, P_2, P_3, P_4, are respectively P_1Q_1, P_2Q_2, P_3Q_3, P_4Q_4.

Let us consider now the following data, which were derived in an experiment in a physics laboratory in connection with the problem of finding the relation between the resistance in ohms of a certain coil of wire and its temperature, the temperature to be kept between 10° and 100° C.

FIGURE 31

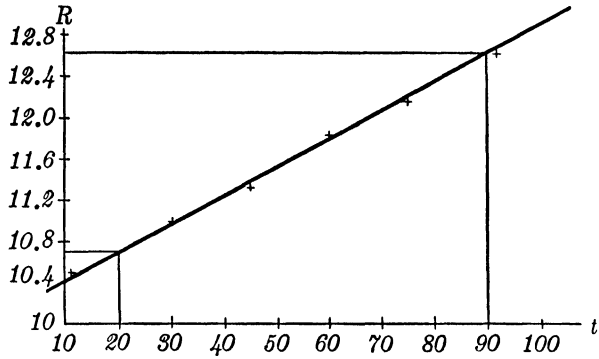


TABLE 42

t	R
10.5	10.42
29.5	10.94
42.7	11.32
60.0	11.80
75.5	12.24
91.1	12.67

When these points are plotted with t as the independent variable and R the dependent variable they lie close to a straight line. (It can be shown by the method of the preceding section that the points do not lie upon a straight line.) Allowing for errors of observation, we may assume that the law connecting resistance and temperature is linear, and our problem now is to determine the equation of the straight line which will best fit the given data.

What is to be regarded as a best fit will depend upon the precise way that we choose to define the term *best*. While there is no unique answer to the question, we shall define the best straight line in accordance with what is called the *principle of least squares*. For the straight line we shall state as follows the principle of least squares: *The straight line best fitting a set of points is that one in which the constants are so determined that they will make the sum of the squares of the residuals a minimum.*[1]

Before we undertake to apply this principle to determine the equation of a straight line best fitting a set of data, let us examine

[1] For other methods of fitting a straight line, see Section 81.

some observed data to which several straight lines have been fitted and, adopting the principle of least squares as a criterion for the goodness of fit, note that we can determine which of several lines is the best.

FIGURE 32

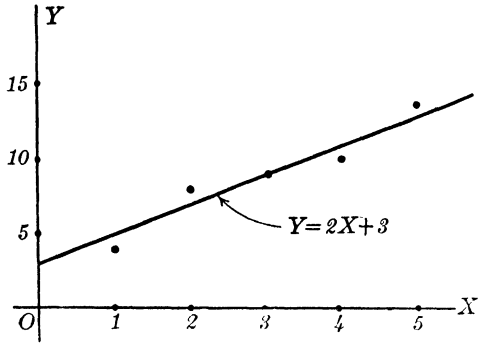


TABLE 43

X	*Observed* Y	*Computed* $Y = 2X + 3$
1	4	5
2	8	7
3	9	9
4	10	11
5	14	13

Consider the observed data given by the first two columns of Table 43. The five points constituting the observed data are plotted on Figure 32. On this set of axes is drawn the line $Y = 2X + 3$. This line has the slope of 2, the Y-intercept of 3, and it passes through the point (3, 9), one of the observed points. Two of the observed points are above the line, two are below the line, and one is on the line. Judging by the graph, the line is a reasonable fit. That is, corresponding to the given values of X the computed values of Y, 5, 7, 9, 11, 13 are reasonably near the corresponding observed values of Y, 4, 8, 9, 10, 14. Stated differently, for the given values of X, the values of Y computed from $Y = 2X + 3$ are reasonably close approximations to the observed values of Y.

Just how near are the observed points, *as a group*, to the line $Y = 2X + 3$? Let us answer this question by applying the principle of least squares to the residuals (see end of page 211). The results are given in Table 44.

We note that, based upon the line $Y = 2X + 3$,

$$\Sigma\rho = 0 \quad \text{and} \quad \Sigma\rho^2 = 4.$$

Suppose that we now consider the line $Y = 2.2X + 2.4$ with the observed values given in Table 45. If the student will plot the observed points and the line $Y = 2.2X + 2.4$ on the same axes, he will observe that this line also passes among the points and that two

of the observed points are above the line, two are below the line, and the line $Y = 2.2X + 2.4$ passes through the observed point (3, 9).

TABLE 44

X	*Observed* Y	*Computed* $Y = 2X + 3$	*Y-Residuals* ρ	*(Y-Residuals)²* ρ^2
1	4	5	− 1	1
2	8	7	+ 1	1
3	9	9	0	0
4	10	11	− 1	1
5	14	13	+ 1	1
			$0 = \Sigma\rho$	$4 = \Sigma\rho^2$

Thus we may say that the line $Y = 2.2X + 2.4$ also fits the data reasonably close. Just how closely does this line fit the data? Again we find the sum of the squares of the residuals by preparing Table 45.

TABLE 45

X	*Observed* Y	*Computed* $Y = 2.2X + 2.4$	*Y-Residuals*	*(Y-Residuals)²*
1	4	4.6	− 0.6	0.36
2	8	6.8	+ 1.2	1.44
3	9	9.0	0.0	0.00
4	10	11.2	− 1.2	1.44
5	14	13.4	+ 0.6	0.36
			$0.0 = \Sigma\rho$	$3.60 = \Sigma\rho^2$

If the algebraical sum of the residuals, $\Sigma\rho$, is adopted as a criterion for the goodness of fit, of the two lines we have considered one fits as well as the other since for each line $\Sigma\rho = 0$. However, if we adopt $\Sigma(Y\text{-residuals})^2$, $\Sigma\rho^2$, as the criterion, then the line $Y = 2.2X + 2.4$ fits more closely than the line $Y = 2X + 3$. As a matter of fact we shall soon have the student show that, based upon the principle of least squares, the line $Y = 2.2X + 2.4$ is the *best fitting* line to the observed data of Table 44.

We shall now proceed to the main problem of the section: adopting as a criterion of the goodness of fit the principle of least squares, to find the values of m and b in order that the line $Y = mX + b$ may best fit a swarm of points. We shall approach the general problem

by considering first a simple set of data, namely, that given in Exercise 16 on page 210. We have the four points as shown in Figure 33.

FIGURE 33

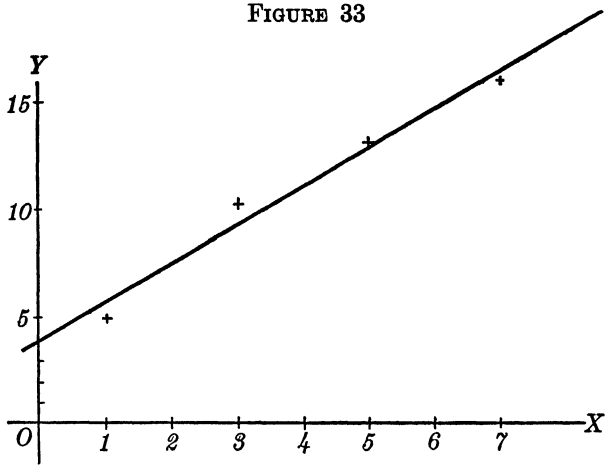


TABLE 46

X	Y
1	5
3	10
5	13
7	16

These data evidently have a straight-line trend. Assume that they can be represented by

$$Y = mX + b$$

for proper values of m and b. Corresponding to $X = 1$, the ordinate of the line is $m \cdot 1 + b$; corresponding to $X = 3$, the ordinate of the line is $m \cdot 3 + b$, and so on. Hence, from the definition:

The 1st Y-residual $= \rho_1 = 5 - (m + b) = 5 - m - b$
The 2nd Y-residual $= \rho_2 = 10 - (3m + b) = 10 - 3m - b$
The 3rd Y-residual $= \rho_3 = 13 - (5m + b) = 13 - 5m - b$
The 4th Y-residual $= \rho_4 = 16 - (7m + b) = 16 - 7m - b$

The values of m and b must be so chosen that the sum of the squares of the Y-residuals is a minimum. The sum of the squares of the Y-residuals is given by:

$$\Sigma\rho^2 = (5 - m - b)^2 + (10 - 3m - b)^2 + (13 - 5m - b)^2 + (16 - 7m - b)^2$$

This result may be written either as a quadratic in b or as a quadratic in m. We have then:

a. $\Sigma\rho^2 = 4b^2 + (32m - 88)b + (84m^2 - 424m + 550)$

b. $\Sigma\rho^2 = 84m^2 + (32b - 424)m + (4b^2 - 88b + 550)$

Recalling the theorem of Section 26 (p. 83) to the effect that $Y = aX^2 + bX + C$ is a minimum when $X = \frac{-b}{2a}$, we note that $\Sigma\rho^2$ in a. is a minimum when:

$$b = \frac{-(32m - 88)}{8} = -4m + 11$$

or
$$4m + b = 11$$

and $\Sigma\rho^2$ in b. is a minimum when:

$$m = \frac{-(32b - 424)}{168} = \frac{-4b + 53}{21}$$

or
$$21m + 4b = 53$$

These equations

$$4m + b = 11$$
$$21m + 4b = 53$$

are called *normal* equations. If they are solved simultaneously, we obtain

$$m = 1.8$$
$$b = 3.8$$

and hence, by the method of least squares, the best-fitting straight line is:

$$Y = 1.8X + 3.8$$

If we give to X in this equation the values 1, 3, 5, 7, we obtain the corresponding *computed* or *most probable values* of Y. Thus:

$$\text{If } X = 1, \quad Y = 1.8 + 3.8 = 5.6$$
$$\text{If } X = 3, \quad Y = 5.4 + 3.8 = 9.2$$

Note that if $X = 4$, $Y = 11$. That is, the point (M_X, M_Y) is on the line.

In the following table

$$\text{any } Y\text{-residual} = Y \text{ observed} - Y \text{ computed}$$

TABLE 47. OBSERVED AND COMPUTED VALUES OF Y COMPARED BY MEANS OF THEIR Y-RESIDUALS

X	Y Observed	Y Computed	Y-Residuals	(Y-Residuals)^2
1	5	5.6	− 0.6	.36
3	10	9.2	0.8	.64
5	13	12.8	0.2	.04
7	16	16.4	− 0.4	.16
$M_X = 4$	$M_Y = 11$		0.0	1.20

EXERCISE

Find the equation of the straight line that best fits the data below. Then find the computed values of Y and compare them with the observed values.

TABLE 48

X	Y
1	1.1
3	6.8
5	12.6
7	19.0

Let us now generalize the procedure by fitting the line

$$Y = mX + b$$

to the data that consist of n sets of values which are given in the table.[1]

FIGURE 34

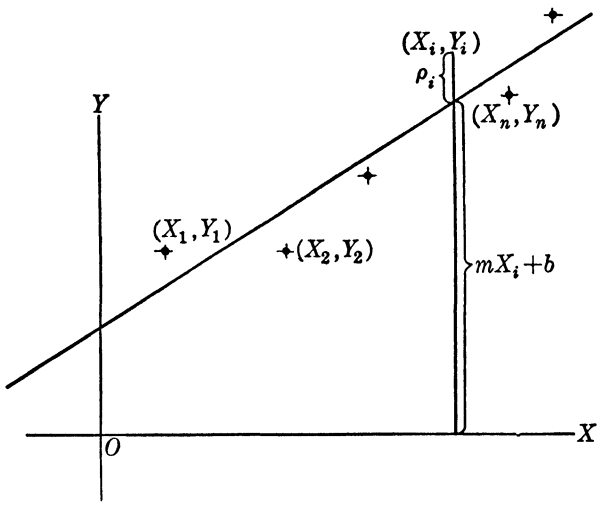


TABLE 49

X	Y
X_1	Y_1
X_2	Y_2
X_3	Y_3
..	..
..	..
X_n	Y_n

We assume that the points have a linear trend, as shown in Figure 34. Our problem is to determine the values of m and b for the best-fitting line.

Corresponding to $X = X_1$, the ordinate of the line is $mX_1 + b$;

[1] The student should note especially that n is the number of pairs of values of Y and X.

corresponding to $X = X_2$, the ordinate of the line is $mX_2 + b$, etc. Hence we have:

$$\rho_1 = \text{1st } Y\text{-residual} = Y_1 - (mX_1 + b) = (Y_1 - mX_1 - b)$$
$$\rho_2 = \text{2nd } Y\text{-residual} = Y_2 - (mX_2 + b) = (Y_2 - mX_2 - b)$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\rho_n = n\text{th } Y\text{-residual} = Y_n - (mX_n + b) = (Y_n - mX_n - b)$$

The values of m and b must be so determined that the sum of the squares of the Y-residuals is a minimum. We will therefore square each residual and find the sum. We have

$$\rho_1^2 = Y_1^2 + m^2X_1^2 + b^2 - 2mX_1Y_1 - 2bY_1 + 2bmX_1$$
$$\rho_2^2 = Y_2^2 + m^2X_2^2 + b^2 - 2mX_2Y_2 - 2bY_2 + 2bmX_2$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\rho_n^2 = Y_n^2 + m^2X_n^2 + b^2 - 2mX_nY_n - 2bY_n + 2bmX_n$$

Adding the above equations, we express $\Sigma\rho^2$ as a quadratic in b and as a quadratic in m. Using the sigma notation, after careful rearrangement of terms, we obtain:

a. $\Sigma\rho^2 = nb^2 + 2[m\Sigma X - \Sigma Y]b + [m^2\Sigma X^2 - 2m\Sigma XY + \Sigma Y^2]$

b. $\Sigma\rho^2 = m^2\Sigma X^2 + 2[b\Sigma X - \Sigma XY]m + [\Sigma Y^2 - 2b\Sigma Y + nb^2]$

From equation a., $\Sigma\rho^2$ is a minimum when

$$b = \frac{-2[m\Sigma X - \Sigma Y]}{2n} = \frac{-m\Sigma X + \Sigma Y}{n}$$

or

$$m\Sigma X + nb = \Sigma Y$$

and from equation b., $\Sigma\rho^2$ is a minimum when:

$$m = \frac{-2[b\Sigma X - \Sigma XY]}{2\Sigma X^2} = \frac{-b\Sigma X + \Sigma XY}{\Sigma X^2}$$

or

$$m\Sigma X^2 + b\Sigma X = \Sigma XY$$

These equations

$$\left.\begin{aligned} m\Sigma X + nb &= \Sigma Y \\ m\Sigma X^2 + b\Sigma X &= \Sigma XY \end{aligned}\right\} \qquad (4)$$

are the *normal equations*.[1] Note that the first can be written by summing the equation $Y = mX + b$, and that the second can be written by multiplying $Y = mX + b$ by X and summing the result.

Solving the normal equations simultaneously, we have:

$$\left.\begin{aligned} m &= \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \\ b &= \frac{\Sigma X^2\Sigma Y - \Sigma X\Sigma XY}{n\Sigma X^2 - (\Sigma X)^2} \end{aligned}\right\} \tag{5}$$

and for the best-fitting straight line

$$Y = \left(\frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}\right)X + \frac{\Sigma X^2\Sigma Y - \Sigma X\Sigma XY}{n\Sigma X^2 - (\Sigma X)^2} \tag{6}$$

The line given by (6) is sometimes called the *line of regression*[2] *of Y on X*.

Let us use the following tabular arrangement to compute the coefficients m and b in (5) and to compare the computed values of Y with the observed values.

TABLE 50

X (1)	*Observed* Y (2)	X^2 (3)	XY (4)	*Computed* Y (5)	*Y-Residuals* (6)	(*Y-Residuals*)^2 (7)
ΣX	ΣY	ΣX^2	ΣXY			

[1] The student familiar with the calculus would derive these equations much more quickly. Thus, if

$$\rho_i = Y_i - (mX_i + b) = \text{the } i\text{th } Y\text{-residual},$$

then: $$\Sigma\rho_i^2 = \Sigma(Y_i - mX_i - b)^2$$

The values of m and b for which $\Sigma\rho^2$ is a minimum are obtained by setting the partial derivatives of $\Sigma\rho^2$ with respect to m and b each equal to zero. We then obtain:

$$m\Sigma X + nb = \Sigma Y$$
$$m\Sigma X^2 + b\Sigma X = \Sigma XY$$

[2] The line of regression of X on Y may be obtained by minimizing the sum of the squares of the X-residuals of the line $X = mY + b$. The properties of this line will be summarized in Section 66 (p. 248).

In connection with Table 50, the following procedure is recommended for *numerical* problems [1]:

1. Compute the values for columns (3) and (4).
2. Find ΣX, ΣY, ΣX^2, and ΣXY.
3. Substitute in (**4**), solve for m and b, and obtain the equation of the best-fitting straight line.
4. Substitute the values of X from column (1) into the equation of the straight line and obtain the *computed* or *most probable* values of Y. This completes column (5).
5. Complete columns (6) and (7) and thus find $\Sigma\rho$ and $\Sigma\rho^2$.

EXERCISE

Apply the foregoing suggestions to the data of Table 46 and Table 48.

B. The Method of Moments. Another very widely used method for fitting a theoretical curve to observed data is the *method of moments.*

Let (X_1,Y_1), (X_2,Y_2), . . ., (X_n,Y_n) be the n points determined by n sets of observed data. If the selected curve is denoted by $Y = f(X)$, the *theoretical* values of Y are $f(X_1)$, $f(X_2)$, . . ., $f(X_n)$. The *principle of moments* (see Section 42, p. 159) says that we shall obtain a good fit if the tth moment about OY, ($t = 0, 1, 2, \ldots, k - 1$), of the n observed values of Y equals the corresponding tth moment about OY of the n theoretical values of Y, k being the number of undetermined constants in the given equation. That is, for $t = 0$ we have the zeroth moments:

$$\Sigma \text{ observed } Y = \Sigma \text{ theoretical } Y$$

or

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} f(X_i)$$

For $t = 1$ we have the first moments:

$$\sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n} X_i f(X_i)$$

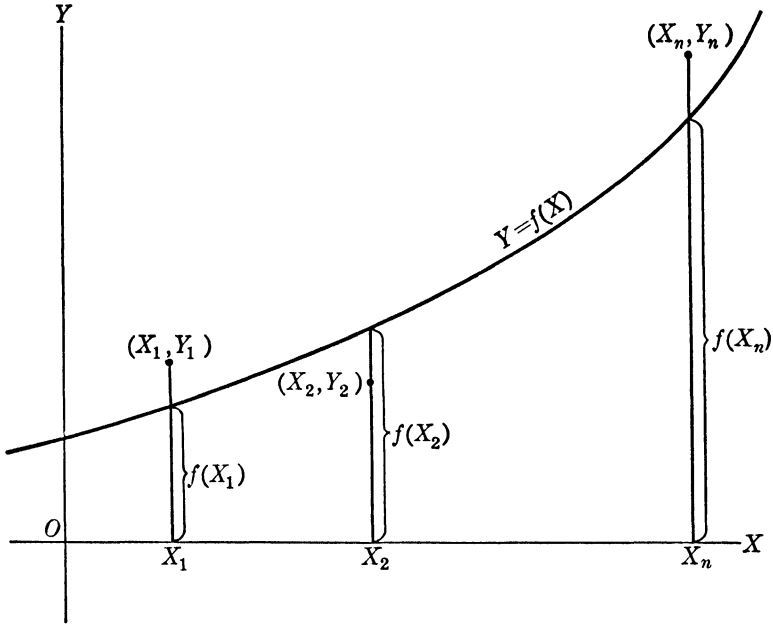
For $t = 2$ we have the second moments:

$$\sum_{i=1}^{n} X_i^2 Y_i = \sum_{i=1}^{n} X_i^2 f(X_i)$$

and so on.

Equations (**5**) and (**6**) are useful for theoretical problems whereas equations (**4**) are better for numerical problems.

FIGURE 35



When the curve $Y = f(X)$ is a straight line $Y = mX + b$, we have

$$\Sigma Y = \Sigma(mX + b) = m\Sigma X + nb$$

and

$$\Sigma XY = \Sigma X(mX + b) = m\Sigma X^2 + b\Sigma X$$

which are the same equations as (**4**).[1] Evidently the suggestions following Table 50 apply here.

EXERCISES

1. Find the equation of the straight line which best fits the temperature resistance data of Table 42 (p. 211).

2. The lengths, l, attained by a certain coiled spring made of steel wire, corresponding to different weights, w, supported by the spring were as shown in the following table. The lengths were measured in centimeters and the weight in grams. Find the linear relation in the form $l = mw + b$.

[1] It can be shown (see "The Method of Moments" by Dunham Jackson, *American Mathematical Monthly*, September, 1923) that if $f(X)$ is a polynomial, the method of moments gives the same solution as the method of least squares.

LENGTHS AND WEIGHTS OF SPRING

w	l	w	l
100	92.2	400	98.3
200	94.3	500	100.2
300	96.2	600	102.3

3. Compute the length of the spring in the table of Exercise 2 for all weights at intervals of 50 grams from 50 grams to 650 grams.

4. Show that the point (M_X, M_Y) is on the line (6), that is, show that the best-fitting line passes through the centroid of the points.

5. Using the values of m and b given in equation (5), show that the sum of the Y-residuals for $Y = mX + b$ is equal to zero.

60. THE STRAIGHT LINE WITH THE ORIGIN AT THE CENTROIDAL POINT

FIGURE 36

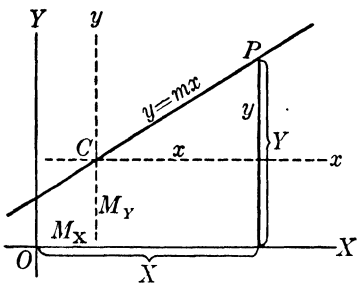
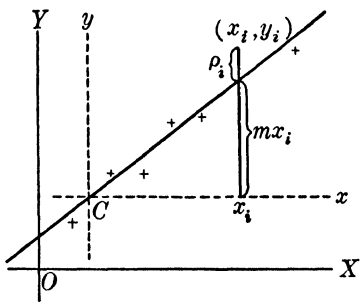


FIGURE 37



The theorem contained in Exercise 4 above states that the best-fitting straight line passes through the centroidal point (M_X, M_Y). Using this point as origin, the equation of the line takes a much simpler form and our further mathematical treatment is greatly simplified.

Denote the centroidal point by C.

If X, Y is any pair of numbers referred to zero as origin, their values referred to C as origin are given by:

$$\left.\begin{aligned} x &= X - M_X \\ y &= Y - M_Y \end{aligned}\right\} \qquad (7)$$

The equation of the line referred to the new origin, C, is of the form

$$y = mx$$

since the y-intercept is zero.

The tabulated data now take the following form:

TABLE 51

X	Y	$x = X - M_X$	$y = Y - M_Y$
X_1	Y_1	x_1	y_1
X_2	Y_2	x_2	y_2
..	..	..	..
..	..	..	..
X_n	Y_n	x_n	y_n

Corresponding to $x = x_1$, the ordinate of the line is mx_1; corresponding to $x = x_2$, the ordinate of the line is mx_2, etc. Hence:

$$\rho_1 = \text{1st } y\text{-residual} = y_1 - mx_1$$
$$\rho_2 = \text{2nd } y\text{-residual} = y_2 - mx_2$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\rho_n = n\text{th } y\text{-residual} = y_n - mx_n$$

We wish to determine m so that the sum of the squares of the y-residuals is a minimum. Evidently:

$$\Sigma\rho_i^2 = \Sigma(y_i - mx_i)^2 = m^2\Sigma x_i^2 - 2m\Sigma x_i y_i + \Sigma y_i^2 \tag{8}$$

The value of m for which $\Sigma\rho_i^2$ is a minimum is, omitting subscripts:

$$m = \frac{\Sigma xy}{\Sigma x^2} \tag{9}$$

and the best-fitting line, referred to the axes through C, is:

$$y = \left(\frac{\Sigma xy}{\Sigma x^2}\right)x \tag{10}$$

If we replace x and y by their values in (7), we obtain the equation of the line referred to axes through $O(0, 0)$ as origin, namely:

$$Y - M_Y = \frac{\Sigma xy}{\Sigma x^2}(X - M_X) \tag{11}$$

We shall illustrate the procedure to be followed should one decide to fit a least-squares line by the x, y method. We shall use data that we have previously considered by the X, Y method. After computing $M_X = 4$ and $M_Y = 11$, we complete the table (see Table 52) by finding the quantities suggested by (9). The algebraic

work follows the table, and we obtain of course the same equation that we found on page 215.

TABLE 52

X	Y	$x = X - 4$	$y = Y - 11$	xy	x^2
1	5	− 3	− 6	18	9
3	10	− 1	− 1	1	1
5	13	1	2	2	1
7	16	3	5	15	9
$M_X = 4$	$M_Y = 11$			36	20

$$m = \frac{\Sigma xy}{\Sigma x^2} = \frac{36}{20} = 1.8$$

$y = 1.8x$ (Equation of line through C as origin)

$Y - 11 = 1.8(X - 4)$

or $Y = 1.8X + 3.8$ (Equation of line referred to axes through O as origin)

We have thus developed two methods of finding the equations of the least-squares line determined by a set of data. We may determine m and b for the line $Y = mX + b$ by using the normal equations (**4**) with the X, Y data, or we may determine m for the line $y = mx$, where x and y are the deviations of X and Y from their respective means: $x = X - M_X$, $y = Y - M_Y$; then replacing x and y by their values we obtain the X, Y equation.

The second method is preferred for *numerical* problems when the values of x and y are such that the arithmetical operations upon them are simpler than when X and Y are used. Thus, if the X, Y data are integral and M_X and M_Y are integral, the values of x and y will be integral and then the table for finding m is decidedly simple to construct. If M_X and M_Y are decimals and the values of x and y are decimals, the second method is to be discouraged.

However for *theoretical purposes* the results of the second, or x, y, method are important and the contents of Section 60 should be mastered.

Let us consider the data of Table 53. We wish to find the X, Y equation for these data.

TABLE 53. THE INDEX NUMBERS OF RETAIL PRICES OF 10 ARTICLES OF FOOD AT TWO DIFFERENT YEARS

Article	*1st year* X	*2nd year* Y	X^2	XY	x	y	x^2	xy
1	88	82	7,744	7,216	4	5	16	20
2	77	71	5,929	5,467	− 7	− 6	49	42
3	91	82	8,281	7,462	7	5	49	35
4	75	70	5,625	5,250	− 9	− 7	81	63
5	95	87	9,025	8,265	11	10	121	110
6	83	77	6,889	6,391	− 1	0	1	0
7	85	77	7,225	6,545	1	0	1	0
8	82	77	6,724	6,314	− 2	0	4	0
9	84	73	7,056	6,132	0	− 4	0	0
10	80	74	6,400	5,920	− 4	− 3	16	12
Total	840 $M_X = 84$	770 $M_Y = 77$	70,898	64,962	0	0	338	282

Using the X, Y values with

$$Y = mX + b$$

the normal equations are

$$840m + 10b = 770$$
$$70{,}898m + 840b = 64{,}962$$

Eliminating b we obtain

$$70{,}560m + 840b = 64{,}680$$
$$70{,}898m + 840b = 64{,}962$$
$$338m = 282$$
$$m = 0.83$$

Substituting we find

$$b = 7.28$$

and our X, Y equation

$$Y = 0.83X + 7.28$$

Using the x, y values with

$$y = mx$$

the normal equation is

$$m = \frac{\Sigma xy}{\Sigma x^2} = \frac{282}{338}$$

$$m = 0.83$$

Our x, y equation is

$$y = 0.83x$$

and our X, Y equation is

$$Y - 77 = 0.83(X - 84)$$

or, simplifying,

$$Y = 0.83X + 7.28$$

Obviously the x, y method leads to the solution more simply than the X, Y method.

The student who has been impressed with the power of the x' method when computing M, σ, α_3, and α_4 will naturally wonder if this method cannot be employed to advantage in this work of fitting straight lines to data. We assure him that the method is an excellent one and we shall present it in the next chapter.

EXERCISES

1. Find the X, Y equation of the least-squares line for each of the following sets of data:

(a)

X	Y
1	2
8	4
14	8
15	9
22	12

(b)

X	Y
5	12
7	15
11	26
13	33
14	34

(c)

X	Y
10	4
8	5
6	7
4	8
2	11

(d)

X	Y
2	47
4	43
6	41
10	37
13	31
15	26
20	20

2. In the following table S is the weight of potassium bromide which will dissolve in 100 grams of water at $T°$ C. Find the relation: $S = mT + b$. Use this equation to estimate S when $T = 50°$.

T	0	20	40	60	80
S	54	65	75	85	96

3. In the following table

X = scores of ten students on a standardized test in secondary algebra taken at the beginning of college

Y = semester grades of the same students in college algebra

Find by two methods the least-squares line for these data. Based upon these data, estimate the semester grade of a student who made 60 on the standardized test.

X	Y	X	Y
54	67	90	91
56	68	63	74
64	74	47	52
33	48	92	90
57	69	34	47

4. A biologist found that the length (in centimeters) of intestines of birds and their weight (in grams) were linearly related. Find the relation $W = mL + b$ for the data given in the following table.

Average Length of Intestines L	*Average Weight of Bird* W	*Average Length of Intestines* L	*Average Weight of Bird* W
4.3	1.5	9.7	6.5
5.8	2.7	10.2	7.3
6.5	3.6	11.0	8.1
7.3	4.2	11.6	8.8
8.4	5.4	12.4	9.7
9.0	5.9	12.6	9.8

5. The latent heat, L, of steam in calories is given for various values of the temperature, T. Find the equation of the best-fitting line for L in terms of T.

T	L	T	L
70	556	110	530
80	550	120	523
90	542	130	515
100	536		

What is the value of L when $T = 75$?

Compare the computed and the observed values of L for the given values of T.

61. FITTING A STRAIGHT LINE TO A TIME SERIES

In Section 17 (p. 43) we encountered series in which time is the independent variable. Several time series were tabulated in Tables 10, 11, 12 (pp. 44–46), and their graphical representations were exhibited in Charts 4, 5, and 6. Further attention to time series has been reserved for this chapter because, after the graphical representation, the next step in the analysis is the determination of the long-time trend, frequently called the *secular trend*, and this is usually accomplished by fitting a straight line to the data. The straight line, of course, should be fitted only to those series which, over a long period, show a general movement in one direction, that is, a general tendency to increase or to decrease.

Over a considerable period of time, many social and economic phenomena show a definite tendency to grow or to decline, that is, they show a definite trend. For example, the population of the

United States (see Table 10) shows a definite tendency to increase, while the percentage decade rate of growth is constantly declining. The production of lumber in the United States (see Table 11) from 1909 to 1922 shows a definite tendency to decline. While the secular trend is usually described by means of a linear function, it must not be supposed that all definite trends are so simply described. Population data, for example, frequently require curves with rather complex equations for their description.

The fact should be emphasized that the secular trend is concerned with the *regular, long-term movements.* True, over a short period of time the movements may vary spasmodically, but the general trend is upward or downward. We are not concerned here with the seasonal variations that are so characteristic of time series, but with the secular trends, and only those that can be described by linear functions.

The computed or trend value of Y at any date is taken as the normal value at that date. It is viewed as the value that would obtain if all temporary and accidental forces were eliminated. The equation

TABLE 54. THE PRODUCTION OF LUMBER IN THE UNITED STATES: COMPUTING THE SECULAR TREND [1]

Year	X	*Production in Board Feet (billions)* Y	X^2	XY	*Computed* Y	ρ
1909	− 6	44.5	36	− 267.0	42.1	2.4
1910	− 5	40.0	25	− 200.0	41.2	− 1.2
1911	− 4	37.0	16	− 148.0	40.3	− 3.3
1912	− 3	39.2	9	− 117.6	39.4	− 0.2
1913	− 2	38.4	4	− 76.8	38.5	− 0.1
1914	− 1	37.3	1	− 37.3	37.6	− 0.3
1915	0	37.0	0	000.0	36.7	0.3
1916	1	39.9	1	39.9	35.8	4.1
1917	2	35.8	4	71.6	34.9	0.9
1918	3	31.9	9	95.7	34.0	− 2.1
1919	4	34.6	16	138.4	33.1	1.5
1920	5	33.8	25	169.0	32.2	1.6
1921	6	27.0	36	162.0	31.3	− 4.3
1922	7	31.6	49	221.2	30.5	1.1
Total	7	508.0	231	51.1		0.4

[1] The data are taken from *Statistical Abstract of the United States*, 1928, p. 689.

of the trend line from which trend values are computed is merely a summarizing expression for the large group of data upon which it is based, and therefore may be used for making estimates of values within the period but may not be at all applicable for making forecasts and predictions. Some new factor may enter at any time and disturb the trend. Therefore, when a trend line is used for extrapolation — that is, for computing values *outside* the given abscissal range — the implications of the line beyond the period of record should be carefully checked against every possible evidence that may influence the factor in question.

The method of fitting a straight line to a time series is illustrated in Table 54. Our problem here is to find the equation of the trend line for the production of lumber in the United States, the data for which were given in Table 11, and graphically presented in Chart 5.

While the origin for X may be chosen at any point, for the sake of simple computation it should be taken at or near the center. If an odd number of years is under consideration, it should be taken at the middle year of the period. If $X = 0$ at 1915, we have

$$n = 14 \qquad \Sigma X^2 = 231$$
$$\Sigma X = 7 \qquad \Sigma XY = 51.1$$
$$\Sigma Y = 508$$

Using formulas (5), we find m and b:

$$m = \frac{14(51.1) - 7(508)}{14(231) - 49} = -0.892$$

$$b = \frac{231(508) - 7(51.1)}{14(231) - 49} = 36.73$$

The equation of the straight line which gives the secular trend is therefore

$$Y = -0.892X + 36.73$$

from which the computed values and the residuals can be found.

Other methods for treating time series will be found in Sections 81 and 87. While the methods we shall present in these later sections make possible the determination of the constants m and b with less arithmetical tedium, we shall present no method that surpasses in precision and reliability that based upon the principle of least squares. In addition to the three important properties (Can you name them?) to which we have referred — casually perhaps — the least-squares

line has the enthusiastic approval of the theory of probability. We cannot say so much for any other line.

EXERCISES

1. The following table gives the annual production of Portland Cement in the United States. [*Statistical Abstract of the United States*, 1930, p. 785.] Find the least-squares line for these data.

Year	*Production (millions of barrels)*	*Year*	*Production (millions of barrels)*
1910	77	1920	100
1911	79	1921	99
1912	82	1922	115
1913	92	1923	137
1914	88	1924	149
1915	86	1925	161
1916	92	1926	165
1917	93	1927	173
1918	71	1928	176
1919	81	1929	171

2. In the following table Y gives the average weekly earnings of shop and office employees in representative New York State factories.

Year	X	Y	*Year*	X	Y
1914	− 3	\$12.48	1918	1	\$20.35
1915	− 2	12.85	1919	2	23.50
1916	− 1	14.43	1920	3	28.15
1917	0	16.37	1921	4	.

(1) Find the equation of the least-squares line for these data.

(2) Based upon this line what were the predicted average weekly earnings in 1921? The actual average weekly earnings were \$25.72.

3. In the following table Y gives (in millions of dollars) the net earnings of the Associated Gas and Electric System, 1920–1928.

Year	X	Y	*Year*	X	Y
1920	− 4	13.4	1925	1	29.5
1921	− 3	16.2	1926	2	33.5
1922	− 2	19.2	1927	3	37.8
1923	− 1	22.7	1928	4	40.6
1924	0	25.1	1929	5	. . .

(1) Find the equation of the least-squares line for these data.
(2) Based upon this line, what were the predicted earnings for 1929? The actual net earnings were 48.5 millions.

4. The number of mules on farms in the United States for the given years is shown in the following table. Choose $X = 0$ at 1932 and find the least-squares line for the data.

Year	*Mules (millions)*	*Year*	*Mules (millions)*
1926	5.9	1933	5.0
1927	5.8	1934	4.9
1928	5.7	1935	4.8
1929	5.5	1936	4.7
1930	5.4	1937	4.6
1931	5.3	1938	4.4
1932	5.1	1939	...

REVIEW EXERCISES

1. Use the relations $X = x + M_X$, $Y = y + M_Y$ with the value of m given by (5) page 218 and thus obtain the value of m given by (9) page 222.

2. State the three most important properties of the least-squares line fitting a swarm of points.

3. Given a set of variates, what is the algebraical sum of the deviations of these variates from their M_X?

4. Given a set of variates, from what value is the sum of the squares of the deviations least?

5. Given a set of variates distributed normally, what per cent of the variates lie within the interval $M_X \pm \sigma_X$? within the interval $M_X \pm 2\sigma_X$? within the interval $M_X \pm 3\sigma_X$?

6. When is it advisable to use the method of Section 60 to find the equation of the least-squares line?

7. What is the unit of measurement of a Y-residual? of a $(Y\text{-residual})^2$? of $\Sigma(Y\text{-residual})^2$ or $\Sigma\rho^2$?

8. Find $\Sigma\rho^2$ for the data on lumber production, Table 54, including the unit of measurement.

9. Do you think $\Sigma\rho^2/n$ can be used to measure the goodness of fit of a curve fitted to a swarm of points? In what unit would it be expressed?

10. What about $\sqrt{\frac{\Sigma\rho^2}{n}}$ as a measure of the goodness of fit? In what unit would it be expressed? Do you think $\sqrt{\frac{\Sigma\rho^2}{n}}$ superior to $\frac{\Sigma\rho^2}{n}$ as a measure of goodness of fit? Why?

11. (a) Show, for the data of lumber production, Table 54, that $\sqrt{\frac{\Sigma\rho^2}{n}} = 2.15$ billions of board feet.

(b) How many values of ρ in Table 54 are numerically less than 2.15?

(c) How many values of ρ are numerically less than 2(2.15)?

12. Can you think of any method whereby we may compare the closeness of fit of straight lines fitted to data expressed in different units, say Tables 53 and 54?

Chapter 8

SIMPLE CORRELATION

62. MEASURES OF CONCENTRATION OF POINTS ABOUT THE LINE OF REGRESSION

In the preceding chapter we have devoted considerable attention to the problem of securing a linear mathematical expression for the relationship between two variables. This equation, called the *line of regression of Y on X*, expresses mathematically the average relationship between the variables.[1] In the exercises and the illustrative examples that we have considered, the points have clustered closely about the regression line. But a line of definite equation may be fitted to points that are quite scattered, widely dispersed with respect to the line. A question immediately presents itself: How can we measure the closeness with which the points cluster about the line? Can we find a measure of the degree of the relationship between the two variables?

This problem is similar to that which arose in connection with the measures of central tendency. We desired to know how great was the concentration of the measures of a distribution about their mean. To measure this concentration we built up several measures of dispersion, recommending especially the standard deviation, which is the square root of the mean of the squares of the deviations from the arithmetic mean.

The line of regression possesses two important properties that are analogous to similar properties of the arithmetic mean. The arithmetic mean is the value such that the sum of deviations from it is zero (see Exercise 5, p. 68); the regression line enjoys the property that the sum of the residuals from it is zero (see Exercise 5 on p. 221). The arithmetic mean is the value such that the sum of the squares of the deviations from it is a minimum (see p. 131); the regression line enjoys the property that the sum of the squares of the residuals from it is a minimum (the principle of least squares).

[1] The line of regression of X on Y will be considered in Section 66 (p. 247).

Owing to the fact that the line of regression possesses the two properties mentioned before, it is frequently called *the line of means.* This name will be more adequately justified in Section 67 (p. 254). Consequently, just as we used

$$\sigma_X = \sqrt{\frac{\Sigma x^2}{N}}$$

(where N is the total frequency) to measure the concentration of the observed X measures about their mean, so we use

$$S_y = \sqrt{\frac{\Sigma \rho_i^2}{n}} \tag{1}$$

(where n is the number of pairs of values of X and Y and where ρ_i = observed Y_i − computed Y_i) to measure the concentration of the observed Y measures about their line of means. S_y is called the *standard error of estimate.* One method of obtaining S_y is illustrated in Table 47 (p. 215) and Table 50 (p. 218).

In Table 47 we have:

$$\Sigma\rho^2 = 1.20 \qquad \text{and} \qquad n = 4$$

therefore

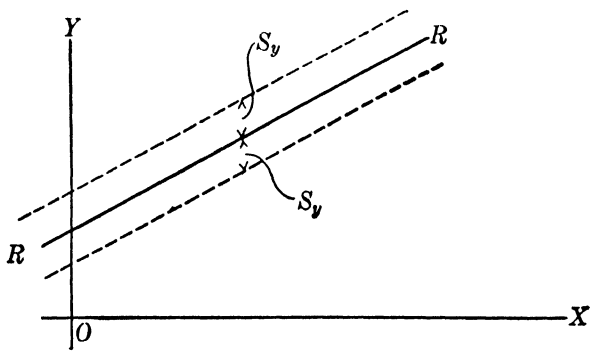
$$S_y = \sqrt{\frac{1.20}{4}} = 0.54772$$

EXERCISE

Find S_y for Exercises 1 and 2 on page 229.

It is evident that the Y-residuals and S_y are expressed in the given Y-unit. To interpret S_y intelligently requires a knowledge of the

FIGURE 38



properties of a normal surface. It is sufficient at this point to state that for a distribution of sufficient size to approximate the normal form, about two-thirds of the points will lie in a strip bounded by two lines on either side of and parallel to the regression line, RR, and a vertical distance, S_y, from it. That is, the odds are 2 to 1 that, for a given X, the observed Y will lie within the zone:

$$(\text{computed } Y) \pm S_y$$

Similarly, a zone established by drawing lines on either side of and parallel to RR and a vertical distance $2S_y$ from it will include about 95 per cent of the points. That is, the odds are 95 to 5 or 19 to 1 that, for a given X, the observed Y will lie within the zone:

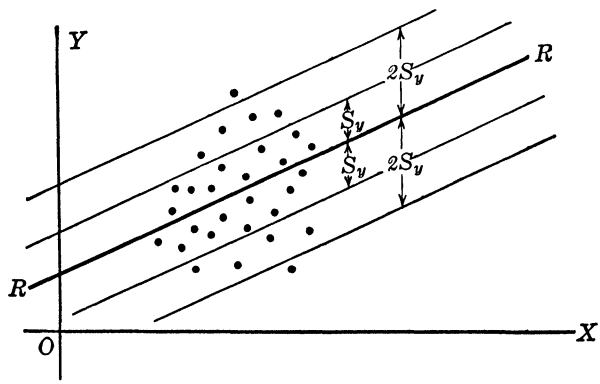
$$(\text{computed } Y) \pm 2S_y$$

If the zone is further enlarged — say $3S_y$ vertically from RR above and below — it is practically certain (odds 385 to 1) that an observed Y will lie within the interval

$$(\text{computed } Y) \pm 3S_y$$

Let us illustrate these statements graphically. On Figure 39 we have plotted thirty points which represent graphically thirty (X, Y) sets of observed data. To these data we have fitted the regression line RR. It will be noted that twenty of the points lie within the zone determined by the parallels to RR and $\pm S_y$ from it. Twenty-eight are within the area determined by the parallels to RR and $\pm 2S_y$ from it. Only two of the points are outside the latter area.

FIGURE 39



Let us look at this matter somewhat differently. It is recalled that the line of regression may be used for purposes of estimating Y for given values of X. (When X is *within* the given abscissal range, we *estimate* Y from the equation; when X is *outside* the given abscissal range, we *predict* Y from the equation.) Thus, the computed value of Y may be an estimate or a prediction. In the language of probability, given an X, the equation gives the best or most probable value of Y.

When we use the regression equation to make estimates or predictions, we naturally are eager to know the degree of confidence to put in our results. Suppose we choose an X and compute Y. The odds are 2 to 1 that the observed Y will not differ numerically from the computed Y by more than S_y. Thus, for Table 54, we have

$$Y = -0.892X + 36.73$$

billions of board feet, and $S_y = 2.15$ billions of board feet. Let $X = 5$. We find $Y = 32.3$ billions of board feet. The odds are 2 to 1 that this value does not differ from the observed $Y(= 33.8)$ by more than $S_y(= 2.15)$. That is the odds are 2 to 1 that the observed Y is within the interval 32.3 ± 2.15 billions of board feet.

Of course if the student wishes to do so, he may use the probable error of estimate instead of the standard error as a measure of the reliability of his estimate. Since the probable error of any parameter is 0.6745 times the standard error of the parameter, we have

$$\begin{matrix}\text{Probable error} \\ \text{of estimate}\end{matrix} = 0.6745 \left(\begin{matrix}\text{Standard error} \\ \text{of estimate}\end{matrix}\right) = 0.6745 S_y$$

There is obviously a consequent change of language. In this case the *chances are even* that for a given X the observed Y will not differ from the estimated Y by more than $\pm 0.6745 S_y$.

EXERCISES

1. Show that $\Sigma\rho^2 = \Sigma[Y - (mX + b)]^2 = \Sigma Y^2 - b\Sigma Y - m\Sigma XY$.
Hint: Make use of the normal equations (4), page 217.

2. (a) Using Exercise 1 above, show that

$$S_y = \sqrt{\frac{\Sigma Y^2 - b\Sigma Y - m\Sigma XY}{n}} \qquad (1')$$

What sigma (Σ) function, not used in finding m and b, is needed to find S_y from formula (1')?

(b) Prove: $S_y = \sigma_\rho$.

3. The following data are taken from *The World Almanac*, 1935, pp. 292 and 310.

X = Savings Bank Deposits in the U.S., 1918–1933.
Y = Number of Strikes and Lockouts in the U.S., 1918–1933.

Year	*Sav. Bk. Dep.* (*billions of* \$'s) X	*S. and L.O.* (*thousands*) Y
1918	5.4	3.3
1919	5.9	3.6
1920	6.5	3.3
1921	6.8	2.4
1922	7.1	1.1
1923	7.7	1.6
1924	8.2	1.2
1925	8.9	1.3
1926	9.3	1.0
1927	9.5	0.7
1928	10.0	0.6
1929	10.1	0.9
1930	10.4	0.7
1931	11.0	0.9
1932	10.9	0.8
1933	10.4	1.6

(1) Find the equation of the regression line.
(2) Interpret the value of m.
(3) Find S_y using formula (1′), and interpret it.
(4) If $X = 8$ find Y. Using S_y interpret your result.
(5) In 1935, $X = 10.6$. Compute Y and compare with the actual $Y = 2.0$.

4.

X	Y
12.5	74
19.8	170
17.3	147
9.9	57
10.9	75
7.5	46
13.7	130
13.1	89
8.5	59
3.8	20
11.9	90
8.6	74
12.1	41
11.9	77
15.6	144

In the adjacent table

X = value of crops (dollars per acre)
Y = value of land and buildings (dollars per acre) in fifteen counties of Illinois in 1930.

(1) Find the equation of the regression line.
(2) Interpret the value of m.
(3) Compute S_y by formula (1′).
(4) Compute Y for $X = 10$, and interpret your result with the aid of S_y.

5. In the following table

X = average yield (bushels per acre) of corn, 1910–1919.
Y = average land value (dollars per acre) on January 1, 1920 in twenty-five counties of Iowa.

X	Y	X	Y
40	87	41	193
36	133	38	203
34	174	38	279
41	285	34	179
39	263	45	244
42	274	34	165
40	235	40	257
31	104	41	252
36	141	42	280
34	208	35	167
30	115	33	168
40	271	36	115
37	163		

(1) Find the equation of the regression line.
(2) Interpret the value of m.
(3) Compute S_y by formula (1′).
(4) Compute Y when $X = 40$, and interpret your result.

63. THE BRAVAIS-PEARSON COEFFICIENT OF CORRELATION

By far the major objection to S_y as a measure of the goodness of fit of a regression line to a swarm of points is this: *it is a concrete number expressed in the given Y-unit.* This fact renders it useless for purposes of comparison. What we really need is an index for measuring the closeness of fit *that is independent of the unit of measure,* a pure number, a relative which will measure the *degree* rather than the amount of the closeness with which the regression line estimates the observed values. We proceed to find such a measure.

To accomplish this end, it is very enlightening to express S_y in terms of the *observed* values. For the sake of simplicity, we shall assume that the observed data are referred to axes through (M_X, M_Y). From equation (8) on page 222 we have:

$$\Sigma\rho^2 = m^2\Sigma x^2 - 2m\Sigma xy + \Sigma y^2$$

where, in terms of the observed values,

$$m = \frac{\Sigma xy}{\Sigma x^2}$$

If this value of m is substituted in the expression for $\Sigma\rho^2$, we obtain:

$$\Sigma\rho^2 = \Sigma y^2 - \frac{2\Sigma xy}{\Sigma x^2} \cdot \Sigma xy + \left(\frac{\Sigma xy}{\Sigma x^2}\right)^2 \cdot \Sigma x^2$$

$$= \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = \Sigma y^2\left[1 - \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}\right]$$

and

$$S_y^2 = \frac{\Sigma\rho^2}{n} = \frac{\Sigma y^2}{n}\left[1 - \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}\right]$$

Recalling that

$$\frac{\Sigma x^2}{n} = \sigma_X^2 \quad \text{and} \quad \frac{\Sigma y^2}{n} = \sigma_Y^2$$

we have:

$$S_y^2 = \sigma_Y^2\left[1 - \frac{(\Sigma xy)^2}{n^2\sigma_X^2\sigma_Y^2}\right] = \sigma_Y^2\left[1 - \left(\frac{\Sigma xy}{n\sigma_X\sigma_Y}\right)^2\right]$$

and finally

$$S_y^2 = \sigma_Y^2(1 - r_{XY}^2) \tag{2}$$

or

$$S_y = \sigma_Y\sqrt{1 - r_{XY}^2}$$

where

$$r_{XY} = \frac{\Sigma xy}{n\sigma_X\sigma_Y} \tag{3}$$

is the well-known Bravais-Pearson coefficient of correlation.[1]

Since x and y are the deviations of the observed values from their respective means, it is evident that r can be very simply computed from the observed values.

The coefficient r plays such an important part in statistical analysis that it is advisable for us to show its relation to the slope, m, of the regression line. Thus:

$$m = \frac{\Sigma xy}{\Sigma x^2} = \frac{nr\sigma_X\sigma_Y}{n\sigma_X^2}$$

or

$$m = r \cdot \frac{\sigma_Y}{\sigma_X} \tag{4}$$

[1] As is our custom we shall omit the subscript XY employing it only for purposes of identification.

The equations of the regression line of Y on X, (10) and (11) of Chapter 7 (p. 222), now become:

$$y = r\frac{\sigma_Y}{\sigma_X} \cdot x \tag{5}$$

and

$$Y - M_Y = r\frac{\sigma_Y}{\sigma_X}(X - M_X) \tag{6}$$

If S_y is taken to be the measure of goodness of fit of the line of regression of Y on X to the observed points, or a measure of the closeness of the relationship of X and Y, it will soon become evident that r is probably a superior measure for this relationship. From equation (2) it is evident that S_y^2 and σ_Y^2 are positive, and therefore r must lie in the interval -1 to $+1$. That is:

$$-1 \leqq r \leqq +1$$

As r approaches unity numerically, S_y decreases toward zero, and this occurs when the points in general cluster closely about the line. As r approaches zero, S_y increases toward its maximum value, σ_Y, and this occurs when the points in general are widely dispersed about the line. If r equals unity numerically, S_y^2 equals zero, hence each residual must equal zero, and the observed points lie upon the line. When r equals unity numerically, we have what is known as *perfect correlation* between the variables X and Y, for the lines of regression then describe the data perfectly.

Therefore a high coefficient of correlation means a small S_y, and consequently a close relationship between Y and X, whereas a low coefficient of correlation means a large S_y, and consequently a poor relationship between X and Y.

Thus we have found our index for we see by (3) that r is a pure number (that is, it is independent of any units of measurement), and hence may be taken as a measure of the *degree* of the relationship between X and Y. It may therefore be used to measure the relationship between variates expressed in any units, as, bushels and dollars, inches and pounds, marks in English and marks in mathematics on different scales, and so on.

In Chapter 7 we learned that if the slope is positive, Y increases as

X increases; if the slope is negative, Y decreases as X increases. From equation (4), since σ_X and σ_Y are always positive, m is positive if r is positive and m is negative if r is negative. Therefore, it follows that if r is positive, Y increases with X, and if r is negative, Y decreases as X increases. The converse of this statement is also true.

The remarks that we have just made about correlation have been from a mathematical standpoint. As we proceed in our study, however, these abstract notions will be clothed with real meaning. We are aware that certain characters tend to rise and fall together as if connected by some direct causal relation — for examples, tall men in general weigh more than short men, young husbands in general are married to young wives, a falling barometer usually signifies an approaching storm, an abnormally small crop in general results in a higher price for the product. In other words, we are aware of the existence of certain persistent relationships between pairs of variables.

The existence of this persistent relationship between paired variables is the important feature of correlation. The variables may in general fluctuate directly or inversely, that is, high values of one variable will in general be paired with high values of the other variable, or high values of one variable will in general be paired with low values of the other — in either case they are said to be correlated.

Therefore:

> Correlation may be defined as tendency toward concomitant variation, and a so-called coefficient is simply a measure of such tendency, more or less adequate according to the circumstances of the case.[1]

In the few preceding pages we have suggested three expressions for this relationship, namely, (1) the equation of the line of regression, (2) the value of the standard error of estimate, and (3) the coefficient of correlation. Each expression has its use, and we shall neglect none of them, but by far the greatest emphasis will be given the coefficient of correlation.[2]

[1] William Brown and G. H. Thomson, *Essentials of Mental Measurement*, 3d ed., 1921, p. 97.

[2] If it is desired, Section 66 (p. 247) may now be read to advantage.

64. COMPUTATION OF r FOR UNGROUPED DATA

Since r plays such an important rôle in the study of relationships, we shall devote considerable attention to its computation and to its interpretation.

The following should be the tabular arrangement for computing r for ungrouped data when the computation is based upon formula (3).

X	Y	$x = X - M_X$	$y = Y - M_Y$	xy	x^2	y^2
$\Sigma X =$ $M_X =$	$\Sigma Y =$ $M_Y =$			Σxy	Σx^2	Σy^2

The following steps should be followed in the arithmetical summary:

1. Find ΣX, then M_X.
2. Find ΣY, then M_Y.
3. Find Σx^2, Σxy, and Σy^2.
4. Find σ_X, σ_Y, and r.
5. Find m from equation (4), or from $m = \Sigma xy/\Sigma x^2$.
6. Write the regression equation of Y on X using equation (6).
7. Obtain the computed values of Y if they are desired.
8. Find S_y from equation (2).

The table on the following page will illustrate the steps recommended in the preceding summary.

We have:

$$n = 15 \qquad \Sigma X = 1402.8 \qquad \Sigma Y = 876.4$$
$$\Sigma xy = -1447.72 \qquad \Sigma x^2 = 2509.21 \qquad \Sigma y^2 = 1852.48$$
$$M_X = 93.5 \text{ bu.} \qquad M_Y = 58.4¢ \qquad \sigma_X = 12.93 \text{ bu.} \qquad \sigma_Y = 11.11¢$$
$$r = -0.672 \qquad S_y = 8.23¢$$
$$m = \frac{-0.672(11.11)}{12.93} = -0.58$$

For the line of regression of Y on X we have:

$$Y - 58.4 = -0.58(X - 93.5)$$

or

$$Y = -0.58X + 112.63$$

TABLE 55. THE AVERAGE YIELD PER ACRE AND THE AVERAGE FARM PRICE PER BUSHEL FOR POTATOES IN THE UNITED STATES, 1900–1914 [1]

Year	*Yield (bushels)* X	*Price (cents)* Y	x	y	xy	x^2	y^2
1900	80.8	43.1	−12.7	−15.3	194.31	161.29	234.09
1901	65.5	76.7	−28.0	18.3	−512.40	784.00	334.89
1902	96.0	47.1	2.5	−11.3	− 28.25	6.25	127.69
1903	84.7	61.4	− 8.8	3.0	− 26.40	77.44	9.00
1904	110.4	45.3	16.9	−13.1	−221.39	285.61	171.61
1905	87.0	61.7	− 6.5	3.3	− 21.45	42.25	10.89
1906	102.2	51.1	8.7	− 7.3	− 63.51	75.69	53.29
1907	95.4	61.8	1.9	3.4	6.46	3.61	11.56
1908	85.7	70.6	− 7.8	12.2	− 95.16	60.84	148.84
1909	106.1	54.1	12.6	− 4.3	− 54.18	158.76	18.49
1910	93.8	55.7	.3	− 2.7	− .81	.09	7.29
1911	80.9	79.9	−12.6	21.5	−270.90	158.76	462.25
1912	113.4	50.5	19.9	− 7.9	−157.21	396.01	62.41
1913	90.4	68.7	− 3.1	10.3	− 31.93	9.61	106.09
1914	110.5	48.7	17.0	− 9.7	−164.90	289.00	94.09
Total	1,402.8	876.4	.3	.4	−1,447.72	2,509.21	1,852.48
Mean	93.5+	58.4+					

We have here a fairly significant coefficient of correlation, $r = -0.672$. Its large numerical value warrants our belief that there does exist a significant relationship between the average yield of potatoes and the corresponding price per bushel. The negative sign, as previously stated, means that as X increases Y decreases. In accordance with our definition of slope, the value of -0.58 for m means that on the average, an increase of one bushel per acre in the yield will mean a diminished price of more than a half a cent per bushel.

Now, let us use our equation for estimating the price that corresponds to a given yield, and S_y for measuring the reliability of the estimate. Let $X = 100$ bu. per acre, then Y estimated is $-0.58(100) + 112.63 = 54.6$ cents. Since $S_y = 8.23$ cents, the odds are 2 to 1 that the observed Y for $X = 100$ does not differ from 54.6 cents by more than 8.23 cents.

[1] The data are taken from *Yearbook of Agriculture*, 1920, p. 616.

EXERCISES

1. The average daily grades and the final examination grades for ten students in a class in calculus are given in the table below.

X = the average daily grade
Y = the grade on the final examination

Find r, the line of regression of Y on X, and S_y. If $X = 90$, $Y = (\ )$.

Student	X	Y	*Student*	X	Y
1	86	71	6	96	94
2	93	76	7	80	71
3	73	61	8	70	60
4	66	52	9	95	85
5	88	75	10	63	55

2. The following table [1] gives the results of experiments performed at Delhi, California, to determine the effect of irrigation upon the yield in alfalfa.

Find r if

X = the total seasonal depth (in inches) of water applied and
Y = the average yield (tons per acre) for the years 1922, 1923, 1924

X	Y	X	Y
12	5.27	36	8.20
18	5.68	42	8.71
24	6.25	48	8.42
30	7.21	60	8.24

3. In the following table [2]

X = the July rainfall (in inches) for the given year for Ohio, and
Y = yield of corn (bushels per acre)

Find r, S_y, and the regression equation. If $X = 4$, $Y = (\ )$. Interpret.

[1] The data are from University of California Experiment Station, *Bulletin* No. 450, p. 8.

[2] The data are taken from *Monthly Weather Review*, Vol. 42 (1914), p. 80.

Year	X	Y	Year	X	Y
1900	4.6	42.6	1905	3.9	37.9
1901	2.7	30.0	1906	5.1	42.2
1902	4.7	38.8	1907	5.4	34.8
1903	3.7	31.5	1908	4.1	36.1
1904	4.1	32.8	1909	3.8	38.7

65. OTHER FORMS OF r

The correlation coefficient, r, as we have defined it by equation (3) of Section 63 is expressed in terms of the deviations of the variates from their respective means, M_X and M_Y. Since M_X and M_Y usually require several decimals for their results, we shall follow the plan that we have used previously in Sections 34 (p. 125) and 44 (p. 164) in computing σ, α_3, and α_4. The labor of computation can be greatly reduced by expressing r in terms of the original variates X and Y, or in terms of x' and y', where x' and y' are deviations in *class units* from some fixed origin (h, k).

In Chapter 4 we have seen that:

$$\sigma_X = \sqrt{\frac{\Sigma X^2}{n} - M_X^2} \quad \text{or} \quad n\sigma_X = \sqrt{n\Sigma X^2 - (\Sigma X)^2}$$

Similarly:

$$\sigma_Y = \sqrt{\frac{\Sigma Y^2}{n} - M_Y^2} \quad \text{or} \quad n\sigma_Y = \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}$$

Also, since

$$x = X - M_X$$
$$y = Y - M_Y$$

we have:

$$xy = XY - M_Y X - M_X Y + M_X M_Y$$

and

$$\Sigma xy = \Sigma XY - M_Y \Sigma X - M_X \Sigma Y + nM_X M_Y$$

Recalling that

$$\Sigma X = nM_X \quad \text{and} \quad \Sigma Y = nM_Y$$

we have:

$$\Sigma xy = \Sigma XY - nM_X M_Y$$

The formula for r can now be expressed in the useful form:

$$r = \frac{\Sigma XY - nM_XM_Y}{\sqrt{\Sigma X^2 - nM_X^2}\sqrt{\Sigma Y^2 - nM_Y^2}} \tag{7}$$

Formula (7) may also be written

$$r = \frac{\dfrac{\Sigma XY}{n} - M_XM_Y}{\sqrt{\dfrac{\Sigma X^2}{n} - M_X^2}\sqrt{\dfrac{\Sigma Y^2}{n} - M_Y^2}} \tag{7'}$$

We shall next derive a formula for r in which the X and Y variates will be expressed in their respective class widths as units and measured from some arbitrary origin (h, k).

Let: C be the centroidal point (M_X, M_Y)

O' be the arbitrary origin (h, k)

w_x = the class width of the X variates

w_y = the class width of the Y variates

$$M_X = h + w_xb_x \quad \text{where} \quad b_x = \frac{\Sigma x'}{n} \quad \text{or} \quad nb_x = \Sigma x'$$

$$\sigma_X = w_x\sqrt{\frac{\Sigma x'^2}{n} - b_x^2}$$

Similarly, we can find:

$$M_Y = k + w_yb_y \quad \text{where} \quad b_y = \frac{\Sigma y'}{n} \quad \text{or} \quad nb_y = \Sigma y'$$

$$\sigma_Y = w_y\sqrt{\frac{\Sigma y'^2}{n} - b_y^2}$$

From Figure 40 we have the following relations:

a. $X = x + M_X$ $\quad Y = y + M_Y$

b. $X = h + w_xx'$ $\quad Y = k + w_yy'$

c. $x = w_xx' - w_xb_x$ $\quad y = w_yy' - w_yb_y$

Applying the relations c. above, we have:

$$xy = w_xw_y(x'y' - b_yx' - b_xy' + b_xb_y)$$

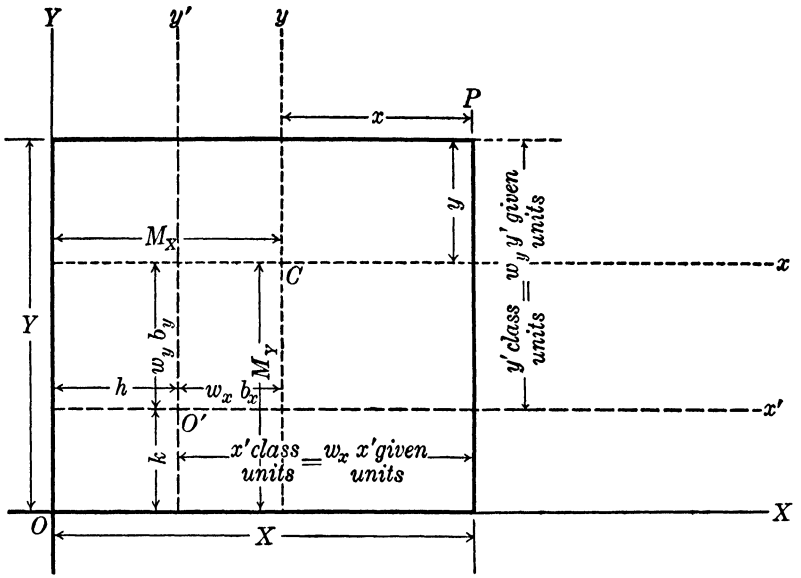
and hence

$$\Sigma xy = w_xw_y(\Sigma x'y' - b_y\Sigma x' - b_x\Sigma y' + nb_xb_y)$$

Substituting $\Sigma x' = nb_x$ and $\Sigma y' = nb_y$, we have:

$$\Sigma xy = w_xw_y(\Sigma x'y' - nb_xb_y)$$

FIGURE 40



Replacing in equation (3) Σxy by the value just found and σ_X and σ_Y by their values in terms of the primed letters, we have:

$$r = \frac{\dfrac{\Sigma x'y'}{n} - b_x b_y}{\sqrt{\dfrac{\Sigma x'^2}{n} - b_x^2}\sqrt{\dfrac{\Sigma y'^2}{n} - b_y^2}} \tag{8}$$

the class widths canceling in the process.

By simple transformations equation (8) reduces to.

$$r = \frac{n\Sigma x'y' - \Sigma x'\Sigma y'}{\sqrt{n\Sigma x'^2 - (\Sigma x')^2}\sqrt{n\Sigma y'^2 - (\Sigma y')^2}} \tag{9}$$

The following example will illustrate the method of procedure for computing r by either formula, (8) or (9).

X = the grade on the first test
Y = the grade on the second test

TABLE 56. GRADES OF TWO TESTS OF 10 STUDENTS IN INTEGRAL CALCULUS

Student	X	Y	x'	y'	$x'y'$	x'^2	y'^2
1	85	77	10	7	70	100	49
2	82	77	7	7	49	49	49
3	91	82	16	12	192	256	144
4	80	74	5	4	20	25	16
5	75	70	0	0	0	0	0
6	95	87	20	17	340	400	289
7	83	77	8	7	56	64	49
8	85	77	10	7	70	100	49
9	88	82	13	12	156	169	144
10	77	71	2	1	2	4	1
Total			91	74	955	1,167	790

Let $h = 75$, $k = 70$, $w_x = 1$, and $w_y = 1$.

We have from the table:

$$\Sigma x' = 91 \quad \Sigma y' = 74 \quad \Sigma x'y' = 955 \quad \Sigma x'^2 = 1167 \quad \Sigma y'^2 = 790 \quad \text{and} \quad n = 10$$

Hence:

$$b_x = 9.1 \qquad b_y = 7.4 \qquad \frac{\Sigma x'y'}{n} = 95.5 \qquad \frac{\Sigma x'^2}{n} = 116.7 \qquad \frac{\Sigma y'^2}{n} = 79$$

Therefore by (8):

$$r = \frac{95.5 - (9.1)(7.4)}{\sqrt{116.7 - 82.81}\,\sqrt{79 - 54.76}} = 0.98$$

66. SUMMARY AND EXTENSION OF THE THEORY OF CORRELATION

In Chapters 7 and 8 we have assumed that our data could be represented by the straight-line equation, $Y = m_1X + b_1$, in which X is the independent variable and Y the dependent variable. By minimizing the sum of the squares of the Y-residuals, we derived the normal equations:

$$m_1\Sigma X + nb_1 = \Sigma Y$$

$$m_1\Sigma X^2 + b_1\Sigma X = \Sigma XY$$

Solving these normal equations for m_1 and b_1, we obtained

$$\left.\begin{aligned} m_1 &= \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \\ b_1 &= \frac{\Sigma X^2\Sigma Y - \Sigma X\Sigma XY}{n\Sigma X^2 - (\Sigma X)^2} \end{aligned}\right\} \quad \text{(5) of Section 59}$$

and hence the equation of the line of regression of Y on X may be found. This equation, for assigned values of X, gives the most probable values for Y. This line (see Exercise 4 on p. 221) passes through the point (M_X, M_Y) and (see Exercise 5 on p. 221) also possesses the property that the sum of the Y-residuals from it is zero.

If the square root of the mean of the squares of the Y-residuals be taken as a measure of the closeness of the concentration of the points about the line, we find:

$$S_y = \sigma_Y\sqrt{1 - r^2} \qquad \text{(2) of Section 63}$$

where

$$r = \frac{\Sigma xy}{n\sigma_X\sigma_Y} \qquad \text{(3) of Section 63}$$

Then:

$$m_1 = r\frac{\sigma_Y}{\sigma_X} \qquad \text{(4) of Section 63}$$

and the equation of the line of regression of Y on X becomes:

$$Y - M_Y = r\frac{\sigma_Y}{\sigma_X}(X - M_X) \qquad \text{(6) of Section 63}$$

In like manner we may arrive at similar results by basing our procedure upon the equation $X = m_2Y + b_2$, where Y is now the independent variable and X is the dependent variable.[1] If we minimize the sum of the squares of the X-residuals we arrive at the normal equations:

$$\begin{aligned} m_2\Sigma Y + nb_2 &= \Sigma X \\ m_2\Sigma Y^2 + b_2\Sigma Y &= \Sigma XY \end{aligned}$$

If these equations be solved for m_2 and b_2, we obtain:

[1] Note that m_2 is not the slope of this line.

$$\left.\begin{aligned} m_2 &= \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2} \\ b_2 &= \frac{\Sigma Y^2\Sigma X - \Sigma Y\Sigma XY}{n\Sigma Y^2 - (\Sigma Y)^2} \end{aligned}\right\} \tag{10}$$

and hence the equation of the line of regression of X on Y may be found. This equation, for assigned values of Y, gives the most probable values of X. This line also passes through the point (M_X, M_Y) and possesses the property that the sum of the X-residuals from it is zero.

If the square root of the mean of the squares of the X-residuals be taken as a measure of the closeness of the concentration of the points about this line of regression of X on Y, we find:

$$S_x = \sigma_X\sqrt{1 - r^2} \tag{11}$$

where, as before,

$$r = \frac{\Sigma xy}{n\sigma_X\sigma_Y}$$

The value of m_2 may now be written

$$m_2 = r\frac{\sigma_X}{\sigma_Y} \tag{12}$$

and the equation of the line of regression of X on Y may be written:

$$x = r\frac{\sigma_X}{\sigma_Y} \cdot y$$

or

$$X - M_X = r\frac{\sigma_X}{\sigma_Y}(Y - M_Y) \tag{13}$$

We can therefore obtain two straight lines which fit the given n points according to the principle of least squares. We can minimize the sum of the squares of the Y-residuals of the line $Y = m_1X + b_1$ and obtain the regression line of Y on X given by equation (6). This line is to be used to find the most probable Y for a given X. We can minimize the sum of the squares of the X-residuals of the line $X = m_2Y + b_2$ and obtain the regression line of X on Y given by equation (13). It is to be used to find the most probable X for a given Y.

Question: For what values of r will the lines (6) and (13) coincide?

The quantities m_1 and m_2 are called *coefficients of regression.* It may be noted that:

$$r^2 = m_1 m_2 \tag{14}$$

If the deviations of the X and Y variates from their respective means be expressed in units of their standard deviations, that is, if

$$t = \frac{x}{\sigma_X} = \frac{X - M_X}{\sigma_X} \quad \text{and} \quad s = \frac{y}{\sigma_Y} = \frac{Y - M_Y}{\sigma_Y},$$

the equation (6) becomes:

$$s = rt \tag{15}$$

That is, r is the slope of the line of regression of Y on X when the variates x and y are expressed in standard units.

EXERCISES

1. Using the data in Table 12 (p. 47), find the correlation between the quantity of beef available for consumption and the price per hundredweight. Let X equal the quantity available and Y equal the price.

2. In the following table the cows considered were of the same breed under the same management. Find r.

VALUE OF FOOD CONSUMED BY 26 COWS AND VALUE OF PRODUCTS PER COW [1]

Value of Feed Consumed X	*Value of Product per Cow* Y	*Value of Feed Consumed* X	*Value of Product per Cow* Y
$99.83	$246.10	$98.93	$174.64
86.42	207.76	82.69	143.61
91.05	216.52	82.94	143.18
94.05	220.01	87.03	150.02
94.06	214.87	89.07	153.51
86.06	183.53	83.52	143.61
84.20	176.39	83.10	140.46
86.70	178.56	89.16	150.68
86.75	178.11	83.01	136.60
86.57	166.70	89.32	145.41
88.52	169.20	82.22	131.35
94.01	179.25	99.74	157.28
86.23	157.20	84.77	122.22

[1] The data are from Horace Secrist, *Readings and Problems in Statistical Methods,* 1920, p. 420.

3. In the following table:

X = production in million of bales

Y = price per pound in cents received by producers December 1

PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES,[1] 1907–1929

Year	X	Y	*Year*	X	Y	*Year*	X	Y
1907	11.1	10.4	1920	13.4	13.9	1917	11.3	27.7
1908	13.2	8.7	1921	8.0	16.2	1918	12.0	27.6
1909	10.0	13.9	1922	9.8	23.8	1919	11.4	35.6
1910	11.6	14.1	1923	10.1	31.0			
1911	15.7	8.8	1924	13.6	22.6			
1912	13.7	11.9	1925	16.1	18.2			
1913	14.2	12.2	1926	18.0	10.9			
1914	16.1	6.8	1927	13.0	19.6			
1915	11.2	11.3	1928	14.3	18.0			
1916	11.5	19.6	1929	14.5	16.4			

a. Find r for the ten-year period, 1907 to 1916 inclusive.

b. Find r for the ten-year period, 1920 to 1929 inclusive. The years 1917, 1918, and 1919 were abnormal years, and may be omitted from the computation.

4. Using the relation

$$(x - y)^2 = x^2 - 2xy + y^2$$

show that:

$$r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$$

In order to compute the value of r by this formula, what are the implied restrictions upon the X and Y units?

5. Verify the value of r for the data of Table 56 by using the formula of Exercise 4.

6. Find the value of r for the "Savings Bank, Strikes and Lockouts" data of Exercise 3, page 236. Is the formula of Exercise 4 applicable to these data?

7. Show that

$$r_{aX+b \quad cY+d} = r_{XY}$$

8. Given

$$r_{XY}^2 = 1 - \frac{S_y^2}{\sigma_Y^2} = 1 - \frac{\sigma_p^2}{\sigma_Y^2},$$

[1] The data are from *Yearbook of Agriculture*, 1928, p. 837; *Commerce Yearbook*, 1930, p. 216.

show that

$$\text{(a) } r_{XY}^2 = 1 - \frac{\Sigma y^2 - m\Sigma xy}{\Sigma y^2}$$

$$\text{(b) } r_{XY} = \pm \frac{\Sigma xy}{n\sigma_X\sigma_Y}$$

9. For a given X we estimate Y (call it $Y_{est.}$) by the equation (see Section 63)

$$Y_{est.} = r\frac{\sigma_Y}{\sigma_X}(X - M_X) + M_Y$$

(a) Show that the arithmetic mean of the estimated values of Y is equal to the arithmetic mean of the observed values of Y, or that

$$M_{Y_{est.}} = M_Y$$

(b) Show that

$$\sigma_{Y_{est.}}^2 = r^2\sigma_Y^2$$

and thus that

$$r = \pm \frac{\sigma_{Y_{est.}}}{\sigma_Y}$$

the sign to be that of m.

10.

X	Y	x	y	x^2	y^2	xy
2	6					
4	8					
6	10					
8	12					
10	14					
12	16					

(1) Plot the data.
(2) Complete the table.
(3) Compute σ_X.
(4) Compute σ_Y.
(5) Compute r.
(6) Compute m.
(7) Find the regression line.

11.

(a)

X	Y
1	10
2	8
3	6
4	4
5	2

(b)

X	Y
1	10
2	7
3	4
4	1
5	− 2

Treat the data in the accompanying tables as you did those in Number 10 above.

12.

(a)

X	Y
0	12
3	5
5	3
7	0
− 3	5
− 5	3
− 7	0

(b)

X	Y
0	7
3	4
4	3
5	0
− 3	4
− 4	3
− 5	0

Treat the data in the accompanying tables as you did those in Number 10 above.

13. The equations of the lines of regression for a set of data are $y = 0.72x$ and $x = 0.64y$. What is the value of r for the data?

14. The equations of the lines of regression for a set of data are $y = -0.8x$ and $x = -0.45y$. What is the value of r?

67. COMPUTATION OF r FOR GROUPED DATA

For sufficiently small values of n, say $n < 30$, one of the methods employed in the preceding sections is usually used in computing the coefficient of correlation.

If n is a very large number, we are compelled to construct a double-entry table. To construct such a table (see Table 57), the sheet is ruled horizontally and vertically, thus dividing the sheet into a system of columns and a system of rows, each of which is a frequency distribution. Each of the rectangles in a row or column is called a *cell*. Along the left-hand margin from bottom to top are laid off the class intervals or the class marks of the Y variates, and along the top of the diagram from left to right are laid off the class intervals or the class marks of the X variates. Very much as we plot points on an ordinary X-, Y-coördinate system, each observed individual may now be located on this sheet, the *preliminary* or *tally sheet*, with respect to the X and Y measures. We shall locate each individual on the preliminary sheet with a + sign placed within the appropriate cell. Since we shall finally concentrate all the measures in a given cell at its center, it is not necessary that the points be plotted with more precision than is necessary to locate them in the appropriate cells. When all the individuals are accurately located we have a *scatter diagram*.

TABLE 57

Y \ X	X_1	X_2	X_3		X_p
Y_q					
			cell		
Y_3					
Y_2					
Y_1					

A *correlation table* may now be obtained from the preliminary sheet by writing within each cell the number of + marks which fall within it. This number is called the *cell frequency*. We shall indicate a cell frequency by $f(x, y)$. The table is now used for a *work-sheet* or a *computation sheet*.

The numbers in a column corresponding to an assigned X, say $X = X_1$, form a Y-array of the type X_1, and those in a row corresponding to an assigned Y, say $Y = Y_1$, form an X-array of type Y_1.

The correlation table may be represented geometrically by a surface. At the center of each cell imagine a vertical erected with a height proportional to the cell frequency. If the tops of these verticals be joined, an irregular surface results. If the cells are made smaller and smaller while the frequencies remain finite, the irregular surface will approach a regular surface which is called a *frequency surface* or a *correlation surface*.

Since in this chapter we are dealing with grouped data, it is advisable that we write our formulas for r in the frequency forms. Thus equations (3), (7′), (8), and (9) become:

$$r = \frac{\Sigma xyf(x, y)}{n\sigma_X\sigma_Y} \tag{16}$$

$$r = \frac{n\Sigma XYf(x, y) - \Sigma Xf(x)\Sigma Yf(y)}{\sqrt{n\Sigma X^2f(x) - [\Sigma Xf(x)]^2}\sqrt{n\Sigma Y^2f(y) - [\Sigma Yf(y)]^2}} \tag{17}$$

$$r = \frac{\dfrac{\Sigma x'y'f(x, y)}{n} - b_xb_y}{\sqrt{\dfrac{\Sigma x'^2f(x)}{n} - b_x^2}\sqrt{\dfrac{\Sigma y'^2f(y)}{n} - b_y^2}} \tag{18}$$

$$r = \frac{n\Sigma x'y'f(x, y) - \Sigma x'f(x)\Sigma y'f(y)}{\sqrt{n\Sigma x'^2f(x) - [\Sigma x'f(x)]^2}\sqrt{n\Sigma y'^2f(y) - [\Sigma y'f(y)]^2}} \tag{19}$$

The data of Table 58 will be used as an example to illustrate the construction of the preliminary sheet, the correlation table, and the method employed in computing r, the regression equations, and the standard error of estimate.

We shall let the percentage of native white population be measured along the horizontal or X-axis, and the percentage of illiteracy be

TABLE 58. PERCENTAGE OF NATIVE WHITE AND PERCENTAGE OF ILLITERATE TEN YEARS OF AGE AND OVER IN THE POPULATION OF PENNSYLVANIA, BY COUNTIES, 1920 [1]

County	*Percentage Native White* X	*Percentage Illiterate* Y	*County*	*Percentage Native White* X	*Percentage Illiterate* Y
Adams	98.7	1.6	Lackawanna	77.1	8.6
Allegheny	74.5	4.8	Lancaster	96.3	1.4
Armstrong	86.4	4.3	Lawrence	80.2	6.5
Beaver	75.9	6.2	Lebanon	95.3	2.5
Bedford	96.9	3.2	Lehigh	89.3	2.3
Berks	93.4	2.7	Luzerne	77.3	9.5
Blair	92.2	2.6	Lycoming	94.4	1.5
Bradford	96.2	2.0	McKean	86.2	1.8
Bucks	87.6	2.5	Mercer	80.0	6.2
Butler	89.6	3.0	Mifflin	96.4	2.4
Cambria	79.2	6.3	Monroe	95.0	2.3
Cameron	89.9	2.6	Montgomery	83.4	3.6
Carbon	82.3	8.3	Montour	94.2	4.8
Center	93.5	1.8	Northampton	81.9	5.2
Chester	81.7	4.5	Northumberland	88.9	4.7
Clarion	96.4	1.8	Perry	98.9	1.5
Clearfield	85.9	4.4	Philadelphia	70.7	4.0
Clinton	93.0	2.5	Pike	90.6	1.3
Columbia	93.0	2.8	Potter	93.0	1.9
Crawford	92.3	1.5	Schuylkill	84.0	7.9
Cumberland	96.2	1.4	Snyder	99.8	2.1
Dauphin	87.8	3.3	Somerset	84.3	6.4
Delaware	75.8	4.4	Sullivan	90.4	4.5
Elk	81.6	3.1	Susquehanna	91.0	2.8
Erie	84.8	4.0	Tioga	93.4	2.4
Fayette	76.3	8.2	Union	99.3	1.4
Forest	94.1	2.7	Venango	92.7	3.5
Franklin	97.0	1.8	Warren	86.2	3.3
Fulton	98 9	2.3	Washington	74.0	7.3
Greene	93.5	4.4	Wayne	90.9	2.8
Huntingdon	93.1	4.0	Westmoreland	77.8	7.6
Indiana	82.4	5.9	Wyoming	96.0	1.6
Jefferson	88.0	3.5	York	97.2	1.6
Juniata	99.1	1.2			

measured along the vertical or Y-axis. The class widths, w_x and w_y, may be selected in accordance with the principles suggested in

[1] The data are from *Fourteenth Census of the United States*, Vol. III, pp. 859–65.

PRELIMINARY SHEET
Percentage Native White

Y \ X	69.95–72.95	75.95–78.95	81.95–84.95	87.95–90.95	93.95–96.95	99.95
9.95		+				
8.95		++	+			
7.95	+	+	+			
6.95	+		+++	+		
5.95			+	+		
4.95	+	++	+	+	++	++	++	+	
3.95			+	+	++	++	+	+	
2.95					+	+++	+++ +++	+++ ++	++
1.95					+	+	+++	+++ ++	+++ +++
0.95									

Percentage Illiterate

Section 13 (p. 30). Since the X variates range from 70.7 to 99.8, we shall choose $w_x = 3$, and since the Y variates range from 1.2 to 9.5, we shall choose $w_y = 1$. Also, since the given measures are accurate to tenths, we shall express our class boundaries to hundredths.[1] Plotting the points, we have the preliminary sheet.

The preliminary sheet is now complete. We are now ready to transcribe the results of the tally to the computation sheet. We then have Table 59.

Having formed the correlation table, which is the part of the table bounded by the double lines, we arrange the computation to

[1] If the student prefers he may use some other method for fixing the class limits. Any method recommended in Section 12 (p. 23) will be satisfactory. Thus the X-class intervals may be 70.0–72.9, 73.0–75.9, etc., and the Y-class intervals may be 1.0–1.9, 2.0–2.9, etc. The class marks will be changed accordingly. The X-class marks will become 71.45, 74.45, etc., and the Y-class marks will become 1.45, 2.45, etc.

simplify as far as possible the somewhat complicated details. We first add the frequencies of the rows and columns and obtain the row marked $f(x)$ and the column marked $f(y)$. Choosing an arbitrary origin (h, k) near the center of the table — in Table 59 $(h, k) = (83.45, 4.45)$ — and the class intervals as units of measurement, we obtain the row marked x' and the column marked y'. That is, we use the familiar transformations $X = h + w_x x'$ and $Y = k + w_y y'$. The next two rows, $x'f(x)$ and $x'^2f(x)$, and the next two columns, $y'f(y)$ and $y'^2f(y)$, are self-explanatory and are used in computing the means, M_X and M_Y, and the standard deviations, σ_X and σ_Y.

The column headed $x'y'f(x, y)$ needs some explanation. Recalling formula (18) for computing r, we note that we must find $\Sigma x'y'f(x, y)$. That is, we must find the $x'y'$ for each individual measured, then find their sum. Since the frequency of any cell is concentrated at the center of the cell, we shall compute the $x'y'$ for the frequency of each cell by multiplying the $x'y'$ of each cell by the cell frequency, and adding the $x'y'$ for all the cells of a given row. In this manner we obtain the numbers in the column headed $x'y'f(x, y)$. By adding the $x'y'$ of all the rows, we obtain the sum of the $x'y'$ of the entire table.[1] Thus:

for row $Y = 8.45$, the total $x'y'$ is $(-2)(4)2 + (0)(4)1 = -16$
The total $x'y'$ for each of the other rows is found in a similar manner.

Consequently, for the entire distribution we have:

$$n = 67, \quad h = 83.45, \quad k = 4.45, \quad w_x = 3, \quad w_y = 1$$
$$\Sigma x'f(x) = 116 \qquad \Sigma y'f(y) = -52 \qquad \Sigma x'^2f(x) = 612$$
$$\Sigma y'^2f(y) = 342 \qquad \Sigma x'y'f(x,y) = -362$$

Therefore:

$$b_x = \frac{116}{67}, \quad M_X = 83.45 + 3\left(\frac{116}{67}\right) = 88.64\%$$

$$b_y = \frac{-52}{67}, \quad M_Y = 4.45 - \frac{52}{67} = 3.67\%$$

$$\sqrt{\frac{\Sigma x'^2f(x)}{n} - b_x^2} = \sqrt{\frac{612}{67} - \left(\frac{116}{67}\right)^2} = 2.48$$

$$\sigma_X = 3(2.48) = 7.44\%$$

[1] A row $x'y'f(x, y)$ is similarly found. It is useful for checking the column $x'y'f(x, y)$.

TABLE 59. COMPUTATION SHEET FOR FINDING r FOR THE DATA OF TABLE 58

Percentage Native White

Percentage Illiterate (Y \ X)	71.45	74.45	77.45	80.45	83.45	86.45	89.45	92.45	95.45	98.45	$f(y)$	y'	$y'f(y)$	$y'^2f(y)$	$x'y'f(x, y)$
9.45			1								1	5	5	25	− 10
8.45			2		1						3	4	12	48	− 16
7.45		1	1		1						3	3	9	27	− 15
6.45		1		3	1						5	2	10	20	− 12
5.45				1	1						2	1	2	2	− 1
4.45	1	2		1	1	2	2	2	1		12	0	0	0	0
3.45				1	1	2	2	1	1		8	− 1	− 8	8	− 12
2.45						1	3	6	5	2	17	− 2	− 34	68	− 110
1.45						1	1	3	5	6	16	− 3	− 48	144	− 186
$f(x)$	1	4	4	6	6	6	8	12	12	8	67		− 52	342	− 362
x'	− 4	− 3	− 2	− 1	0	1	2	3	4	5					
$x'f(x)$	− 4	− 12	− 8	− 6	0	6	16	36	48	40	116				
$x'^2f(x)$	16	36	16	6	0	6	32	108	192	200	612				
$x'y'f(x, y)$	0	− 15	− 32	− 6	0	− 7	− 22	− 66	− 104	− 110	− 362				

Check

$$\sqrt{\frac{\Sigma y'^2 f(y)}{n} - b_y^2} = \sqrt{\frac{342}{67} - \left(\frac{-52}{67}\right)^2} = 2.12\%$$

$$\sigma_Y = 2.12\%$$

Using equation (18) we have:

$$r = \frac{\frac{-362}{67} - \left(\frac{116}{67}\right)\left(\frac{-52}{67}\right)}{(2.48)(2.12)} = -0.77$$

The equations of the lines of regression can now be found:

$$m_1 = r\frac{\sigma_Y}{\sigma_X} = \frac{-0.77(2.12)}{3(2.48)} = -0.22$$

Using

$$Y - M_Y = m_1(X - M_X)$$

we obtain the equation for the regression of Y on X with its S_y. It is:

$$Y - 3.67 = -0.22(X - 88.64) \quad \text{or} \quad Y = -0.22X + 23.17$$

$$S_y = 2.12\sqrt{1 - (.77)^2} = 1.35\%$$

For a given value of X, this equation gives the best (that is, the most probable) value for Y. This most probable value of Y is the mean of the Y-array corresponding to a given value of X. Hence, the equation above gives the expected mean [1] of the Y-array for a given X. We use S_y to measure its reliability.

For example, if $X = 86.45$, we obtain $Y = 4.15$ for the estimated or expected mean of the Y-array. We may compare this with the observed mean for $X = 86.45$ by computing the mean of the distribution in the usual manner. We find the observed mean for the Y-array corresponding to $X = 86.45$ to be 3.28.

When $X = 86.45$ we found the estimated Y, $Y_{est.}$, to be 4.15. Combining this value with its measure of reliability $S_y = 1.35$ we have this fact: the odds are 2 to 1 that the observed Y for $X = 86.45$ does not differ numerically from $Y_{est.} = 4.15$ by more than 1.35.

[1] It may be shown that the line of regression of Y on X is the line which best fits the points which designate the means of the Y-arrays or columns, and that the line of regression of X on Y best fits the points which designate the means of the X-arrays or rows.

In other words, the odds are 2 to 1 that for $X = 86.45\%$ the observed Y will lie in the interval $4.15 \pm 1.35\%$.

$$m_2 = r\frac{\sigma_X}{\sigma_Y} = \frac{-0.77(7.44)}{2.12} = -0.27$$

Using

$$X - M_X = m_2(Y - M_Y)$$

we obtain the equation for the regression of X on Y with its S_x. It is:

$$X - 88.64 = -0.27(Y - 3.67)$$

or

$$X = -0.27Y + 89.63$$

$$S_x = 7.44\sqrt{1 - (.77)^2} = 4.75\%$$

For a given value of Y, this equation gives the most probable value for X. That is, for a given Y, this equation gives the expected mean of the corresponding X-array.

For example, if $Y = 3.45$, we obtain $X = 88.70$ for the estimated mean of the array. The observed mean of the X-array corresponding to $Y = 3.45$ is $X = 87.95$. We use S_x to measure the reliability of the estimate. Thus the odds are 2 to 1 that for $Y = 3.45\%$ the observed X will lie within the interval $88.70 \pm 4.75\%$.

This completes the theory of simple linear correlation. A word about the reliability of r may be in order. If n is fairly large and if the surface described on page 254 is closely normal, the reliability of r may be tested by either of the formulas:

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}$$

$$E_r = 0.6745\sigma_r = 0.6745\frac{1 - r^2}{\sqrt{n}}$$

with the interpretation of σ_r and E_r similar to that employed in Section 37. Since the assumptions underlying these formulas are rather severe, they are to be used with care.

EXERCISES

1. The data for the table on page 261 are taken from the *Yearbooks of Agriculture:* 1920, pp. 753 and 537; 1935, pp. 568 and 379.

X = price of corn per bushel (cents)
Y = value of hogs per head (dollars)

Construct a correlation table with the X-classes: 20 a.u. 35, 35 a.u. 50, etc., and the Y-classes 3.00 a.u. 6.00, 6.00 a.u. 9.00, etc.

Find r, the equation of the regression line of Y on X, and S_y.

Estimate Y when $X = 75$ and give the odds that measure the reliability of the estimate.

Year	*Corn Cents per bu.* X	*Hogs Dollars per head* Y	*Year*	*Corn Cents per bu.* X	*Hogs Dollars per head* Y
1870	49	$5.80	1905	41	$5.99
1871	43	5.61	1906	40	6.18
1872	35	4.01	1907	52	7.62
1873	44	3.67	1908	61	6.05
1874	58	3.98	1909	58	6.55
1875	37	4.80	1910	48	9.17
1876	34	6.00	1911	62	9.37
1877	35	5.66	1912	49	8.00
1878	32	4.85	1913	69	9.86
1879	38	3.18	1914	64	10.40
1880	40	4.28	1915	58	9.87
1881	64	4.70	1916	89	8.40
1882	49	5.97	1917	128	11.75
1883	42	6.75	1918	137	19.54
1884	36	5.57	1919	135	22.02
1885	33	5.02	1920	68	19.08
1886	37	4.26	1921	53	12.99
1887	44	4.48	1922	75	10.06
1888	34	4.98	1923	84	11.58
1889	28	5.79	1924	105	9.72
1890	51	4.72	1925	70	12.38
1891	41	4.15	1926	75	15.21
1892	39	4.60	1927	85	15.97
1893	37	6.41	1928	84	12.03
1894	46	5.98	1929	80	12.24
1895	25	4.97	1830	59	12.73
1896	22	4.35	1931	32	10.75
1897	26	4.10	1932	32	5.80
1898	29	4.39	1933	52	3.99
1899	30	4.40	1934	85	3.92
1900	36	5.00			
1901	61	6.20			
1902	40	7.03			
1903	43	7.78			
1904	44	6.15			

2. The accompanying table shows the scores on placement examinations of 326 freshmen at Bucknell University. Find r and the equations of the lines of regression.

EXAMINATION SCORES IN MATHEMATICS AND ENGLISH

Mathematics

English \ X	2.5	7.5	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5	72.5
252.5							1	1		1	1				
237.5							1		3				1		
222.5			1	1	1		1	1		3	2	2			1
207.5			1		2	3	3	5	2	1	1	2			
192.5				3	4	7	4	6	2	2	1	1			
177.5			2	5	3	2	1	6	4	1	1	1	1		
162.5		1	1	6	6	8	18	7	6	5	3	3			
147.5	1			3	9	3	5	5	3	2	1	2			
132.5		2	4	3	3	8	8	3	2	1	1				
117.5			4	5	6	5	9	4	1	3	1				
102.5		1	2	4	3	4	3	3		1	1				
87.5	1	1	1	3	3	2		1	1						
72.5		2	3	3	5		2	1							
57.5		1	1		1	1									
42.5					1										

3. In the following table

X = the number of minutes required to solve a group of arithmetical exercises by each of forty employees

Y = the executive ratings, in per cent, of the same employees

X	Y	X	Y	X	Y	X	Y
12.4	90	17.2	77	24.0	67	18.3	72
14.0	85	8.8	94	12.4	91	8.5	96
15.5	83	11.6	90	20.2	74	10.4	92
25.0	70	20.6	68	16 2	82	17.6	80
15.8	80	9.8	91	12.2	88	22.4	72
23.4	74	11.2	89	16.0	82	15.3	78
22.3	78	8.7	96	13.3	87	21.5	70
13.5	88	25.8	65	9.2	92	13.2	87
17.8	82	12.6	88	12.4	87	17.6	75
14.4	92	16.5	74	26.3	60	9.5	94

Construct a double entry table with the X-classes designated as 8.0 a.u. 12.0, 12.0 a.u. 16.0, etc., the Y-classes designated as 60 a.u. 65, 65 a.u. 70, etc.

Find r, the regression line of Y on X, and S_y.

What is the estimated value of Y for $X = 20$, and what is the reliability of the estimate?

68. CORRELATION BY RANKS

When two series of values are expressed according to their *ranks* and not in terms of their *actual values* or *scores*, we can easily find the approximate correlation between them. Such correlation is used to find the relation between the paired scores when their number is small or when the data do not warrant an application of the cross product method to the actual values. Also, the method is useful in finding the correlation between series that may be arranged according to size and yet may not be subjected to exact measurement.

In such correlation as we are here describing we must keep in mind that the (X, Y) values are the *rank* or *position* numbers of some characteristics. We shall arrange the values in *ascending* order. To the smallest value we assign 1, to the next in order 2, etc. We may then find the rank correlation by employing any of our formulas for r_{XY} with the data arranged according to ranks. However, a formula may be easily derived for this special case by a method which we shall indicate at the end of this section. When ranks are used we indicate the coefficient by ρ_{XY} or by r_{XY} (rank).

To illustrate the problem we are presenting, let us consider the heights and weights of the five boys A, B, C, D, E.

TABLE 60. HEIGHTS AND WEIGHTS OF FIVE BOYS

Boy	*Height (inches)*	*Weight (pounds)*	*Rank in Height X*	*Rank in Weight Y*
A	60	137	1	2
B	62	132	2	1
C	63	148	3	3
D	65	157	4	5
E	68	153	5	4

For the height-weight data given in columns 2 and 3, $r_{\text{height weight}} = 0.77$.

Let us find the cross product coefficient for the rank data given in columns 4 and 5 of Table 60.

X	Y	x	y	x^2	y^2	xy	X^2	Y^2	XY
1	2	−2	−1	4	1	2	1	4	2
2	1	−1	−2	1	4	2	4	1	2
3	3	0	0	0	0	0	9	9	9
4	5	1	2	1	4	2	16	25	20
5	4	2	1	4	1	2	25	16	20
15	15	0	0	10	10	8	55	55	53

$$M_X = M_Y = \tfrac{15}{5} = 3 \qquad \sigma_X = \sigma_Y = \sqrt{\tfrac{10}{5}} = \sqrt{2}$$

$$\rho_{XY} = r_{XY}\ (\text{rank}) = \frac{\Sigma xy}{n\sigma_X\sigma_Y} = \frac{8}{5\sqrt{2}\,\sqrt{2}} = \frac{8}{10} = 0.80$$

We may also find $\rho_{XY} = r_{XY}$ (rank) by using formula (7) page 245, $r = \dfrac{\Sigma XY - nM_XM_Y}{n\sigma_X\sigma_Y}$. We have

$$\sigma_X = \sqrt{\frac{\Sigma X^2}{n} - (M_X)^2} = \sqrt{\frac{55}{5} - 9} = \sqrt{2}$$

$$\sigma_Y = \sqrt{\frac{\Sigma Y^2}{n} - (M_Y)^2} = \sqrt{\frac{55}{5} - 9} = \sqrt{2}$$

Hence

$$\rho_{XY} = r_{XY}\ (\text{rank}) = \frac{53 - 5(3)(3)}{5\sqrt{2}\,\sqrt{2}} = 0.80$$

Thus we see that the so-called rank difference method is merely the cross-product correlation between the rank numbers of the variates. As might be suspected, frequently certain complications arise to interrupt the apparently simple ranking of the values. Generally there are several scores of the same size, or there exist *ties* in the ranks. In such cases it is customary to give each the mean of the ranks of the positions that they occupy. Thus, suppose 3 tied for fifth place. Had there been no ties, the ranks would have been 5, 6, 7. We arbitrarily assign to each place the rank number 6, which is the mean of 5, 6, and 7. If 2 scores tied for the eighth place, we would assign each the rank number 8.5.

We shall now proceed to develop a formula for finding the rank coefficient ρ_{XY}.

Evidently the X-values are the numbers 1, 2, 3, . . ., n, and the Y-values are the same numbers but probably arranged in a different order. Hence

$$\Sigma X = \Sigma Y = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

$$M_X = M_Y = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{(n+1)}{2}$$

Also

$$\Sigma X^2 = \Sigma Y^2 = 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

Hence, using formula (7) page 128,

$$\sigma_X = \sigma_Y = \sqrt{\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2} = \sqrt{\frac{n^2-1}{12}}$$

From

$$\Sigma(X - Y)^2 = \Sigma X^2 - 2\Sigma XY + \Sigma Y^2$$

we obtain, substituting values for ΣX^2 and ΣY^2 above,

$$\Sigma XY = \frac{n(n+1)(2n+1)}{6} - \frac{\Sigma(X - Y)^2}{2}$$

Now, substituting in (7′) page 245, the values found, we obtain after simplifying

$$\rho_{XY} = r_{XY}\,(\text{rank}) = 1 - \frac{6\Sigma(X - Y)^2}{n(n^2 - 1)} \qquad (20)$$

Thus, after our data are ranked the computation of ρ_{XY} is decidedly simple. To illustrate the use of formula (20) let us return to the height-weight data. We have the following table with headings suitable to the use of formula (20).

RANK IN HEIGHT AND WEIGHT OF FIVE BOYS

Boy	*Rank in Height* X	*Rank in Weight* Y	$X - Y$	$(X - Y)^2$
A	1	2	− 1	1
B	2	1	1	1
C	3	3	0	0
D	4	5	− 1	1
E	5	4	1	1

$$n = 5 \qquad \Sigma(X - Y)^2 = 4$$

$$\rho_{XY} = 1 - \frac{6(4)}{5(25 - 1)} = 1 - \frac{24}{5(24)} = \frac{4}{5} = 0.80$$

EXERCISES

1. Ten examination papers in algebra were read by two judges and ranked according to merit. The following table shows the results of the rankings. Find ρ_{XY}.

Examination Paper	*Rank by Judge No. 1* X	*Rank by Judge No. 2* Y
1	6	5
2	2	1
3	4	6
4	8	9
5	1	2
6	3	3
7	7	7
8	5	4
9	10	8
10	9	10

2. The following table gives the ranks of 10 salesmen by the sales manager of a corporation and also the ranks of the 10 salesmen on a psychological test. Find ρ_{XY}.

Salesman	*Rank by Sales Manager*	*Rank on Test*
Jones	1	1
Smith	2	3
Brown	3	2
Kelly	4	6
Sanders	5	7
Benson	6	4
Owens	7	8
Miller	8	5
Borden	9	9
Peterson	10	10

3. From the following table, by the method of ranks find the correlation between the grades in Test I and Test II; between the grades in Test I and Test III; between the grades in Test II and Test III.

GRADES OF 21 STUDENTS IN THREE TESTS IN INTEGRAL CALCULUS

Student	*Test I*	*Test II*	*Test III*	*Student*	*Test I*	*Test II*	*Test III*
1	80	45	55	12	99	87	99
2	60	50	80	13	82	70	72
3	94	81	95	14	98	83	92
4	93	85	90	15	97	95	93
5	87	80	70	16	34	55	30
6	95	90	100	17	96	96	96
7	74	60	60	18	74	20	40
8	61	79	85	19	62	72	75
9	92	82	71	20	63	94	91
10	67	84	97	21	88	78	94
11	100	86	98				

69. CORRELATION AND CAUSATION

The correlation coefficient, as we have used the term, is a mathematical expression which measures the mathematical relationship — *based upon linear regression*, or the best-fitting straight line to the data — that exists between two variables X and Y. It must not be supposed that a low coefficient of correlation *proves* a lack of relationship between the two variables. Consider the data of the Table 61. We note that for these data:

$$M_X = 0$$

and

$$\Sigma XY = 0$$

Hence by equation (7):

$$r = 0$$

TABLE 61

X	Y	XY
0	5	00
3	4	12
4	3	12
5	0	00
− 3	4	− 12
− 4	3	− 12
− 5	0	00
0	19	00

That is, based upon the best-fitting straight line the data show a very poor relationship or a straight line of very poor fit.

But based upon the semicircle, $Y = +\sqrt{25 - X^2}$, we have *perfect correlation*, since each point is on the curve. This simple illustration emphasizes a fact that we should keep in mind, namely, that the *Bravais-Pearson cross-product formula is based upon straight-line regression.*

It should also not be supposed that the existence of high coefficient of correlation between two variables *proves* any necessary and inherent causal relationship between the two — that is, that one is the absolute cause of the other. Consider the following table:

TABLE 62 [1]

Year	X	Y	X^2	Y^2	XY
1870	38	30	1,444	900	1,140
1875	55	38	3,025	1,444	2,090
1880	56	51	3,196	2,601	2,856
1885	73	69	5,329	4,761	5,037
1890	92	97	8,464	9,409	8,924
1895	114	114	12,996	12,996	12,996
1900	138	135	19,044	18,225	18,630
1905	177	169	31,329	28,561	29,913
1910	254	205	64,516	42,025	52,070
Total	997	908	149,283	120,922	133,656

Applying formula (7) we obtain

$$r = 0.98$$

which is so astoundingly large that we are tempted to believe that we have a direct and dependent cause-and-effect relationship. As a matter of fact

X = the total salaries paid school superintendents and teachers in millions of dollars

and

Y = the total consumption of wines and liquors in the United States in ten million gallons

for the given years.

This illustration shows almost perfect correlation, yet no one believes that the consumption of wines and liquors increased necessarily because teachers' salaries were increasing, nor that teachers' salaries were increasing necessarily because more wines and liquors were being consumed.

A high coefficient of correlation proves a close *linear* mathematical relationship between the two variables. It *proves* nothing more. It

[1] The data are from *Statistical Abstract of the United States*, 1918, pp. 830 and 835.

suggests the probability of a cause-and-effect relationship between the two variables, but the investigator must search further for the explanation. Measurement of correlation is one part of the problem; interpretation of the results is a more difficult part of the problem.[1]

Before the subject of statistical analysis had reached its present development, John Stuart Mill stated in his *Logic:*

> Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or *is connected with it through some fact of causation.*[2]

The suggestion in the last clause of Mill's statement may assist us in explaining the above paradoxical relationship between teachers' salaries and the consumption of wines and liquors. The period from 1870 to 1910 was one of rapid development in the United States. Population increased rapidly; foreign and domestic commerce, agriculture, and the manufacturing industries grew by leaps and bounds. The total amounts paid for the salaries of school superintendents and school-teachers and the total amount of wines and liquors consumed merely kept step with the development in other lines. As a matter of fact, we are not at all astonished that the two do show a surprisingly large coefficient. We term such correlation "spurious."

In the interpretation of the coefficient of correlation it is better not to consider it as a *measure of causal dependence* but rather to consider it as a mathematical expression for the *degree of association* between the factors. In this regard Professor Chaddock says,

> Therefore, we no longer search for cause and effect relations as fixed and unvarying laws. Association or correlation between occurrences tends to replace the older idea of causation in scientific investigation. We have seen that variation is a universal characteristic of phenomena. We can secure relative likeness in phenomena by a process of classification which places similar things together and disregards minor variations. The problem of science is to find out how the variation in one group of facts is associated with or contingent upon the variation in other groups, and to measure the degree of the association.
>
> The aim is to find the series of facts which are most closely correlated in order to enable the investigator to predict future experience. *Causation*

[1] See Rietz and others, *op. cit.*, p. 138.

[2] Book III, Chap. VIII, Sect. 6. (Italics my own.)

becomes a descriptive concept reached by statistical processes applied to the facts of experience.[1]

As a final word we wish to reëmphasize that the preceding chapter, Linear Trends, and this chapter, Simple Correlation, have been concerned with the problem of expressing the relationship between the sets of data by means of *linear* regression. We have assumed Y to be a linear function of a *single* independent variable X — or X to be a linear function of Y. The close restrictions imposed necessarily limit the range of application of the method. As an illustration, in considering the problem of July rainfall in Ohio and its effect upon the yield of corn in that state — Exercise 3, page 243 — the thoughtful student must have wondered about the effect of other natural causes, such as the rainfall for May,the rainfall for June, the temperatures for May, June, July, and August. And well he may wonder. The yield of corn may be considered as a function (or effect) of the several variables (or causes) mentioned. A study of problems of this character in which the dependent variable is a *linear function of several independent variables* belongs to the subject of *multiple correlation*, whereas problems in which the dependent variable is a *linear function of a single independent variable* belong to the subject of *simple correlation*. The subject of multiple correlation is treated in Chapter 9. If the reader desires he may, without loss of continuity, begin its study now; or he may defer it.

Further, we may consider that the relationship between the dependent variable and the single independent variable can be described by some simple curve other than a straight line. Such correlation based upon curvilinear regression will be considered in Chapter 10.

EXERCISES

1. For the Water Depth–Alfalfa Yield data of Exercise 2, page 243, the following is a summary:

$M_X = 33.75$ inches	$M_Y = 7.25$ tons	$m = 0.075$
$\sigma_X = 14.98$ inches	$\sigma_Y = 1.26$ tons	$r = 0.89$

(1) Find the equation of the regression line of Y on X.
(2) Is the value of r sufficiently large to warrant confidence in the regression line for purposes of estimation?
(3) Find Y in (1) if $X = 40$.
(4) Find S_y and interpret your result for the value found in (3).

[1] R. E. Chaddock, *Principles and Methods in Statistics*, 1925, p. 250.

2. Find the correlation of the yield of a plant of oats with the number of kernels per plant for the data of the accompanying table.

X = the number of kernels per plant. Y = the yield in grams.

KERNELS PER PLANT [1]

Yield in grams (Y) \ X	25	75	125	175	225	275	325	375	425	475
8.5										1
7.5								1		1
6.5								3	4	
5.5						12	26	4		
4.5				4	30	39	7			
3.5			4	47	51	7				
2.5			61	45						
1.5		20	30							
0.5	2	1								

3. The following table is a correlation table for the lengths and the breadths of 60 leaves. X = breadths and Y = lengths, in millimeters.[2]

BREADTHS

Lengths (Y) \ X	16	19	22	25	28	31	34
52						1	1
47				2	3	1	1
42			1	3	5	3	
37			2	5	4	3	
32		1	3	4	3	2	
27		1	3	3	1		
22	1	2	1				

Find r and the regression lines for the data.

4. Find r for the Savings Bank Deposits–Strikes and Lockouts data of page 236. Is this value of r sufficiently large to warrant your using with confidence the regression equations for purposes of estimation?

5. As in Exercise 4 above, treat the Value of Crops–Value of Land (Illinois) data of page 236.

6. Similarly, treat the Value of Crops–Value of Land (Iowa) data of page 237.

[1] The data are from A. S. Gale and C. W. Watkeys, *Elementary Functions and Applications*, 1920, p. 432.

[2] Gavett, *First Course in Statistical Method*, p. 234.

7. The following correlation table gives the scores of 104 freshmen at Georgetown College. X = scores in mathematics. Y = scores in intelligence.

SCORES IN INTELLIGENCE AND MATHEMATICS TESTS OF 104 STUDENTS

Intelligence Y \ Mathematics X	2.5	7.5	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5
145							1			1	1	1
135					1		2	2		1	1	1
125				1	5		2	3	4	1	1	1
115		1	2	1	5	1	3	2	2	1		1
105		3	1	4	1	4	2	4				
95		4	2	2	4	2	3		1			
85	1		3	3	2	1						
75		1	2	1	1	1						
65				2								

Find r and the regression lines for the data.

8. In an investigation of the resemblance of fathers and sons with respect to stature, the following summary was obtained:

Stature of fathers	*Stature of sons*
X	Y
$M_X = 67.7$ inches	$M_Y = 68.7$ inches
$\sigma_X = 3.21$ inches	$\sigma_Y = 2.71$ inches

$$r = 0.51$$

What is the most probable height of the sons of a group of selected fathers whose mean height is 6 feet? Discuss the reliability of this estimate by means of S_y.

9. Are the following correlations positive or negative?

(1) The speed of an auto and the distance required to bring the car to rest when the brakes are applied.
(2) Age of applicants for life insurance and cost of insurance.
(3) Age of an automobile and its trade-in value.
(4) Family income and cost of the family car.
(5) Marriage rate and index of unemployment.
(6) Age and blood pressure.
(7) Age of husbands and age of wives.
(8) Index of unemployment and amount of goods purchased.
(9) The soot content in the air at Pittsburgh and the production of pig iron.

(10) Total production of wheat and the average farm price per bushel.
(11) Per cent illiteracy and the per cent foreign population in the counties in Pennsylvania.
(12) Crime, as measured by the number of indictable offences tried, and the index of unemployment.
(13) Value of crops per acre and the value of land per acre in Illinois.
(14) Amount of savings deposits and the number of strikes and lock-outs in the United States.
(15) Number of hogs slaughtered per month at Chicago and monthly price of pork at Chicago.
(16) Marriage rate and the index of industrial activity. (See Groves and Ogburn: *American Marriage and Family Relationships*, Chapter XVIII.)
(17) Scholarship and success in life. (See Gifford: "Does Business Want Scholars?" *Harpers*, May, 1928.)

10. In the following table (F. C. Mills: *Statistical Methods*, p. 381)

X = Federal Reserve Banks' Discount Rates (per cent).
Y = Commercial Banks' Discount Rates (per cent).

FEDERAL RESERVE BANKS' DISCOUNT RATES (PER CENT)

Commercial Banks' Discount Rates (per cent)

Y \ X	4.00	4.50	5.00	5.50	6.00	6.50	7.00
8.00					1	1	2
7.50					7	9	1
7.00			5	4	63	9	36
6.50		2	9	10	22	1	3
6.00	1	90	29	6	30		
5.50	11	110	5				
5.00	10	24					
4.50	2	1					

(1) Choose (h, k) at (5.50, 6.50), compute r and the regression of Y on X.
(2) Find the estimated value of Y if $X = 5.00$.
(3) Compute the arithmetic mean of the Y-array for $X = 5.00$ and compare with the value found in (2).
(4) Find the regression equation of X on Y.
(5) Find the estimated value of X if $Y = 7.00$.
(6) Find the arithmetic mean of the X-array for $Y = 7.00$ and compare with the value found in (4).

(7) Find S_y and S_x for the estimated values in (2) and (5) and interpret them.

11. The data in the following table were taken from the *Handbook of Labor Statistics*, 1936 Edition, pages 132 and 673.

X = Index of Wholesale Prices in the United States. (U.S. Dept. of Labor. Monthly Average, 1926 = 100.)

Y = General Index of Employment. (U.S. Dept. of Labor. 3-year average, 1923–1925 = 100.)

Compute r.

Year	X	Y	*Year*	X	Y	*Year*	X	Y
1919	139	107	1925	104	99	1931	73	77
1920	154	108	1926	100	101	1932	65	64
1921	98	82	1927	95	99	1933	66	69
1922	97	91	1928	97	99	1934	75	79
1923	101	104	1929	95	105	1935	80	82
1924	98	97	1930	86	92			

12. The following data are taken from the *Yearbook of Agriculture*, 1935, pp. 363–364.

X = supply of wheat in the U.S., July 1.

Y = price of wheat at Chicago.

Year	*Supply (million bu.)* X	*Price (cents)* Y	*Year*	*Supply (million bu.)* X	*Price (cents)* Y
1919	77	227	1927	122	138
1920	145	216	1928	124	117
1921	126	128	1929	247	130
1922	114	113	1930	303	84
1923	137	106	1931	326	53
1924	144	139	1932	385	53
1925	115	161	1933	393	94
1926	105	140			

(1) Draw a chart for these data similar to Chart 6, p. 48.
(2) Compute r and interpret it.
(3) Compute m and interpret it.
(4) Write the equation of the regression line of Y on X.
(5) Find the estimated values of Y if X = 100, 200, and 300.
(6) Find S_y of the estimates, and interpret.

13. The following table gives the average number of kernels per culm per oat plant and the average height of the oat plants (Love-Leighty). Find r.

NUMBER OF KERNELS

Height Y \ X	35	45	55	65	75	85	95	105	115	125	$f(y)$
87.5								3	2	2	7
82.5					1	12	26	23	9	2	73
77.5				2	16	40	38	23	3		122
72.5			1	13	30	59	32	5			140
67.5			7	22	9	6	1				45
62.5		4	7								11
57.5	1		1								2
$f(x)$	1	4	16	37	56	117	97	54	14	4	400

14. In the following table:

X = price per bushel in cents received by producers December 1 for corn

Y = price per bushel in cents received by producers December 1 for wheat

Find r and discuss its significance. Would you say this correlation is spurious?

PRICE OF CORN AND PRICE OF WHEAT IN THE UNITED STATES,[1] 1909–1928

Year	X	Y	*Year*	X	Y
1909	58.6	98.4	1919	134.5	214.9
1910	48.0	88.3	1920	67.0	143.7
1911	61.8	87.4	1921	42.3	92.6
1912	48.7	76.0	1922	65.8	100.7
1913	69.1	79.9	1923	72.6	92.3
1914	64.4	98.6	1924	98.2	129.9
1915	57.5	91.9	1925	67.4	141.6
1916	88.9	160.3	1926	64.2	119.8
1917	127.9	200.8	1927	72.3	111.5
1918	136.5	204.2	1928	75.1	97.2

15. If X = Income in dollars per capita in Texas in 1932,
Y = Retail sales in dollars per capita in Texas in 1932,
$r = 0.875$, and $m = 0.746$,

(1) Comment on the estimative value of the line of regression $Y = 0.746X + 8.33$.

[1] The data are from *Yearbook of Agriculture*, 1928, pp. 670 and 702.

(2) If $X = \$175$, compute Y, and compare with the observed value, \$133.
(3) If X increases \$1.00, what is the expected change in Y?

16. In the following table:

X = scores of 32 students on the Bucknell test in intermediate algebra.
Y = scores of the same students on a standardized test in intermediate algebra.
Z = the semester grades of the same students in intermediate algebra.

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
54	56	67	90	94	91	27	46	35	88	95	90
55	64	67	63	79	77	78	54	76	72	70	82
64	67	74	43	56	60	10	19	20	55	59	61
33	43	48	69	48	70	49	39	60	61	68	76
57	55	60	47	48	52	46	58	50	33	52	50
42	59	60	62	59	67	70	41	62	65	45	65
88	84	81	92	64	90	45	37	50	84	82	88
85	84	86	75	68	85	95	99	92	55	60	52

Verify the following analysis:

$$M_X = 61 \qquad M_Y = 61 \qquad M_Z = 67$$
$$\sigma_X = 20.4 \qquad \sigma_Y = 17.9 \qquad \sigma_Z = 17.0$$
$$r_{XY} = 0.78 \qquad r_{XZ} = 0.94 \qquad r_{YZ} = 0.84$$

Which test was given the greater weight in the determination of the students' semester grades?

17. The following data are taken from the 1935 *World Almanac*, pp. 479, 499.

Column II gives the average attendance (in thousands) in New York City schools for the given years.

Column III gives the number (in thousands) arraigned before the Magistrates Courts in New York City in the same years.

Find ρ_{XY} or r_{XY} (rank) for these data. Would you say that this correlation is spurious? Explain.

Year	*Col. II*	*Col. III*	*Year*	*Col. II*	*Col. III*
1918	700	202	1923	853	420
1919	712	282	1924	870	455
1920	736	355	1925	891	440
1921	779	367	1926	910	437
1922	814	434	1927	926	527

Chapter 9

MULTIPLE CORRELATION

70. PRELIMINARY EXPLANATION

Our previous work in correlation has been concerned with problems involving only two variables, an independent variable X and a dependent variable Y. Such correlation is called "bivariate." It is obvious that many types of phenomena are affected by more than one factor and that the variations in the dependent variable may be due to the interaction of many forces.

In bivariate correlation we measure the relationship between the dependent variable Y and a single independent variable X, completely ignoring the influence upon Y of other forces that may be just as potent as X. Thus, on page 243 we measured the influence of July rainfall X upon the production of corn in Ohio Y. We found r to be 0.61 which shows that July rainfall does exert a significant influence upon the production of corn. But we may wonder if it exerts a greater influence than June rainfall or June temperature or July temperature. We are thus aware that the production of corn may be dependent upon several variables, and a consideration of the production in this regard would present a problem in multiple correlation. *Multiple correlation is then concerned with the combined influence of several independent variables upon a single dependent variable.*

As another illustration, suppose we have the scores made by a group of students on objective tests in English, Mathematics, and Intelligence. By means of simple correlation we can measure the relationship between the scores in Intelligence and those in Mathematics, between the scores in Intelligence and those in English, and between the scores in English and the scores in Mathematics. What we now need is a method of combining two factors, say English and Mathematics, in order that an estimate may be made of their influence *in combination* upon the third factor, Intelligence.

The method of procedure by which this may be accomplished is similar to that used in simple correlation.

71. THE CASE OF THREE VARIABLES

Let us assume that X_1, X_2, X_3 are three variable quantities which represent three interacting forces. Any one variable may be considered mathematically a function of the other two. As in the case of bivariate correlation, we shall assume that the relationships are linear, that is, that the N observed points representing the N observed sets of data are distributed about the plane

$$X_1 = b_{12}X_2 + b_{13}X_3 + c \tag{1}$$

in which X_2 and X_3 are independent variables and X_1 is the dependent variable.[1]

We shall determine the constants in accordance with the Principle of Least Squares: The plane best fitting a set of points is that one in which the constants are so determined that the sum of the squares of the X_1-residuals is a minimum.

An X_1-residual is defined by the equation

$$\rho = X_1 - (b_{12}X_2 + b_{13}X_3 + c) \tag{2}$$

We shall determine b_{12}, b_{13}, and c so that

$$\Sigma\rho^2 = \Sigma[X_1 - (b_{12}X_2 + b_{13}X_3 + c)]^2 \tag{3}$$

shall be a minimum. The conditions for this are that the first partial derivatives of $\Sigma\rho^2$ with respect to c, b_{12}, and b_{13} shall be equal to zero. Equating to zero these derivatives, we obtain the *normal equations*

$$\left.\begin{aligned} b_{12}\Sigma X_2 + b_{13}\Sigma X_3 + Nc &= \Sigma X_1 \\ b_{12}\Sigma X_2^2 + b_{13}\Sigma X_2X_3 + c\Sigma X_2 &= \Sigma X_1X_2 \\ b_{12}\Sigma X_2X_3 + b_{13}\Sigma X_3^2 + c\Sigma X_3 &= \Sigma X_1X_3 \end{aligned}\right\} \tag{4}$$

from which, by simultaneous solution, the values of b_{12}, b_{13}, and c may be determined in terms of the observed values X_1, X_2, X_3.

Thus, suppose we wish to find the plane

$$X_1 = b_{12}X_2 + b_{13}X_3 + c$$

that best fits the ten points (X_1, X_2, X_3) given in Table 63.

[1] The first subscript affixed to the regression coefficient b_{ij} will be the subscript of the letter X on the left (the dependent variable), and the second will be the subscript of the X to which it is attached.

TABLE 63

X_3	X_2	X_1	X_3^2	X_2^2	X_1^2	X_1X_2	X_1X_3	X_2X_3
2	2	11	4	4	121	22	22	4
3	4	17	9	16	289	68	51	12
4	6	26	16	36	676	156	104	24
5	5	28	25	25	784	140	140	25
6	8	31	36	64	961	248	186	48
7	7	35	49	49	1,225	245	245	49
9	10	41	81	100	1,681	410	369	90
10	11	49	100	121	2,401	539	490	110
11	13	63	121	169	3,969	819	693	143
13	14	69	169	196	4,761	966	897	182
70	80	370	610	780	16,868	3,613	3,197	687
$M_3 = 7$	$M_2 = 8$	$M_1 = 37$						

We complete the table to find the Σ functions that we need in the normal equations (4). Substituting in (4) we obtain

$$80b_{12} + 70b_{13} + 10c = 370$$
$$780b_{12} + 687b_{13} + 80c = 3613$$
$$687b_{12} + 610b_{13} + 70c = 3197$$

To solve these equations we divide each equation by the coefficient of b_{12} of that equation. We obtain

$$b_{12} + .875b_{13} + .125c = 4.625$$
$$b_{12} + .881b_{13} + .103c = 4.632$$
$$b_{12} + .888b_{13} + .102c = 4.654$$

Next we subtract the first equation from the second and the second equation from the third. We obtain

$$.006b_{13} - .022c = .007$$
$$.007b_{13} - .001c = .022$$

or, multiplying by 1,000

$$6b_{13} - 22c = 7$$
$$7b_{13} - c = 22$$

Solving these equations and substituting we obtain $b_{12} = 1.735$, $b_{13} = 3.223$, $c = 0.561$. The equation of the best fitting plane is

$$X_1 = 1.735X_2 + 3.223X_3 + 0.561$$

Exercise. Show that the point $(M_1, M_2, M_3) = (37, 8, 7)$ is on this plane.

We may test the goodness-of-fit of the plane by finding the computed values of X_1 for the given values of X_2 and X_3, and the X_1-residuals. The computed values of X_1, the (X_1-residuals), and the (X_1-residuals)^2 are shown in Table 64.

TABLE 64

X_3	X_2	X_1	*Computed* X_1	X_1-*residuals* ρ	(X_1-*residuals*)^2 ρ^2
2	2	11	10.477	+ 0.523	.274
3	4	17	17.170	− 0.170	.029
4	6	26	23.863	+ 2.137	4.567
5	5	28	25.351	+ 2.649	7.017
6	8	31	33.779	− 2.779	7.723
7	7	35	35.267	− 0.267	.071
9	10	41	46.918	− 5.918	35.023
10	11	49	51.876	− 2.876	8.271
11	13	63	58.569	+ 4.431	19.634
13	14	69	66.750	+ 2.250	5.062
				− 0.020	87.671 = $\Sigma\rho^2$

We note that five points are above the plane and five points are below it, and that the sum of the residuals is essentially zero. The sum of the squares of the X_1-residuals, $\Sigma\rho^2$, plays a rôle in multiple correlation similar to that played by $\Sigma\rho^2$ in simple correlation. (See p. 233.) It assists us in finding the *standard error* of estimate, $S_{1(23)}$. As we did in simple correlation, we define the standard error of estimate by the equation [1]

$$S_{1(23)} = \sqrt{\frac{\Sigma\rho^2}{N}}$$

This is a quantity which, when combined with the computed value of X_1, makes possible our measuring the confidence or the reliability we may place in values of X_1 estimated from the equation for given values of X_2 and X_3. Thus, the odds are 2 to 1 that, for given values of X_2 and X_3, the observed X_1 will lie within the interval

$$(\text{computed } X_1) \pm S_{1(23)}$$

[1] The subscript before the parenthesis designates the variable estimated (the dependent variable) and the subscripts within the parentheses designate the variables from which the estimate has been made.

Similarly, the odds are 19 to 1 that the observed X_1 will lie within

$$(\text{computed } X_1) \pm 2S_{1(23)}$$

and 385 to 1 that the observed X_1 will lie within

$$(\text{computed } X_1) \pm 3S_{1(23)}$$

For the problem we are considering

$$S_{1(23)} = \sqrt{\frac{87.671}{10}} = 2.689$$

It will be noted that only two of the ten points have residuals numerically larger than 2.689, and only one point has a residual numerically larger than 2(2.689).

In a later section we will discuss the *coefficient of multiple correlation* which is an expression that measures the degree of the relation between a single dependent variable, say X_1, and several independent variables, X_2 and X_3, in combination. We shall show that this coefficient $R_{1(23)}$ may be found from the formula

$$R_{1(23)} = \sqrt{1 - \frac{S^2_{1(23)}}{\sigma_1^2}}$$

where σ_1 means σ_{X_1}.

From Table 63 we find

$$\sigma_1 = \sqrt{\frac{\Sigma X_1^2}{N} - M_1^2} = \sqrt{\frac{16868}{10} - 37^2} = 17.83$$

Hence

$$R_{1(23)} = \sqrt{1 - \frac{8.767}{317.8}} = \sqrt{1 - .0275} = \sqrt{.9725} = 0.986$$

Thus, we have completed the analysis of the data of Table 63. This analysis has included finding (1) the best fitting plane, (2) the standard error of estimate, and (3) the coefficient of multiple correlation between X_1 and (X_2 and X_3) in combination.

EXERCISES

1. For the values of b_{12}, b_{13}, and c determined by (4), show that
(a) the algebraical sum of the X_1-residuals is equal to zero, and that
(b) the point (M_1, M_2, M_3) is on (1).

Note. The quantities M_1, M_2, and M_3 are the means of the variables X_1, X_2, and X_3 respectively.

2.

X_1	X_2	X_3
10	2	5
15	4	7
17	6	8
19	8	9
25	9	12
22	10	10
26	11	13
31	12	15
30	13	14
35	15	17

(1) Find the regression equation for these data with X_1 as dependent upon X_2 and X_3.
(2) Find the computed values of X_1 for the given values of X_2 and X_3.
(3) Find the X_1-residuals.
(4) Find $S_{1(23)}$ and $R_{1(23)}$.
(5) How many of the points are within (X_1 computed) $\pm S_{1(23)}$?

72. THE CASE OF THREE VARIABLES CONTINUED

Secondary Explanation

The method employed in the preceding section is satisfactory when the number, N, of sets of values is small, say less than forty. When N is large, as it usually is, we need a more systematic procedure. Further, the development of a theory in terms of the original variates, X_1, X_2, and X_3 is rather complex and tedious.

A simpler and more elegant procedure is to show that the centroidal point (M_1, M_2, M_3) is on the best-fitting plane, then to transform our variates to this centroidal point as origin. (We shall indicate the means of the variables X_1, X_2, and X_3 by M_1, M_2, and M_3 respectively, and their standard deviations by σ_1, σ_2, and σ_3.) We shall prove that (M_1, M_2, M_3) satisfies equation (1) for the values of b_{12}, b_{13}, and c determined by equations (4).

If the first of equations (4) be divided by N we have

$$b_{12}\frac{\Sigma X_2}{N} + b_{13}\frac{\Sigma X_3}{N} + c = \frac{\Sigma X_1}{N}$$

or

$$b_{12}M_2 + b_{13}M_3 + c - M_1 = 0$$

which is the condition that (M_1, M_2, M_3) is on (1).

We now translate our data to the centroidal point as origin and take the equation of the plane through this point to be

$$x_1 = b_{12}x_2 + b_{13}x_3$$

where

$$x_1 = X_1 - M_1, \qquad x_2 = X_2 - M_2, \qquad x_3 = X_3 - M_3$$

For this form of the regression plane any x_1-residual is given by

$$\rho = x_1 - (b_{12}x_2 + b_{13}x_3)$$

and by equating to zero the first partial derivatives of

$$\Sigma\rho^2 = \Sigma[x_1 - (b_{12}x_2 + b_{13}x_3)]^2 \tag{5}$$

with respect to b_{12} and b_{13}, we obtain the normal equations

$$\left.\begin{aligned} b_{12}\Sigma x_2^2 + b_{13}\Sigma x_2x_3 &= \Sigma x_1x_2 \\ b_{12}\Sigma x_2x_3 + b_{13}\Sigma x_3^2 &= \Sigma x_1x_3 \end{aligned}\right\} \tag{6}$$

Let σ_i be the standard deviation of the N values of X_i, and let r_{pq} be the correlation coefficient of the N given pairs of values of X_p and X_q. Thus $\Sigma x_2^2 = N\sigma_2^2$, $\Sigma x_3^2 = N\sigma_3^2$, $\Sigma x_1x_2 = N\sigma_1\sigma_2r_{12}$, $\Sigma x_1x_3 = N\sigma_1\sigma_3r_{13}$, $\Sigma x_2x_3 = N\sigma_2\sigma_3r_{23}$.

By expressing the summations in terms of the standard deviations and correlation coefficients, the normal equations (6) after simplification become

$$\left.\begin{aligned} b_{12}\sigma_2 + b_{13}\sigma_3r_{23} &= \sigma_1r_{12} \\ b_{12}\sigma_2r_{23} + b_{13}\sigma_3 &= \sigma_1r_{13} \end{aligned}\right\} \tag{7}$$

Solving the normal equations (7) we obtain the regression coefficients

$$\left.\begin{aligned} b_{12} &= \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\,\frac{\sigma_1}{\sigma_2} \\ b_{13} &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\,\frac{\sigma_1}{\sigma_3} \end{aligned}\right\} \tag{8}$$

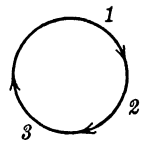
and the regression plane is thus

$$\frac{x_1}{\sigma_1}(1 - r_{23}^2) = \frac{x_2}{\sigma_2}(r_{12} - r_{13}r_{23}) + \frac{x_3}{\sigma_3}(r_{13} - r_{12}r_{23}) \tag{9}$$

In terms of the original variates X_1, X_2, X_3 the equation of the regression plane is

$$\frac{(X_1-M_1)}{\sigma_1}(1-r_{23}^2) = \frac{(X_2-M_2)}{\sigma_2}(r_{12}-r_{13}r_{23}) + \frac{(X_3-M_3)}{\sigma_3}(r_{13}-r_{12}r_{23}) \tag{10}$$

Equation (10) gives the most probable value of X_1 for assigned values of X_2 and X_3. Analogous equations may be written with X_2 and X_3 as the dependent variables by cyclically permuting the subscripts 1, 2, and 3; that is, replacing 1 by 2, 2 by 3, and 3 by 1, as if one were going around the circle in the direction indicated by the figure.



We have thus reached a result which gives an effective summary of the manner in which X_2 and X_3 in combination affect X_1. Further, it is delightful to observe that this summarizing equation involves nothing more complicated than simple correlation coefficients.

EXERCISES

1.

X_1	X_2	X_3
2	26	1
4	20	2
6	20	3
9	17	4
5	7	5
5	5	6
11	3	7

(1) Verify the following:

$M_1 = 6$ $\quad M_2 = 14$ $\quad M_3 = 4$

$\sigma_1 = 2.828$ $\quad \sigma_2 = 8.246$ $\quad \sigma_3 = 2$

$r_{12} = -0.551$ $\quad r_{13} = 0.707$ $\quad r_{23} = -0.970$

(2) Find the regression plane with X_1 as dependent on X_2 and X_3.

(3) Find $R_{1(23)}$ and $S_{1(23)}$.

2.

X_1	X_2	X_3
5	4	5
4	5	2
5	6	4
6	4	9
9	5	8
10	6	4
9	6	10
12	7	11
11	9	10
9	8	7

(1) Verify the following:

$M_1 = 8$ $\quad M_2 = 6$ $\quad M_3 = 7$

$\sigma_1 = 2.646$ $\quad \sigma_2 = 1.549$ $\quad \sigma_3 = 2.933$

$r_{12} = .683$ $\quad r_{13} = .696$ $\quad r_{23} = .374$

(2) Find the regression plane with X_1 as dependent on X_2 and X_3.

(3) Find the computed values of X_1 and the X_1-residuals.

(4) Find $R_{1(23)}$ and $S_{1(23)}$.

(5) How many of the points are within (X_1 computed) $\pm S_{1(23)}$?

3. In the following table

X_1 = the semester grades of 32 students in intermediate algebra

X_2 = the scores of the same students on a standardized test in intermediate algebra

X_3 = the scores of the same students on the Bucknell test in intermediate algebra

X_3	X_2	X_1	X_3	X_2	X_1	X_3	X_2	X_1	X_3	X_2	X_1
54	56	67	90	94	91	27	46	35	88	95	90
55	64	67	63	79	77	78	54	76	72	70	82
64	67	74	43	56	60	10	19	20	55	59	61
33	43	48	69	48	70	49	39	60	61	68	76
57	55	60	47	48	52	46	58	50	33	52	50
42	59	60	62	59	67	70	41	62	65	45	65
88	84	81	92	64	90	45	37	50	84	82	88
85	84	86	75	68	85	95	99	92	55	60	52

(1) Verify the following values:

$M_1 = 67$ $\quad M_2 = 61$ $\quad M_3 = 61$
$\sigma_1 = 17.0$ $\quad \sigma_2 = 17.9$ $\quad \sigma_3 = 20.4$
$r_{12} = 0.84$ $\quad r_{13} = 0.94$ $\quad r_{23} = 0.78$

(2) Find $R_{1(23)}$ and $S_{1(23)}$.
(3) Find the equation of the regression plane with X_1 dependent upon X_2 and X_3. Show that the point (M_1, M_2, M_3) is on this plane.
(4) What meaning do you attach to the values of b_{12} and b_{13}?
(5) Estimate X_1 for $X_2 = 84$ and $X_3 = 81$. Use your value of $S_{1(23)}$ to interpret this estimate.

4. The following table gives a summary of the fundamental statistical constants that were obtained from scores made on objective tests in English, Mathematics, and Intelligence by 343 Bucknell freshmen.

(1) Find the equation for the regression of Intelligence on English and Mathematics.
(2) What is the estimated Intelligence score for an individual whose English score was 172 and whose Mathematics score was 40?
(3) What is the estimated Mathematics score of an individual whose English score was 160 and whose Intelligence score was 150?

FUNDAMENTAL CONSTANTS FROM 343 SCORES IN THE TESTS GIVEN IN ENGLISH, MATHEMATICS, AND INTELLIGENCE

		English	*Mathematics*	*Intelligence*
Correlations	English	1.00	0.30	0.65
	Mathematics	0.30	1.00	0.46
	Intelligence	0.65	0.46	1.00
Arithmetic Means		151	34	140
Standard Deviations		44	12	45

73. COEFFICIENT OF MULTIPLE CORRELATION

Three Variables

It is evident that the value of equation (9) or (10) as a tool for purposes of estimation depends upon the closeness of fit of the plane to the points. As was suggested in the preceding section we shall use for measuring the goodness of fit of the plane to the points the standard error of estimate, $S_{1(23)}$,

$$S_{1(23)} = \sqrt{\frac{\Sigma\rho^2}{N}}$$

where $\Sigma\rho^2$ is determined from the values of b_{12} and b_{13} in (7) or (8).

From (5)

$$\begin{aligned}\Sigma\rho^2 &= \Sigma[x_1 - (b_{12}x_2 + b_{13}x_3)]^2 \\ &= \Sigma x_1^2 + b_{12}^2\Sigma x_2^2 + b_{13}^2\Sigma x_3^2 - 2b_{12}\Sigma x_1x_2 - 2b_{13}\Sigma x_1x_3 + 2b_{12}b_{13}\Sigma x_2x_3\end{aligned}$$

which may be written in the form

$$\Sigma\rho^2 = N[\sigma_1^2 + b_{12}^2\sigma_2^2 + b_{13}^2\sigma_3^2 - 2b_{12}\sigma_1\sigma_2r_{12} - 2b_{13}\sigma_1\sigma_3r_{13} + 2b_{12}b_{13}\sigma_2\sigma_3r_{23}] \qquad (11)$$

We desire the value of $\Sigma\rho^2$ for the values of b_{12} and b_{13} given by (7) or (8). This may be easily found by multiplying the normal equations (7) by $b_{12}\sigma_2$ and $b_{13}\sigma_3$ respectively, adding the results, and substituting the results in (11). The value for $\Sigma\rho^2$ then becomes

$$\Sigma\rho^2 = N[\sigma_1^2 - b_{12}\sigma_1\sigma_2r_{12} - b_{13}\sigma_1\sigma_3r_{13}] \qquad (12)$$

If now the values of b_{12} and b_{13} given by (8) are substituted, we have

$$S_{1(23)}^2 = \sigma_1^2\left[1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}\right] \qquad (13)$$

or

$$S_{1(23)} = \sigma_1\sqrt{1 - R_{1(23)}^2} \qquad (14)$$

where

$$R_{1(23)} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \qquad (15)$$

is the "coefficient of multiple correlation" of X_1 on X_2 and X_3.

By permuting the subscripts we may write down the values of $R_{2(13)}$ and $R_{3(12)}$. Due to the fact that we have no mathematical method of attaching a meaning to the algebraical sign of $R_{1(23)}$, it is customary to write it without sign.

From (13) we may note that since $S^2_{1(23)}$ is a positive quantity, $0 \leqq R^2_{1(23)} \leqq 1$. When $R_{1(23)}$ is equal to unity numerically, that is, when

$$r^2_{12} + r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23} = 1$$

we have perfect multiple correlation. In this case the points are on the regression plane.

The coefficient of multiple correlation is an expression which measures the degree of relationship between a single dependent variable and a number of independent variables in combination. It is more accurately defined as the ordinary cross-product coefficient of correlation between the X_1 estimated by (10) and the observed X_1, or between x_1 estimated by (9) and the observed x_1. (See Exercise 1 of the next list of exercises.)

EXERCISES

1. (a) The estimated value of x_1, say x_{1e}, may be found from $x_{1e} = b_{12}x_2 + b_{13}x_3$ where b_{12} and b_{13} are given by (7) or (8). Show that the standard deviation σ_{1e} of x_{1e} is given by $\sigma^2_{1e} = b_{12}\sigma_1\sigma_2 r_{12} + b_{13}\sigma_1\sigma_3 r_{13}$.

Hint: Use $\sigma^2_{1e} = \dfrac{\Sigma x^2_{1e}}{N} - \left[\dfrac{\Sigma x_{1e}}{N}\right]^2$ and equations (7).

(b) Show that $S^2_{1(23)} = \sigma^2_1 - \sigma^2_{1e}$.

Hint: Use equation (12) and (a).

(c) Show that $R_{1(23)} = \dfrac{\sigma_{1e}}{\sigma_1}$.

(d) Show that $\Sigma x_1 x_{1e} = N\sigma^2_{1e}$.

Hint: Multiply the value of x_{1e} in (a) by x_1, and sum. Change the Σ quantities on the right-hand side into statistical symbols.

(e) Show that $r_{1\,1e} = \dfrac{\sigma_{1e}}{\sigma_1}$.

(f) Show that $R_{1(23)} = r_{1\,1e}$.

2. Show that $R^2_{1(23)} = \dfrac{b_{12}\Sigma x_1 x_2 + b_{13}\Sigma x_1 x_3}{\Sigma x^2_1}$.

3. Show that $R^2_{1(23)} = \dfrac{1}{\sigma_1}[r_{12}b_{12}\sigma_2 + r_{13}b_{13}\sigma_3]$.

4. Show that, for the least-squares plane, the algebraical sum of the residuals is zero.

5. State three important properties of the least-squares plane fitting a set of N points.

6. (Davies and Crowder.) The following table gives the rankings of the specified states in 1860.

You can save labor by using the values of ΣX and ΣX^2 given on pages 9 and 10, or by using (20) page 265.

X_1 = rank of the specified state in notables
X_2 = rank of the specified state in education
X_3 = rank of the specified state in capital

State	X_1	X_2	X_3	*State*	X_1	X_2	X_3
Alabama	24	23	24	Mississippi	28	17	27
Arkansas	29	27	29	Missouri	19	18	19
Connecticut	2	2	3	New Hampshire	5	5	7
Delaware	8	19	8	New Jersey	9	13	4
Florida	27	29	28	New York	7	7	6
Georgia	25	25	23	North Carolina	22	28	22
Illinois	14	14	16	Ohio	10	11	10
Indiana	17	16	14	Pennsylvania	12	10	5
Iowa	16	12	26	Rhode Island	4	8	1
Kentucky	20	22	15	South Carolina	21	21	21
Louisiana	26	20	25	Tennessee	23	24	18
Maine	6	4	12	Vermont	3	3	11
Maryland	13	15	9	Virginia	18	26	13
Massachusetts	1	1	2	Wisconsin	15	6	20
Michigan	11	9	17				

(1) Verify the values:

$$M_1 = 15 \qquad M_2 = 15 \qquad M_3 = 15$$
$$\sigma_1 = 8.367 \qquad \sigma_2 = 8.367 \qquad \sigma_3 = 8.367$$
$$r_{12} = 0.867 \qquad r_{13} = 0.886 \qquad r_{23} = 0.670$$

(2) Find $R_{1(23)}$ and $S_{1(23)}$.

74. DETERMINANTS

A. Determinants of the Second Order.

If we solve the equations

$$a_1x_1 + b_1y_1 = c_1$$
$$a_2x_1 + b_2y_1 = c_2$$

simultaneously, we obtain the solutions:

$$x_1 = \frac{c_1b_2 - c_2b_1}{a_1b_2 - a_2b_1} \qquad y_1 = \frac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1}$$

By adopting the shorthand notation

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2 - a_2b_1 \qquad \begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix} = c_1b_2 - c_2b_1 \text{ etc.}$$

we may write the solutions

$$x_1 = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} \qquad y_1 = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}$$

The square arrays defined above are called *determinants.* Since there are two rows and two columns, we call the arrays *determinants of the second order.* The letters a_1, a_2, b_1, b_2, etc. are called the *elements* of the determinant. The elements a_1, b_2 constitute the *principal diagonal* of the determinant found in the denominators of x_1 and y_1.

We note that the denominators of x_1 and y_1 are the same determinant, that formed from the coefficients as they stand in the given equations. Further, we note that the numerator for x_1 may be obtained from the denominator by replacing a_1, a_2, which are coefficients of x_1 in the given equations, by the terms c_1, c_2. Similarly, the numerator for y_1 is the determinant of the denominator with b_1, b_2 replaced by c_1, c_2 respectively. The determinant of the denominator is called the *determinant of the system.*

Example. Solve by determinants:

$$\begin{aligned} x + y &= 3 \\ 2x + 3y &= 1 \end{aligned}$$

Solution:

$$x = \frac{\begin{vmatrix} 3 & 1 \\ 1 & 3 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 2 & 3 \end{vmatrix}} = \frac{9 - 1}{3 - 2} = 8 \qquad y = \frac{\begin{vmatrix} 1 & 3 \\ 2 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 2 & 3 \end{vmatrix}} = \frac{1 - 6}{3 - 2} = -5$$

EXERCISES

Solve the following pairs of equations using determinants:

1. $x + y = 2$
$2x + 3y = 7$

2. $x - 3y = 6$
$4x - 5y = 24$

3. $0.3x + 0.2y = 4.0$
$0.7x - 0.6y = 26.4$

4. $4x - 8y = 17$
$12x + 16y = -9$

B. Determinants of the Third Order. The solution of three equations in three unknowns is also facilitated by the use of determinants. In this case we have square arrays of three rows and three columns or *determinants of the third order.* The square array in the left-hand member of the equality

$$D = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - b_1 \begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + c_1 \begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix}$$

is a determinant of the third order. It is defined in terms of determinants of the second order as in the right-hand member of the above equality which is called the *expansion* of the determinant. The elements a_1, b_2, c_3 constitute the *principal* diagonal.

The second order determinants in the above equality are called *minors* of the elements a_1, b_1, c_1 respectively. The minor to a_1 is the determinant that remains after crossing out the row and the column in which a_1 lies. Similarly the minor for any other element is found.

The above determinant was expanded *according to the elements of the first row.* We may also expand it according to the elements of the first column. Thus,

$$D = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & c_1 \\ b_3 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}$$

It is obvious that the complete development of a determinant of the third order has six terms. Thus,

$$D = a_1b_2c_3 + a_2b_3c_1 + a_3b_1c_2 - a_1b_3c_2 - a_3b_2c_1 - a_2b_1c_3$$

If we solve by elementary algebra the equations

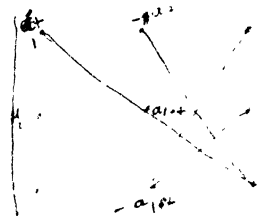
$$\begin{aligned} a_1x + b_1y + c_1z &= d_1 \\ a_2x + b_2y + c_2z &= d_2 \\ a_3x + b_3y + c_3z &= d_3 \end{aligned}$$

for x, we obtain

$$x = \frac{d_1b_2c_3 + d_2b_3c_1 + d_3b_1c_2 - d_1b_3c_2 - d_3b_2c_1 - d_2b_1c_3}{a_1b_2c_3 + a_2b_3c_1 + a_3b_1c_2 - a_1b_3c_2 - a_3b_2c_1 - a_2b_1c_3}$$

The denominator is the development of the determinant D, above, and the numerator is the same as the denominator with a_i replaced by d_i, $i = 1, 2, 3$. Hence we can write

$$x = \frac{\begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}$$



In a similar way we can find y and z:

$$y = \frac{\begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}} \qquad z = \frac{\begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}$$

We note that the denominators of x, y, and z are the same, the determinant of the system. The determinant in the numerator of any unknown can be obtained from the denominator by replacing the column of the coefficients of this unknown by the corresponding known terms, d_1, d_2, d_3.

In the expansion of D we note that the sign preceding the minor of a_1 is plus, that preceding the minor of a_2 is minus, that preceding the minor of a_3 is plus. The sign preceding a minor corresponding to an element is easy to remember. Consider an element in the h-row and k-column. If $(h + k)$ is even the sign prefixed to the minor is plus, and if $(h + k)$ is odd the sign prefixed to the minor

is minus. The minor of an element with its sign attached is called the *co-factor* of the element. We note that D is equal to the sum of the products of any row (or column) and their respective co-factors.

EXERCISES

1. Evaluate the determinant $\begin{vmatrix} 1 & 3 & 4 \\ 2 & 7 & 3 \\ 1 & 3 & 5 \end{vmatrix}$ by expanding (a) according to the elements in the first row, and (b) according to the elements in the first column.

Solve for x, y, and z the equations:

2.
$$x - y - z = -6$$
$$2x + y + z = 0$$
$$3x - 5y + 8z = 13$$

3.
$$x + 2y - z = 6$$
$$2x - y + 3z = -13$$
$$3x - 2y + 3z = 16$$

C. Determinants of Any Order. We defined a determinant of the third order in terms of the elements of a row (or column) and their minors. Similarly we may define determinants of the fourth and higher orders. Thus, the following determinant of the fourth order

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 & d_2 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & c_1 & d_1 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & c_1 & d_1 \\ b_2 & c_2 & d_2 \\ b_4 & c_4 & d_4 \end{vmatrix} - a_4 \begin{vmatrix} b_1 & c_1 & d_1 \\ b_2 & c_2 & d_2 \\ b_3 & c_3 & d_3 \end{vmatrix}$$

is defined in terms of the elements of the first column and their minors. The sign preceding a minor of an element in h-row and k-column is plus or minus according as $(h + k)$ is even or odd. A minor of an element with its sign attached is the co-factor of the element. The value of a determinant is the sum of the products of the elements of a row (or column) and their co-factors.

Just as we define determinants of the third and fourth orders in terms of the elements of a row or column and their co-factors, so we define a *determinant of any order to be the sum of the products of the elements of a row* (or column) *and their respective co-factors.*

EXERCISES

1. Expand the following determinants (a) according to the elements of the first row, and (b) according to the elements of the first column.

$$(1)\quad \begin{vmatrix} 2 & 4 & -2 & 3 \\ 1 & -2 & 1 & 0 \\ -2 & 0 & -1 & 3 \\ 2 & 3 & -2 & 3 \end{vmatrix} \qquad (2)\quad \begin{vmatrix} -2 & 1 & 3 & 0 \\ 5 & -3 & 3 & 1 \\ 4 & 0 & 2 & 4 \\ 1 & 2 & 3 & 3 \end{vmatrix}$$

2. The following theorems are true for determinants of any order We ask the student to prove them for determinants of the third order.

(1) If the corresponding rows and columns of D be interchanged, D is unchanged in value.
(2) If any two rows (or columns) of D be interchanged, D becomes $-D$.
(3) If any two rows (or columns) be identical, $D = 0$.
(4) If each element of a row (or column) of D be multiplied by k, the value of the resulting determinant is kD.
(5) If to each element of a row (or column) of D is added k times the corresponding element of another row (or column), D is unchanged in value.

75. APPLICATION OF DETERMINANTS

Three Variables

The results of the analysis of the foregoing sections on multiple correlation can be expressed in very simple forms by the use of determinants.

Let

$$D = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

where r_{hk} is the element in the h-row and the k-column. Evidently $r_{hh} = r_{kk} = 1$, and $r_{hk} = r_{kh}$.

A *minor* D_{hk} of the element r_{hk} is the determinant formed by the elements that remain after striking out all the coefficients in the row and the column common to r_{hk}. Thus, for examples,

$$D_{11} = \begin{vmatrix} r_{22} & r_{23} \\ r_{32} & r_{33} \end{vmatrix} = 1 - r_{23}^2$$

$$D_{12} = \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & r_{33} \end{vmatrix} = r_{12} - r_{13}r_{23}$$

$$D_{13} = \begin{vmatrix} r_{21} & r_{22} \\ r_{31} & r_{32} \end{vmatrix} = r_{12}r_{23} - r_{13}$$

A *co-factor* A_{hk} of the element r_{hk} *is the minor* D_{hk} *with the sign* that would be prefixed to it when the determinant D is expanded. The sign that is prefixed to the minor is positive or negative according as $(h + k)$ is even or odd. That is

$$A_{hk} = (-1)^{h+k} D_{hk}$$

Expanding D according to the elements of the first row, we have

$$\left.\begin{aligned} D &= r_{11}D_{11} - r_{12}D_{12} + r_{13}D_{13} \\ &= r_{11}A_{11} + r_{12}A_{12} + r_{13}A_{13} \\ &= 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \end{aligned}\right\} \tag{16}$$

Now let us solve equations (7) by determinants and note the simplicity of the results.

$$\left.\begin{aligned} b_{12}\sigma_2 + b_{13}\sigma_3 r_{23} &= \sigma_1 r_{12} \\ b_{12}\sigma_2 r_{23} + b_{13}\sigma_3 &= \sigma_1 r_{13} \end{aligned}\right\} \tag{7}$$

We obtain

$$b_{12} = \frac{\begin{vmatrix} \sigma_1 r_{12} & \sigma_3 r_{23} \\ \sigma_1 r_{13} & \sigma_3 \end{vmatrix}}{\begin{vmatrix} \sigma_2 & \sigma_3 r_{23} \\ \sigma_2 r_{23} & \sigma_3 \end{vmatrix}} = \frac{\sigma_1}{\sigma_2} \frac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}$$

$$b_{13} = \frac{\begin{vmatrix} \sigma_2 & \sigma_1 r_{12} \\ \sigma_2 r_{23} & \sigma_1 r_{13} \end{vmatrix}}{\begin{vmatrix} \sigma_2 & \sigma_3 r_{23} \\ \sigma_2 r_{23} & \sigma_3 \end{vmatrix}} = \frac{\sigma_1}{\sigma_3} \frac{\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}$$

The regression coefficients in the determinant notation are

$$\left.\begin{aligned} b_{12} &= \frac{D_{12}}{D_{11}} \frac{\sigma_1}{\sigma_2} = \frac{-A_{12}}{A_{11}} \frac{\sigma_1}{\sigma_2} \\ b_{13} &= \frac{-D_{13}}{D_{11}} \frac{\sigma_1}{\sigma_3} = \frac{-A_{13}}{A_{11}} \frac{\sigma_1}{\sigma_3} \end{aligned}\right\} \tag{17}$$

and the equations of the regression planes (9) and (10) are

$$\frac{x_1}{\sigma_1}A_{11} + \frac{x_2}{\sigma_2}A_{12} + \frac{x_3}{\sigma_3}A_{13} = 0$$

and

$$\frac{(X_1 - M_1)}{\sigma_1}A_{11} + \frac{(X_2 - M_2)}{\sigma_2}A_{12} + \frac{(X_3 - M_3)}{\sigma_3}A_{13} = 0 \tag{18}$$

or

$$\sum_{i=1}^{3} t_i A_{1i} = 0$$

where

$$t_i = \frac{x_i}{\sigma_i}.$$

Applying the determinant notation to equation (13) we get

$$S_{1(23)} = \sigma_1 \sqrt{\frac{D}{D_{11}}} \tag{19}$$

which, when substituted in (14), leads to

$$R_{1(23)} = \sqrt{1 - \frac{D}{D_{11}}} \tag{20}$$

EXERCISES

1. Analyze the data of Exercise 3, page 284, using determinants.
2. Analyze the data of Exercise 4, page 285, using determinants.
3. Analyze the data of Exercise 6, page 288, using determinants.

76. PARTIAL CORRELATION

Sometimes a correlation between two factors is due to the influence of one or more other factors rather than to any inherent relationship between the two themselves.[1] For this reason it is necessary to eliminate as far as possible those uncontrolled factors which, through their common relation to the variables to be correlated, tend to influence unduly the true correlation. This is accomplished by a technique known as *partial* correlation.

It is desirable, therefore, to obtain the correlation between X_1 and X_2, say, when X_3 has a fixed value. For example, we can find the correlation between English and Mathematics (p. 285) when Intelligence is constant, say 100, but *not completely ignored.*

In bivariate correlation, it will be recalled that the values of the regression coefficients b_{12} and b_{21} of the regression equations

$$X_1 = b_{12}X_2 + c_1 \quad \text{and} \quad X_2 = b_{21}X_1 + c_2$$

were found to be[2]

$$b_{12} = r_{12}\frac{\sigma_1}{\sigma_2} \quad \text{and} \quad b_{21} = r_{12}\frac{\sigma_2}{\sigma_1}$$

[1] See page 268.

[2] See pages 248 and 249.

The quantity b_{12} measures the regression of X_1 on X_2 and b_{21} measures the regression of X_2 on X_1 *when all other factors are ignored.* We also found that

$$r_{12}^2 = b_{12} \cdot b_{21} \tag{21}$$

Similarly, $b_{12.3}$ and $b_{21.3}$ measure the regression of X_1 on X_2 and of X_2 on X_1 respectively in the equations [1]

$$X_1 = b_{12.3}X_2 + b_{13.2}X_3 + c_1 \quad \text{and} \quad X_2 = b_{21.3}X_1 + b_{23.1}X_3 + c_2$$

when X_3 is held constant but not ignored. Since the conditions leading to equation (21) in bivariate correlation are exactly paralleled here, we define the partial correlation coefficient $r_{12.3}$ between X_1 and X_2 for an assigned value of X_3 by the equation

$$r_{12.3}^2 = b_{12.3} \cdot b_{21.3} \quad \text{or} \quad r_{12.3} = \sqrt{b_{12.3}b_{21.3}}$$

In terms of the constants previously determined in (17) we find

$$r_{12.3} = \pm\sqrt{\frac{D_{12}}{D_{11}}\frac{\sigma_1}{\sigma_2} \cdot \frac{D_{21}}{D_{22}}\frac{\sigma_2}{\sigma_1}} = \frac{\pm D_{12}}{\sqrt{D_{11}D_{22}}} = \frac{\pm A_{12}}{\sqrt{A_{11}A_{22}}} \tag{22}$$

since the major determinant is symmetrical about the principal diagonal and hence $D_{12} = D_{21}$. The sign attached to $r_{12.3}$ is that of $b_{12.3}$ or b_{12}.

It is noted that $r_{12.3}$ is generally unequal to r_{12}. The quantity r_{12} measures the degree of correlation between X_1 and X_2 when all other factors are completely ignored whereas $r_{12.3}$ measures the degree of correlation between X_1 and X_2 when X_3 is held fixed but not ignored. The principal application of partial correlation is thus approximating what the correlation between two variables would be if the influence of other variables was eliminated.

Professor Sorenson [2] gives an interesting illustration that shows the influence of the third variable on the correlation between the other two variables. In his illustration

X_1 represents the carpal area of children
X_2 represents the mental age of children
X_3 represents the chronological age of children

[1] The subscripts following the point merely indicate the variables that are held fixed in the development. They may frequently be omitted from the detail.

[2] Herbert Sorenson: *Statistics for Students of Psychology and Education*, p. 252.

The following simple correlations were obtained.

$$r_{12} = 0.83 \qquad r_{13} = 0.92 \qquad r_{23} = 0.88$$

Naturally we are impressed by the apparently large correlation between the skeletal development (carpal area) of children and their mental age. When we "partial out" or remove the influence of the third factor, chronological age, we find

$$r_{12.3} = \frac{D_{12}}{\sqrt{D_{11}D_{22}}} = \frac{0.0204}{\sqrt{(0.2256)(0.1536)}} = 0.11$$

which indicates very slight, if any, correlation.

EXERCISES

1. Express $r_{12.3}$ in terms of simple correlation coefficients.
2. Write down the values of $r_{13.2}$ and $r_{23.1}$.
3. Show that: $S_{1(23)} = \sigma_1\sqrt{(1 - r_{12}^2)(1 - r_{13.2}^2)}$
$= \sigma_1\sqrt{(1 - r_{13}^2)(1 - r_{12.3}^2)}$
4. By permuting the subscripts in number 3 preceding, write down the values for $S_{2(13)}$ and $S_{3(12)}$.
5. In a certain study of a group of students' grades

X_1 denotes the percentage grades in mathematics
X_2 denotes the percentage grades in chemistry
X_3 denotes the percentage grades in history

$M_1 = 72$	$\sigma_1 = 8$	$r_{12} = .6$
$M_2 = 68$	$\sigma_2 = 10$	$r_{13} = .4$
$M_3 = 78$	$\sigma_3 = 7$	$r_{23} = .3$

What is the probable grade in chemistry of a student whose grades are: mathematics, 80%; history, 70%?

77. THE CASE OF FOUR VARIABLES

In the preceding sections we have considered in great detail the case of multiple and partial correlation based upon three variables. We shall greatly abbreviate the theory for the case of four variables leaving the details to be supplied by the reader.

Assume that we have N sets of data in the four variables X_1, X_2, X_3, X_4 and that we wish to determine the regression coefficients

b_{12}, b_{13}, b_{14}, and the constant c so that X_1 computed from the hyperplane

$$X_1 = b_{12}X_2 + b_{13}X_3 + b_{14}X_4 + c \tag{23}$$

may be the best estimate of X_1 for assigned values of X_2, X_3, X_4. Adopting the least-squares criterion, we may determine the regression coefficients so that

$$\Sigma\rho^2 = \Sigma[X_1 - (b_{12}X_2 + b_{13}X_3 + b_{14}X_4 + c)]^2 \tag{24}$$

shall be a minimum.

Equating to zero the first partial derivatives of $\Sigma\rho^2$ with respect to c, b_{12}, b_{13}, b_{14}, we obtain the normal equations

$$\left.\begin{aligned} b_{12}\Sigma X_2 + b_{13}\Sigma X_3 + b_{14}\Sigma X_4 + Nc &= \Sigma X_1 \\ b_{12}\Sigma X_2^2 + b_{13}\Sigma X_2X_3 + b_{14}\Sigma X_2X_4 + c\Sigma X_2 &= \Sigma X_1X_2 \\ b_{12}\Sigma X_2X_3 + b_{13}\Sigma X_3^2 + b_{14}\Sigma X_3X_4 + c\Sigma X_3 &= \Sigma X_1X_3 \\ b_{12}\Sigma X_2X_4 + b_{13}\Sigma X_3X_4 + b_{14}\Sigma X_4^2 + c\Sigma X_4 &= \Sigma X_1X_4 \end{aligned}\right\} \tag{25}$$

By dividing the first of equations (25) by N, we may show that the hyperplane (23) for the values of b_{12}, b_{13}, b_{14}, c given by (25), passes through the point (M_1, M_2, M_3, M_4).

Referring our data to this point as origin our regression equation becomes

$$x_1 = b_{12}x_2 + b_{13}x_3 + b_{14}x_4 \tag{26}$$

where $$x_i = X_i - M_i, \quad i = 1, 2, 3, 4$$

That is, our regression equation is of the form (26) when the variables are deviations from their respective means.

By minimizing the sum of the squares of the x_1-residuals,

$$\Sigma\rho^2 = \Sigma[x_1 - (b_{12}x_2 + b_{13}x_3 + b_{14}x_4)]^2$$

we obtain the normal equations

$$\left.\begin{aligned} b_{12}\Sigma x_2^2 + b_{13}\Sigma x_2x_3 + b_{14}\Sigma x_2x_4 &= \Sigma x_1x_2 \\ b_{12}\Sigma x_2x_3 + b_{13}\Sigma x_3^2 + b_{14}\Sigma x_3x_4 &= \Sigma x_1x_3 \\ b_{12}\Sigma x_2x_4 + b_{13}\Sigma x_3x_4 + b_{14}\Sigma x_4^2 &= \Sigma x_1x_4 \end{aligned}\right\} \tag{27}$$

Expressing the summations in terms of standard deviations and coefficients of correlation, equations (27) become

$$\left.\begin{aligned} b_{12}\sigma_2 + b_{13}\sigma_3 r_{23} + b_{14}\sigma_4 r_{24} &= \sigma_1 r_{12} \\ b_{12}\sigma_2 r_{23} + b_{13}\sigma_3 + b_{14}\sigma_4 r_{34} &= \sigma_1 r_{13} \\ b_{12}\sigma_2 r_{24} + b_{13}\sigma_3 r_{34} + b_{14}\sigma_4 &= \sigma_1 r_{14} \end{aligned}\right\} \tag{28}$$

Let D denote the major determinant:

$$D = \begin{vmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{vmatrix}$$

Further, let D_{hk} be the minor and A_{hk} the co-factor of r_{hk} so that $A_{hk} = (-1)^{h+k} D_{hk}$. Then the solutions of (28) become

$$\left.\begin{aligned} b_{12} &= \frac{\sigma_1}{\sigma_2}\frac{D_{12}}{D_{11}} = -\frac{\sigma_1}{\sigma_2}\frac{A_{12}}{A_{11}} \\ b_{13} &= -\frac{\sigma_1}{\sigma_3}\frac{D_{13}}{D_{11}} = -\frac{\sigma_1}{\sigma_3}\frac{A_{13}}{A_{11}} \\ b_{14} &= \frac{\sigma_1}{\sigma_4}\frac{D_{14}}{D_{11}} = -\frac{\sigma_1}{\sigma_4}\frac{A_{14}}{A_{11}} \end{aligned}\right\} \tag{29}$$

and the equations of the regression hyperplane become

$$\frac{x_1}{\sigma_1}A_{11} + \frac{x_2}{\sigma_2}A_{12} + \frac{x_3}{\sigma_3}A_{13} + \frac{x_4}{\sigma_4}A_{14} = 0 \tag{30}$$

and

$$\frac{(X_1 - M_1)}{\sigma_1}A_{11} + \frac{(X_2 - M_2)}{\sigma_2}A_{12} + \frac{(X_3 - M_3)}{\sigma_3}A_{13} + \frac{(X_4 - M_4)}{\sigma_4}A_{14} = 0 \tag{31}$$

expressed in terms of the deviations from their respective means and the original variates respectively.

If the respective deviations from the means be expressed in units of their standard deviations, that is, if

$$t_i = \frac{x_i}{\sigma_i} = \frac{X_i - M_i}{\sigma_i}, \quad i = 1, 2, 3, 4$$

equations (30) and (31) become

$$A_{11}t_1 + A_{12}t_2 + A_{13}t_3 + A_{14}t_4 = 0 \tag{32}$$

or

$$\sum_{i=1}^{4} t_i A_{1i} = 0$$

Adopting as a measure of the accuracy of fit of (30), (31), or (32) to the given observed values the quantity

$$S_{1(234)} = \sqrt{\frac{\Sigma\rho^2}{N}}$$

after some rather tedious algebraic operations we find

$$S_{1(234)} = \sigma_1\sqrt{\frac{r_{11}D_{11} - r_{12}D_{12} + r_{13}D_{13} - r_{14}D_{14}}{D_{11}}} \tag{33}$$

$$= \sigma_1\sqrt{\frac{D}{D_{11}}} \tag{34}$$

Equation (33) may also be written

$$S_{1(234)} = \sigma_1\sqrt{1 - R^2_{1(234)}} \tag{35}$$

where

$$R_{1(234)} = \sqrt{\frac{r_{12}D_{12} - r_{13}D_{13} + r_{14}D_{14}}{D_{11}}}$$

$$= \sqrt{1 - \frac{D}{D_{11}}} \tag{36}$$

Defining $r_{12.34}$, the partial coefficient of correlation between X_1 and X_2 when the variables X_3 and X_4 are held fixed, by the equation

$$r_{12.34} = \pm\sqrt{(b_{12.34})(b_{21.34})}$$

we immediately obtain

$$r_{12.34} = \pm\frac{D_{12}}{\sqrt{D_{11}D_{22}}} = \pm\frac{A_{12}}{\sqrt{A_{11}A_{22}}}$$

Similarly

$$r_{13.24} = \pm\frac{D_{13}}{\sqrt{D_{11}D_{33}}} = \pm\frac{A_{13}}{\sqrt{A_{11}A_{33}}}$$

$$r_{14.23} = \pm\frac{D_{14}}{\sqrt{D_{11}D_{44}}} = \pm\frac{A_{14}}{\sqrt{A_{11}A_{44}}}$$

The signs of these values, $r_{12.34}$, $r_{13.24}$, etc. are the same as b_{12}, b_{13}, etc.

The following steps are recommended in the computation of the constants, assuming that the arithmetic means, the standard deviations, and the simple correlation coefficients have been computed.

(1) Write down D.

(2) Compute D_{11}, D_{22}, D_{33}, D_{44}.

(3) Compute D_{12}, D_{13}, D_{14}, D_{23}, D_{24}, D_{34}.
(4) Compute A_{12}, A_{13}, etc. from $A_{hk} = (-1)^{h+k}D_{hk}$.
(5) Compute the value of D from the formula

$$D = r_{11}A_{11} + r_{12}A_{12} + r_{13}A_{13} + r_{14}A_{14}$$

(6) Compute b_{12}, b_{13}, etc.
(7) Compute $r_{12.34}$, $r_{13.24}$, $r_{14.23}$.
(8) Compute $R_{1(234)}$ from equation (36).
(9) Compute $S_{1(234)}$ from equation (34).
(10) Write down the regression equation.

EXERCISES

1. The following table gives the fundamental constants obtained from the measurement of 450 eggs.[1]

		Length (mm.)	*Breadth (mm.)*	*Bulk (cc.)*	*Weight (gm.)*
Correlations	Length	1.0000	0.0837	0.5751	0.5797
	Breadth	0.0837	1.0000	0.8602	0.8357
	Bulk	0.5751	0.8602	1.0000	0.9804
	Weight	0.5797	0.8357	0.9804	1.0000
Arithmetic Means		56.3222	41.9167	51.8400	55.2400
Standard Deviations		2.3862	1.3777	4.2438	4.5923

(a) Find the regression of weight upon length and breadth.
(b) What is the estimated weight of an egg of the following measurements: length 56.03 mm., breadth 42.02 mm.?
(c) Find the regression equation of weight on length and bulk.
(d) Find the regression equation of weight on bulk and breadth.
(e) Find the standard errors of estimate for (a), (c), and (d).
(f) What is the best combination for estimating weight?

2. The data of the following table were secured from measurements of 450 freshmen at Syracuse University[2]:

X_1 = Academic success as measured by the number of honor points earned by the student during the first semester in college.

[1] Pearl and Surface: *A Biometrical Study of Egg Production in the Domestic Fowl*, Part III.

[2] May, Mark: Predicting Academic Success. *Journal of Educational Psychology*, Volume XIV, pp. 429–440.

X_2 = General intelligence based upon standardized tests.
X_3 = Industry and application as measured by the number of hours per week spent in study.
X_4 = Quality of preparatory work based upon average high school grade.

(a) Find the regression equation of X_1 on X_2, X_3, and X_4.
(b) Estimate X_1 when $X_2 = 108$, $X_3 = 32$, and $X_4 = 82$.
(c) Find $R_{1(234)}$ and $S_{1(234)}$.
(d) Find $r_{12.34}$, $r_{13.24}$, and $r_{14.23}$.

		X_1	X_2	X_3	X_4
Correlations	X_1	1.00	0.60	0.32	0.40
	X_2	0.60	1.00	− 0.35	0.36
	X_3	0.32	− 0.35	1.00	0.11
	X_4	0.40	0.36	0.11	1.00
M's		18.5	100.6	24	79
σ's		11.2	15.8	6	7.5

3. Show that $$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{(1 - r^2_{14.3})(1 - r^2_{24.3})}}$$

4. By permuting the subscripts in number 3 preceding, write down the values of $r_{13.24}$ and $r_{14.23}$.

5. In the following table the values are monthly averages.

X_1 = Wholesale price of butter, 92 score, in ¢ per lb.
X_2 = Apparent consumption, millions of pounds.
X_3 = Factory production, millions of pounds.
X_4 = Stocks in cold storage at end of month, millions of pounds.

FACTORS AFFECTING WHOLESALE PRICE OF CREAMERY BUTTER

Year	X_1	X_2	X_3	X_4	*Year*	X_1	X_2	X_3	X_4
1919	61	68	72	67	1929	45	130	133	82
1920	61	73	72	60	1930	37	134	133	83
1921	43	90	88	53	1931	28	142	139	55
1922	41	98	96	51	1932	21	142	141	50
1923	47	106	103	47	1933	22	139	147	92
1924	43	111	113	74	1934	26	147	141	69
1925	45	115	114	62	1935	30	138	136	71
1926	44	123	121	68	1936	33	135	136	60
1927	47	124	125	71	1937	34	138	135	64
1928	47	124	124	62	1938	28	142	149	111

(1) Find the values:

$M_1 =$ $M_2 =$ $M_3 =$ $M_4 =$
$\sigma_1 =$ $\sigma_2 =$ $\sigma_3 =$ $\sigma_4 =$
$r_{12} =$ $r_{13} =$ $r_{14} =$
$r_{23} =$ $r_{24} =$ $r_{34} =$

(2) Find the equation of the regression hyperplane with X_1 dependent upon X_2, X_3, X_4.

(3) Find $R_{1(234)}$ and $S_{1(234)}$.

78. THE CASE OF n VARIABLES

We shall give a summary of the results of multiple regression for the case of n variables leaving all of the details to be carried out by the student.

Let
$$X_1 = b_{12}X_2 + b_{13}X_3 + \cdots + b_{1n}X_n + c \tag{37}$$
be the equation which gives the best value to X_1 for any set of values of $X_2, X_3, \ldots, X_n$.

If the regression coefficients b_{12}, b_{13}, etc., and c are determined so that the sum of the squares of the X_1-residuals is a minimum, the point $(M_1, M_2, \ldots, M_n)$ is on the hyperplane (37). Transferring our data to this centroidal point as origin, the regression equation becomes
$$x_1 = b_{12}x_2 + b_{13}x_3 + b_{14}x_4 + \cdots + b_{1n}x_n \tag{38}$$
where
$$x_i = X_i - M_i, \quad i = 1, 2, 3, \ldots, n$$

Based upon the principle of least squares, the normal equations for the determination of the regression coefficients are

$$\left.\begin{aligned}
b_{12}\sigma_2 + b_{13}\sigma_3 r_{23} + \cdots + b_{1n}\sigma_n r_{2n} &= \sigma_1 r_{12} \\
b_{12}\sigma_2 r_{23} + b_{13}\sigma_3 + \cdots + b_{1n}\sigma_n r_{3n} &= \sigma_1 r_{13} \\
\cdots\cdots\cdots\cdots\cdots\cdots &= \cdots \\
\cdots\cdots\cdots\cdots\cdots\cdots &= \cdots \\
b_{12}\sigma_2 r_{2n} + b_{13}\sigma_3 r_{2n} + \cdots + b_{1n}\sigma_n &= \sigma_1 r_{1n}
\end{aligned}\right\} \tag{39}$$

Defining the major determinant D by the equation

$$D = \begin{vmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & r_{nn} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{vmatrix} \tag{40}$$

we find

$$\left.\begin{aligned} b_{12} &= \frac{\sigma_1 D_{12}}{\sigma_2 D_{11}} = -\frac{\sigma_1 A_{12}}{\sigma_2 A_{11}} \\ b_{13} &= -\frac{\sigma_1 D_{13}}{\sigma_3 D_{11}} = -\frac{\sigma_1 A_{13}}{\sigma_3 A_{11}} \\ \ldots &= \ldots\ldots\ldots = \ldots\ldots\ldots \\ b_{1n} &= (-1)^n \frac{\sigma_1 D_{1n}}{\sigma_n D_{11}} = -\frac{\sigma_1 A_{1n}}{\sigma_n A_{11}} \end{aligned}\right\} \tag{41}$$

where D_{hk} is the minor and A_{hk} is the co-factor of r_{hk},

$$A_{hk} = (-1)^{h+k} D_{hk}$$

The regression equation for determining the best value of x_1 for given values of $x_2, x_3, \ldots, x_n$, is

$$\frac{x_1}{\sigma_1} A_{11} + \frac{x_2}{\sigma_2} A_{12} + \ldots + \frac{x_n}{\sigma_n} A_{1n} = 0 \tag{42}$$

In terms of the original variates the equation of regression is

$$\frac{(X_1 - M_1)}{\sigma_1} A_{11} + \frac{(X_2 - M_2)}{\sigma_2} A_{12} + \cdots + \frac{(X_n - M_n)}{\sigma_n} A_{1n} = 0 \tag{43}$$

Equations (42) and (43) may be written

$$\sum \frac{x_i}{\sigma_i} A_{1i} = \sum \frac{(X_i - M_i)}{\sigma_i} A_{1i} = \Sigma t_i A_{1i} = 0 \tag{44}$$

where

$$t_i = \frac{x_i}{\sigma_i} = \frac{X_i - M_i}{\sigma_i}, \quad i = 1, 2, 3, \ldots, n$$

Adopting as a measure of the goodness of fit of (42) to the given data the quantity

$$S_{1(23\cdots n)} = \sqrt{\frac{\Sigma \rho^2}{N}}$$

where $\rho = x_1 - (b_{12}x_2 + b_{13}x_3 + \cdots + b_{1n}x_n)$, and the values of b_{1i}, $i = 1, 2, 3, \ldots, n$, are given by (41), we find

$$\begin{aligned} S_{1(23\cdots n)} &= \sigma_1 \sqrt{\frac{D}{D_{11}}} \\ &= \sqrt{1 - R^2_{1(23\cdots n)}} \end{aligned} \tag{45}$$

where

$$R_{1(23\cdots n)} = \sqrt{1 - \frac{D}{D_{11}}} \tag{46}$$

Defining the partial coefficient of correlation $r_{1k.23\cdots n}$ by the equation

$$r_{1k.23\cdots n} = \pm \sqrt{b_{1k.23\cdots n} \cdot b_{k1.23\cdots n}}$$

we find

$$r_{1k.23\cdots n} = \pm \frac{D_{1k}}{\sqrt{D_{11}D_{kk}}} = \pm \frac{A_{1k}}{\sqrt{A_{11}A_{kk}}} \tag{47}$$

The sign of $r_{hk.ab\cdots n}$ is the same as that of b_{hk}.

Chapter 10

NONLINEAR TRENDS: CURVE-FITTING

79. INTRODUCTION

The investigator in any branch of science is frequently confronted with quantitative data which, when plotted, seem to lie near a smooth curve and hence to obey, approximately at least, some mathematical law. Thus the following table gives the area Y (in square centimeters) of a wound at the end of X days.

FIGURE 41

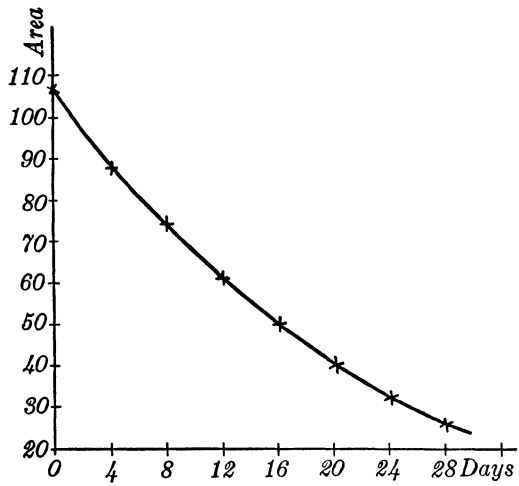


TABLE 65

Number of Days X	*Area of Wound* Y
0	107.0
4	88.0
8	74.2
12	61.8
16	51.0
20	41.6
24	33.6
28	26.9

The fact that these data when plotted [Figure 41] lie very near a smooth curve leads us to suspect that they can be represented, approximately, by the equation of a curve. Such an equation, whose form is inferred from the results of experiment or observation and whose constants are determined from experimental or observational data, is known as an *empirical equation.* The empirical equation, once it is derived, is a summarizing expression for the observed data, and it may be used to obtain a good approximation to the value of

the true ordinate for a given abscissa *within the range of values used in its determination.*

The problem of determining the type of equation to be used is an indeterminate one, for a number of curves can be drawn to pass very near the plotted points and hence a number of equations can be found to represent the data approximately. The choice of the proper mathematical function depends a great deal upon the investigator's knowledge of the properties of curves and his experience in curve-fitting. Fortunately, there are a number of simple tests that may be employed to enable us to make an intelligent choice of the type of equation to be used. Of course one can select an equation in which the number of undetermined constants equals the number of the observations and thus have the resulting curve pass through the observed points exactly, but this process emphasizes the minor fluctuations that represent simply errors of observation and renders impossible the discovery of a simple law. A better procedure is to select a simple type of function involving only a few constants and thus allow for fluctuations due to sampling.

Having chosen a particular type of function with which to graduate the data, our specific question is: *How can the constants of the equation be determined in order to obtain the curve of that type of best fit?* The method employed depends upon the desired degree of accuracy. We may employ one or more of four methods: (1) the *method of selected points*, (2) the *method of averages*, (3) the *method of least squares*, or (4) the *method of moments*. Of these methods the first is the simplest; the second requires more computation than the first but usually gives better results; the third requires considerable computation but gives the best results and a *unique* answer to our question; the fourth gives a unique answer that is identical to that obtained by the third for polynomial functions.

80. THE PROCESS OF DIFFERENCING

In the preceding section we alluded to certain simple tests that may be employed to assist us in choosing the appropriate type of equation to represent our data. Inasmuch as these tests will frequently be stated in the language of *differences*, it may be well that we digress at this point from our general problem to learn the rudiments of this language. Consider the following table:

TABLE 66

X (1)	Y (2)	ΔY (3)	$\Delta^2 Y$ (4)	$\Delta^3 Y$ (5)	$\Delta^4 Y$ (6)
0	1				
		3			
1	4		3		
		6		1	
2	10		4		0
		10		1	
3	20		5		0
		15		1	
4	35		6		0
		21		1	
5	56		7		
		28			
6	84				

Corresponding values of X and Y, where Y is some undetermined function of X, are given in columns (1) and (2). In column (3), headed ΔY, we have the *first differences* of Y. Any value of ΔY is found by subtracting a value of Y from its successor. Thus, $3 = 4 - 1$, $6 = 10 - 4$, etc. Similarly column (4), headed $\Delta^2 Y$, is obtained by subtracting each ΔY from its successor. These values are called the *second differences* of Y. Other differences are found in a similar manner. In the table we are considering it may be noted that the values of $\Delta^2 Y$ are in arithmetic progression, those of $\Delta^3 Y$ are constant and hence all higher differences are zero.

EXERCISE

Begin at the right-hand side of Table 66, work back to the left and show that when $X = 7$, $Y = 120$.

The values of X may differ by amounts other than unity. In general we may indicate by ΔX the difference in X. When the difference in successive X's is the same — that is, when ΔX is constant — and Y is a function of X:

$$\Delta Y_x = Y_{x+\Delta x} - Y_x \tag{1}$$

In the following table, where again Y is an undetermined function of X, we have, for example, $\Delta X = 2$.

TABLE 67

X	Y	ΔY	$\Delta^2 Y$	$\Delta^3 Y$
0	0			
		2		
2	2		4	
		6		0
4	8		4	
		10		0
6	18		4	
		14		
8	32			

In experimental data involving two variables, the independent variable is usually subject to the control of the experimenter, and the values of the independent variable are frequently given in arithmetic progression. That is, if X is the independent variable, ΔX is frequently constant. We shall see that this precaution on the part of the experimenter may greatly simplify the discovery of an appropriate equation.

Consider the straight line:

$$Y = mX + b$$

We have from (1):

$$\Delta Y = m(X + \Delta X) + b - (mX + b)$$
$$\Delta Y = m \cdot \Delta X$$

From this result it is seen that if ΔX is constant, ΔY is also constant; further, $\Delta Y/\Delta X$ is constant (compare Section 57, p. 204).

Consider now the parabola:

$$Y = aX^2 + bX + c$$

Applying (1):

$$\Delta Y = a(X + \Delta X)^2 + b(X + \Delta X) + c - (aX^2 + bX + c)$$
$$\Delta Y = 2aX\Delta X + b\Delta X + a(\Delta X)^2$$
$$\Delta(\Delta Y) = \Delta^2 Y = 2a(X + \Delta X)\Delta X + b\Delta X + a(\Delta X)^2 - 2aX\Delta X - b\Delta X - a(\Delta X)^2$$
$$\Delta^2 Y = 2a(\Delta X)^2$$

From this result we see that if ΔX is constant, the second difference of the polynomial $aX^2 + bX + c$ is also constant.

One may continue this process and show that *if ΔX is constant, the nth difference of a polynomial of the nth degree is also constant.*

The converse of this theorem is also true, namely:

If for a constant ΔX, $\Delta^n Y$ is also constant, then Y is a polynomial in X of degree n.[1]

The nth differences of the values of Y obtained from observational data are seldom constant. If, however, the nth differences of Y are approximately constant, ΔX being constant, we can represent the data approximately by:

$$Y = aX^n + bX^{n-1} + \cdots + k$$

EXERCISES

1. If $Y = c$, show that $\Delta Y = 0$.

2. If $Y = X^3$, show that $\Delta^3 Y = 6(\Delta X)^3$.

3. Prepare a table for the function $Y = 2X^2 - 3X + 4$ for $X = 0, 1, 2, 3, 4$ and find the second differences from the table.

4. Prepare a table for the function $Y = X^3 - X^2 + 8X + 2$ for $X = 1, 3, 5, 7, 9$ and find the third differences from the table.

5. In the following table, ΔX is constant $(= 1)$ and $\Delta^2 Y$ is constant $(= 2)$. Hence Y is a quadratic function of X:

$$Y = aX^2 + bX + c$$

Find the values of a, b, and c. Find Y when $X = 5$.

Hint:

$$\Delta^2 Y = 2 = 2a(\Delta X)^2 = 2a(1)^2 = 2a$$

X	Y	ΔY	$\Delta^2 Y$
0	2		
		-1	
1	1		2
		1	
2	2		2
		3	
3	5		2
		5	
4	10		

[1] For a proof, see T. R. Running, *Empirical Formulas*, p. 18.

X	Y	ΔY	$\Delta^2 Y$	$\Delta^3 Y$
0	0			
1	−1			
2	4			
3	21			
4	56			

6. Complete the accompanying table. Find the function that represents the data. Find Y when $X = 5$.

7. Prove that if a sequence of numbers is in geometric progression, their logarithms are in arithmetic progression.

8. Prove that if a sequence of numbers is in geometric progression, their first differences are in geometric progression.

81. FITTING A STRAIGHT LINE TO OBSERVED DATA

A large portion of Chapter 7 was devoted to the problem of fitting a straight line to observed data by the method of least squares. We desired at that time to emphasize the method of least squares because we were then interested in finding a unique line for which we could secure a test for the goodness of fit and thus arrive at the Bravais-Pearson cross-product coefficient of correlation. Since one may frequently not desire so accurate a solution as is given by the method of least squares — especially at the price of tedious computation one must pay to secure it — we shall discuss two other less accurate methods.

A. The Method of Selected Points. To apply this method we must plot the observed data carefully. We then draw a straight line among the points which will pass as near as possible to each of them. Since the straight-line equation

$$Y = mX + b$$

has two undetermined constants, m and b, we must obtain two equations with m and b as unknowns from which to determine them. If the line happens to pass through two of the plotted points or through any other two points whose coördinates can be determined approximately, we can substitute their coördinates in the given equation and solve the two resulting equations for m and b. In any case the points so used should be as far apart as possible.

Consider again the temperature-resistance data of Table 42, to which we have previously given attention in Section 59 (p. 211).

FIGURE 42

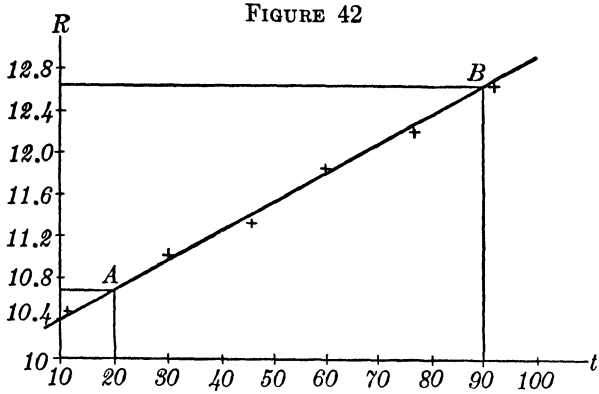


TABLE 68

t	R
10.5	10.42
29.5	10.94
42.7	11.32
60.0	11.80
75.5	12.24
91.1	12.67

These data, when plotted, present six points that may seem to lie upon a straight line. Let us seek further evidence by applying the test for straight-line data. We have learned in the preceding section that if $\Delta Y/\Delta X$ is constant the data can be fitted to a straight-line equation. In Table 69 we have computed the several values of

TABLE 69

Δt	t	R	ΔR	$\Delta R/\Delta t$
	10.5	10.42		
19.0			0.52	0.0274
	29.5	10.94		
13.2			0.38	0.0289
	42.7	11.32		
17.3			0.48	0.0277
	60.0	11.80		
15.5			0.44	0.0284
	75.5	12.24		
15.6			0.43	0.0276
	91.1	12.67		
			Mean = 0.0280	

$\Delta R/\Delta t$. Since they are approximately constant we are justified in concluding that the data may be fitted approximately to a straight-line equation:

$$R = mt + b$$

The line we have drawn does not pass through any of the given points. However, it seems to pass through the points $A(20, 10.7)$ and $B(90, 12.6)$ whose ordinates we have estimated from the graph. Substituting in the given equation the coördinates of the points we have:

$$b + 20m = 10.7$$
$$b + 90m = 12.6$$

from which we obtain

$$m = 0.027 \qquad b = 10.157$$

Hence the required relation is:

$$R = 0.027t + 10.157$$

The least-square solution (Exercise 1 on p. 220) gives:

$$R = 0.02799t + 10.122$$

EXERCISE

Assume the line passes through the first and last points, (10.5, 10.42) and (91.1, 12.67), and find its equation.

It will be noted that the arithmetic mean of the values of $\Delta R/\Delta t$ in Table 69 is 0.0280. How may this be used in finding an equation for a line fitting our data approximately?

If we take this average slope as the slope of our required line we have:

$$R = 0.0280t + b$$

We can now substitute the coördinates of each of the six given points and thus determine six values of b. Their mean may be taken as *the* value of b for the required line. We shall leave the computation as an exercise for the student. He should receive for an answer:

$$R = 0.0280t + 10.1216$$

B. The Method of Averages. The fundamental principle of the method of averages is that an empirical curve of given type best fitting a given group of points is one for which the algebraic sum of the residuals is zero. (It will be recalled that this criterion was satisfied by the line determined by the method of least squares.[1]) From Section 59 (p. 210), if ρ_i is any residual:

$$\rho_i = Y_i - mX_i - b$$

and

$$\Sigma\rho_i = \Sigma Y_i - m\Sigma X_i - nb$$

[1] See Exercise 5 on p. 221.

Since the sum of the residuals is zero we have:

$$m\Sigma X + nb = \Sigma Y$$

In order to obtain two equations which may be solved for the unknowns, m and b, we divide our data, Table 68 (p. 312), into two groups each containing three sets of data. For the first group we choose the first three sets of data for which $\Sigma t = 82.7$, $n = 3$, $\Sigma R = 32.68$, and for the second group the remaining three sets of data for which $\Sigma t = 226.6$, $n = 3$, $\Sigma R = 36.71$. We then have the equations:

$$82.7m + 3b = 32.68$$
$$226.6m + 3b = 36.71$$

from which we obtain

$$m = 0.0280 \qquad b = 10.121$$

Hence the required relation is:

$$R = 0.0280t + 10.121$$

C. The Method of Least Squares. Curve-fitting by the method of least squares is based upon the principle that the empirical curve of a given type best fitting a given set of points is that one in which the constants are so determined that they will make the sum of the squares of the residuals a minimum. Since the squares of the residuals are positive quantities, the requirement that their sum shall be a minimum gives assurance that the numerical values of the residuals will be such that the best-fitting curve will pass as close as possible to all the points.

Inasmuch as Section 59 (p. 210) was devoted to the problem of fitting the line

$$Y = mX + b$$

to a set of points by the method of least squares, we shall merely recapitulate here the findings of that section. By minimizing

$$\Sigma\rho_i^2 = \Sigma(Y_i - mX_i - b)^2$$

where ρ_i is the Y-residual of the ith point, we obtain the normal equations

$$m\Sigma X + nb = \Sigma Y$$
$$m\Sigma X^2 + b\Sigma X = \Sigma XY$$

which, when solved, gave:

$$m = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

$$b = \frac{\Sigma X^2\Sigma Y - \Sigma X\Sigma XY}{n\Sigma X^2 - (\Sigma X)^2}$$

EXERCISES

1. Show that the point (M_X, M_Y) is on a line determined by the method of averages.

2. The following table gives the population of France at each census from 1806 to 1866. Determine by the method of averages a straight line well adapted to the data, choosing $X = 0$ at 1836.

POPULATION OF FRANCE, 1806–1866

Year	*Population (millions)*	*Year*	*Population (millions)*
1806	29.11	1851	35.78
1821	30.46	1856	36.04
1831	32.57	1861	37.39
1836	33.54	1866	38.07
1846	35.40		

3. Find by the method of least squares the equation of the best-fitting straight line to the data of the following table. What are the predicted net earnings, based upon this line, for the year 1929? The actual net earnings were 48.5 millions.

ANNUAL EARNINGS OF THE ASSOCIATED GAS AND ELECTRIC SYSTEM, 1920–1928 [1]

Year	*Net Earnings (millions of dollars)*	*Year*	*Net Earnings (millions of dollars)*
1920	13.4	1925	29.5
1921	16.2	1926	33.5
1922	19.2	1927	37.8
1923	22.7	1928	40.6
1924	25.1	1929	....

[1] The data are from *Time*, Jan. 27, 1930.

82. THE EXPONENTIAL FUNCTION $Y = ab^X$

Excepting the linear function, probably no expression with two undetermined constants is more useful in characterizing observed data than the exponential function $Y = ab^X$. It may be described as that function whose rate of change is proportional to the value of the function. The rate of change may be positive or negative, that is, Y may increase with X or Y may decrease as X increases. Because the accumulated amount of a sum of money placed at compound interest at a given rate for a given time is expressed by this function, it is known as the *compound interest law*. Thus, if \$100 is placed at compound interest for X years at 5 per cent the accumulated amount Y is given by:

$$Y = 100(1.05)^X$$

We represent this function graphically.

FIGURE 43

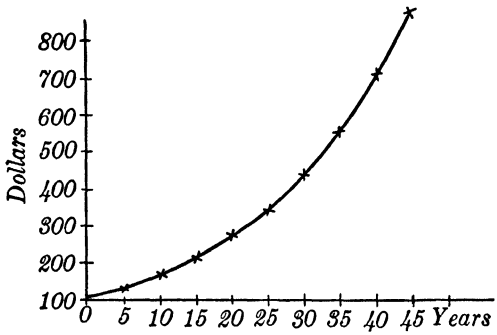


TABLE 70

X	Y
0	100.00
5	127.63
10	162.89
15	207.89
20	265.33
25	338.64
30	432.19
35	551.60
40	704.00
45	898.50

The exponential function is also called the *law of organic growth* because many biological phenomena obey closely this law of growth. For examples, a culture of bacteria, or populations of mice, of rabbits, of human beings, when placed in environments conducive to growth, will increase for a time in approximate accordance with this law.

The exponential law is applicable to many other types of data. Many data from the commercial and the economic fields show exponential trends. We find the law especially applicable to data on

production and to data on the periodic earnings of many industrial organizations.

A simple test for the exponential function is contained in the following:

Theorem I. *If the values of X are in arithmetic progression and the corresponding values of Y are in geometric progression, the relation between the variables is expressed by the equation:*

$$Y = ab^X$$

TABLE 71

X	Y
X_1	Y_1
$X_2 = X_1 + \Delta X$	$Y_2 = rY_1$
$X_3 = X_1 + 2\Delta X$	$Y_3 = r^2Y_1$
. . . . .	. . .
. . . . .	. . .
$X_n = X_1 + (n-1)\Delta X$	$Y_n = r^{n-1}Y_1$

From the hypothesis we have the data as shown in the accompanying table. Since

$$X_n = X_1 + (n-1)\Delta X$$

we have:

$$n - 1 = \frac{X_n - X_1}{\Delta X}$$

and hence

$$Y_n = Y_1 r^{\frac{X_n - X_1}{\Delta X}}$$

$$= Y_1 r^{\frac{-X_1}{\Delta X}} \left(r^{\frac{1}{\Delta X}}\right)^{X_n}$$

or

$$Y_n = ab^{X_n}$$

when

$$a = Y_1 r^{\frac{-X_1}{\Delta X}} \quad \text{and} \quad b = r^{\frac{1}{\Delta X}}$$

That is, any X is connected with the corresponding Y by the relation:

$$Y = ab^X \tag{2}$$

Illustrative Problem 1. Consider Table 72, which gives the population of the United States at each ten-year census from 1800 to 1890.

TABLE 72. POPULATION OF THE UNITED STATES, 1800–1890

	Year	*Population (millions)*	*Ratio of Each Population to the One Above*
t	X	P	
0	1800	5.3	
1	1810	7.2	1.36
2	1820	9.6	1.33
3	1830	12.9	1.34
4	1840	17.1	1.33
5	1850	23.2	1.36
6	1860	31.4	1.35
7	1870	38.6	1.23
8	1880	50.2	1.30
9	1890	63.0	1.25
		Mean	1.3167

Let $t = (X - 1800)/10$.

Here we note that the values of X (and t) are in arithmetical progression and that the values of P are approximately in a geometric progression since the ratio of any population to the one preceding is approximately constant. Hence we may assume that the data follow approximately the exponential law: $P_t = ab^t$.

If we assume that the point $t = 0$, $P = 5.3$ is on the curve, and that the decade rate of increase is the arithmetic mean of the ratios, we can immediately obtain a first approximation formula:

$$P_o = 5.3 = ab^o = a$$
$$a = 5.3$$

Since by definition

$$P_t = a(1.3167)^t = ab^t$$

we have:

$$b = 1.3167$$

Hence

$$P_t = 5.3(1.3167)^t$$

may be considered a first or crude approximation. By assigning $t = 0, 1, 2, 3, \ldots, 9$, the computed values of P_t, which can be compared with the observed values, can be found.

For a closer approximation we proceed as follows. From

$$P = ab^t$$

we have:

$$\log P = (\log b)t + \log a$$

Now let $Y = \log P, \quad m = \log b, \quad k = \log a.$

We then have:

$$Y = mt + k$$

which is a straight line. Therefore we can fit the curve $P = ab^t$ to the given data by fitting the line $Y = mt + k$ to the corresponding $(t, Y = \log P)$ data. We shall do this by the method of least squares.

From Sections 59 (p. 218) and 81 (p. 315) we have:

$$m = \frac{n\Sigma tY - \Sigma t\Sigma Y}{n\Sigma t^2 - (\Sigma t)^2} = \frac{n\Sigma t \log P - \Sigma t\Sigma \log P}{n\Sigma t^2 - (\Sigma t)^2}$$

$$k = \frac{\Sigma t^2\Sigma Y - \Sigma t\Sigma tY}{n\Sigma t^2 - (\Sigma t)^2} = \frac{\Sigma t^2\Sigma \log P - \Sigma t\Sigma t \log P}{n\Sigma t^2 - (\Sigma t)^2}$$

We shall use the following form with eight-place logarithms to assist in finding m and k.

TABLE 73

t	P	$\log P$	t^2	$t \log P$	*Computed* P
0	5.3	0.7242759	0	00.0000000	5.5
1	7.2	0.8573325	1	00.8573325	7.3
2	9.6	0.9822712	4	1.9645424	9.6
3	12.9	1.1105897	9	3.3317691	12.7
4	17.1	1.2329961	16	4.9319844	16.8
5	23.2	1.3654880	25	6.8274400	22.2
6	31.4	1.4969296	36	8.9815776	29.3
7	38.6	1.5865873	49	11.1061111	38.6
8	50.2	1.7007037	64	13.6056296	51.0
9	63.0	1.7993405	81	16.1940645	67.3
45		12.8565145	285	67.8004512	

$$m = \log b = \frac{10(67.8004512) - 45(12.8565145)}{10(285) - (45)^2} = 0.1205592$$

$$b = 1.319967$$

$$k = \log a = \frac{285(12.8565145) - 45(67.8004512)}{10(285) - (45)^2}$$

$$\log a = 0.7431349$$
$$a = 5.535186$$

Hence our law is:

$$P = 5.535186(1.319967)^t$$

or

$$\log P = 0.1205592t + 0.7431349$$

By assigning $t = 0, 1, 2, \ldots, 9$ we obtain the computed values of P which are found in Table 73.

We can use this formula to predict the populations in 1900, 1910, 1920 by assigning $t = 10, 11, 12$. We find the predicted populations to be 88.9, 117.3, and 154.8, whereas the actual populations were 76.0, 92.0, and 105.7. This shows that an empirical formula must be used with caution for values outside the given abscissal range. In this particular case the exponential law ceased to operate after 1870 and we began then to approach the point of saturation.

We shall leave it as an exercise for the student to find the law for the population based upon the method of averages.

It frequently happens that the data are not given with the values of the independent variable in arithmetic progression and hence the test of Theorem I will not apply. In such cases we can use the following:

Theorem II. *If the variables X and Y are so related that $\Delta \log Y/\Delta X$ is constant, then the relation between them can be expressed by the formula:*

$$Y = ab^X$$

Since by hypothesis

$$\frac{\Delta \log Y}{\Delta X} = m,$$

we have by Section 80 (p. 310):

$$\log Y = mX + k$$

or, if $k = \log a$,

$$\log Y = mX + \log a$$
$$\log Y - \log a = mX$$
$$\log (Y/a) = mX$$
$$Y/a = 10^{mX} = (10^m)^X = b^X$$

or

$$Y = ab^X$$

where $10^m = b$.

We shall apply this theorem to

Illustrative Problem 2. The following table shows the amount A of a substance remaining in a reacting chemical system at the expiration of a given time t (Harcourt and Esson).

TABLE 74

t	A	$\log A$	Δt	$\Delta \log A$	$\frac{\Delta \log A}{\Delta t}$
2	94.8	1.9768083			
			3	− 0.0328194	− 0.0109
5	87.9	1.9439889			
			3	− 0.0338984	− 0.0113
8	81.3	1.9100905			
			3	− 0.0356087	− 0.0119
11	74.9	1.8744818			
			3	− 0.0375251	− 0.0125
14	68.7	1.8369567			
			3	− 0.0307767	− 0.0103
17	64.0	1.8061800			
			10	− 0.1133331	− 0.0113
27	49.3	1.6928469			
			4	− 0.0493942	− 0.0123
31	44.0	1.6434527			
			4	− 0.0512759	− 0.0128
35	39.1	1.5921768			
			9	− 0.0924897	− 0.0103
44	31.6	1.4996871			

The values of $\Delta \log A/\Delta t$ are fairly constant and we conclude therefore that the data may be represented approximately by

$$A = ab^t$$

or

$$\log A = (\log b)t + \log a$$

or

$$Y = mt + k$$

when $Y = \log A$, $m = \log b$, and $k = \log a$.

We shall use the method of averages to determine the constants. Dividing the data into two groups, the first five sets of data for the first group and the remaining five sets of data for the second group, we obtain:

$$\Sigma Y = \Sigma \log A = 9.5423262, \quad \Sigma t = 40, \quad n = 5$$
$$\Sigma Y = \Sigma \log A = 8.2343435, \quad \Sigma t = 154, \quad n = 5$$

Recalling that in the method of averages the sum of the residuals is zero; that is,

$$\Sigma(Y - mt - k) = 0$$

or

$$\Sigma Y = m\Sigma t + nk$$

we have upon substituting the above values:

$$5k + 40m = 9.5423262$$
$$5k + 154m = 8.2343435$$

Solving, we obtain:

$$m = \log b = -0.0114735$$
$$k = \log a = 2.0002532$$
$$b = 0.973927$$
$$a = 100.0586$$

Hence the law is:

$$A = 100.0586(0.973927)^t$$

or

$$\log A = -0.0114735t + 2.0002532$$

By substituting the given values of t we obtain the computed values of $\log A$ from which we obtain the computed values of A which are shown in the following table.

t	*Observed* A	*Computed* $\log A$	*Computed* A	*Residuals*
2	94.8	1.9773062	94.9	− 0.1
5	87.9	1.9428857	87.7	0.2
8	81.3	1.9084652	81.0	0.3
11	74.9	1.8740447	74.8	0.1
14	68.7	1.8396242	69.1	− 0.4
17	64.0	1.8052037	63.9	0.1
27	49.3	1.6904687	49.0	0.3
31	44.0	1.6445747	44.1	− 0.1
35	39.1	1.5986807	39.7	− 0.6
44	31.6	1.4954192	31.3	0.3

Exercise. Solve this problem by method of averages using four-place logarithms.

EXERCISES [1]

1. Show that an exponential curve may give a satisfactory fit for the data of Table 65 (p. 306). Fit an exponential curve to these data and estimate from the equation the values of Y when $X = 32$ and when $X = 36$. Compare these results with the actual values: $Y = 21.3$ when $X = 32$, and $Y = 16.8$ when $X = 36$.

Answer: Method of least squares gives $Y = 108.8035(0.952348)^X$.

2. In the following table p is the barometric pressure in inches of a column of mercury at distance h in feet above the sea level. Show that an exponential curve, $p = ab^h$, may be appropriately applied to these data. Find the equation of the best-fitting curve and the values of p when h = 1,000 ft., 2,000 ft., 5,000 ft.

h	0	886	2,753	4,763	6,942	10,593
p	30	29	27	25	23	20

3. The following table exhibits the values of the temperature T reached by a cooling body at the expiration of various times t. Determine the best-fitting curve of the type $T = ab^t$ for the data of this table.

t	0	3.79	11.93	21.23	31.68	44.11	59.12
T	17.9	17.0	15.2	13.4	11.6	9.8	8.0

4. Fit an exponential curve to the data of Exercise 1, page 91.

5. The following observations were made on a growing plant. The time is reckoned in days from the first observation. What is the law of growth?

Days	0	1	2	3	4	5	6	7	8
Height (inches)	0.75	1.20	1.75	2.50	3.45	4.70	6.20	8.25	11.50

83. THE POWER FUNCTION $Y = aX^b$

In the preceding sections of this chapter we have dealt with the problems which involved fitting the linear function $Y = aX + b$ and the exponential function $Y = ab^X$ to observed data. A third function with two undetermined constants, the power function $Y = aX^b$, finds frequent application. Owing to the fact that the constants can be determined approximately by *rectifying* the curve, that is, by transforming it into a straight-line equation — as was done

[1] We leave it to the discretion of the teacher to suggest the method that is to be used in solving these exercises.

with the exponential function — the power curve is not difficult to employ.

The power function is parabolic in form when b is positive and hyperbolic if b is negative. The parabolic curves all pass through the points (0, 0) and (1, a) and also enjoy the property that Y increases with X. The hyperbolic curves all pass through the point (1, a), have the coördinate axes as asymptotes, and enjoy the property that Y decreases as X increases.

EXERCISE

Plot on the same coördinate axes for $X > 0$ the curves:

a. $Y = 2X^2$
b. $Y = 2X$
c. $Y = 2X^{0.5}$
d. $Y = 2X^{-2}$
e. $Y = 2X^{-1}$
f. $Y = 2X^{-0.5}$

A simple test — not always applicable — for determining if the power function is applicable is contained in:

Theorem I. *If the values of X are in geometrical progression and the corresponding values of Y are also in geometrical progression, then the relation between the variables is expressed by the formula:*

$$Y = aX^b$$

TABLE 75

X	Y
X_1	Y_1
$X_2 = rX_1$	$Y_2 = RY_1$
$X_3 = r^2X_1$	$Y_3 = R^2Y_1$
. . .	. . .
. . .	. . .
$X_n = r^{n-1}X_1$	$Y_n = R^{n-1}Y_1$

From the hypothesis we have the data as in Table 75. Since

$$X_n = r^{n-1}X_1 \quad \text{and} \quad Y_n = R^{n-1}Y_1,$$

we have, applying logarithms:

$$n - 1 = \frac{\log X_n - \log X_1}{\log r}$$

and

$$n - 1 = \frac{\log Y_n - \log Y_1}{\log R}$$

Equating these values of $(n-1)$ and writing $\log R/\log r = b$, we have

$$\frac{\log Y_n - \log Y_1}{\log X_n - \log X_1} = b,$$

that is:

$$\log Y_n/Y_1 = \log (X_n/X_1)^b$$

or

$$Y_n = (Y_1/X_1^b)X_n^b = aX_n^b$$

where

$$b = \log R/\log r \qquad \text{and} \qquad a = Y_1/X_1^b$$

That is, for any set of corresponding values we have:

$$Y = aX^b$$

There is an evident practical difficulty with this beautiful theorem. There is rarely any reason, or even an opportunity, for the observer to gather his data with the values of one variable in geometric progression and thus make possible a test to determine if the other variable is also in geometric progression. In general the observer has no predilections as to the law; he gathers the data and may hope to discover a law. Very frequently, however, the careful observer will, if possible, secure data with the independent variable ordered in some definite manner, most frequently in arithmetic progression.

When Theorem I may not be applicable we may be able to use the following:

Theorem II. *If the values of* X *and* Y *are so related that* $\Delta \log Y/\Delta \log X$ *is constant, then the relation between the variables is expressed by:*

$$Y = aX^b$$

Since

$$\frac{\Delta \log Y}{\Delta \log X} = b, \text{ a constant,}$$

we have by Section 80 (p. 310):

$$\log Y = b \log X + c$$
$$\log Y = \log X^b + \log a \qquad (\text{if } c = \log a)$$
$$\log Y = \log aX^b$$

or

$$Y = aX^b$$

Consider Table 76, which shows the currents, i, in amperes passing through an 118-volt tungsten lamp for various terminal voltages, e.

TABLE 76

e	i	*Ratio of Any i to Preceding*
2	0.0245	
4	0.0370	1.51
8	0.0570	1.54
16	0.0855	1.50
32	0.1295	1.51
64	0.2000	1.54
128	0.3035	1.53

We note that the independent variable, e, is given in a geometric progression. We find, as the table shows, that the corresponding values of i are also essentially in geometric progression. Therefore the data follow the law:

$$i = ae^b \tag{4}$$

Since

$$\log i = b(\log e) + \log a \tag{5}$$

if we let $Y = \log i, \quad X = \log e, \quad k = \log a$

we have:

$$Y = bX + k \tag{6}$$

which is a straight line. Therefore we may approximately fit the curve (4), $i = ae^b$, to the given data by fitting the line (5), $Y = bX + k$, to the corresponding ($X = \log e$, $Y = \log i$) data.

We shall first use the method of averages.

TABLE 77

e	i	$\log e = X$	$\log i = Y$
2	0.0245	0.3010300	$\bar{2}.3891661$
4	0.0370	0.6020600	$\bar{2}.5682017$
8	0.0570	0.9030900	$\bar{2}.7558749$
16	0.0855	1.2041200	$\bar{2}.9319661$
32	0.1295	1.5051500	$\bar{1}.1122698$
64	0.2000	1.8061800	$\bar{1}.3010300$
128	0.3035	2.1072100	$\bar{1}.4821587$

Dividing the data up into two groups, the first four sets constituting the first group and the last three sets the second group, we have:

$$n = 4, \quad \Sigma X = 3.0103000, \quad \Sigma Y = \bar{6}.6452088$$
$$n = 3, \quad \Sigma X = 5.4185400, \quad \Sigma Y = \bar{3}.8954585$$

Substituting these values in the residual equation

$$b\Sigma X + nk = \Sigma Y$$

we have:

$$3.01030b + 4k = 4.6452088 - 10$$
$$5.41854b + 3k = 7.8954585 - 10$$

Solving, we have:

$$b = 0.6047655$$
$$k = \log a = \bar{2}.2061708$$
$$a = 0.016076$$

Hence by the method of averages the required relation is:

$$i = 0.016076e^{0.6047655}$$

or

$$\log i = 0.6047655 \log e + \bar{2}.2061708$$

The computed values by this equation are given in Table 79, page 329. As an exercise the student should carry this problem through by averages, using four-place logarithms, and compare his results with ours.

We shall now solve this exercise by the method of least squares. The exercise affords an excellent opportunity for illustrating a short method. Continuing Table 77, we shall employ the following substitutions:

$$x' = \frac{X - 1.2041200}{0.3010300} \quad \text{and} \quad y' = Y + 1$$

or

$$X = 0.3010300x' + 1.2041200 \quad \text{and} \quad Y = y' - 1 \qquad (7)$$

Equation (6) will then become:

$$y' - 1 = b(0.3010300x' + 1.2041200) + k$$

or

$$y' = (0.30103b)x' + (1.20412b + k + 1)$$

or

$$y' = mx' + k' \tag{8}$$

where $m = 0.30103b$ and $k' = 1.20412b + k + 1$

From Section 59 (p. 218) we find m and k' by:

$$m = \frac{n\Sigma x'y' - \Sigma x'\Sigma y'}{n\Sigma x'^2 - (\Sigma x')^2}$$

$$k' = \frac{\Sigma x'^2\Sigma y' - \Sigma x'\Sigma x'y'}{n\Sigma x'^2 - (\Sigma x')^2}$$

We therefore continue Table 77 according to (7) and obtain Table 78.

TABLE 78

x'	y'	x'^2	$x'y'$
− 3	− 0.6108339	9	1.8325017
− 2	− 0.4317983	4	0.8635966
− 1	− 0.2441251	1	0.2441251
0	− 0.0680339	0	0.0000000
1	0.1122693	1	0.1122698
2	0.3010300	4	0.6020600
3	0.4821587	9	1.4464761
0	− 0.4593327	28	5.1010293

We can now find m and k':

$$m = 0.30103b = \frac{7(5.1010293)}{7(28)} = 0.18217961$$

$$b = 0.6051875$$

$$k' = 1.20412b + k + 1 = \frac{(28)(-0.4593327)}{7(28)} = -0.06561896$$

$$k = \log a = \bar{2}.2056627$$

$$a = 0.0160568$$

Hence by the method of least squares the required relation is:

$$i = 0.0160568e^{0.6051875}$$

or

$$\log i = 0.6051875 \log e + \bar{2}.2056627$$

In the following table we show the computed values which have been found from the equation determined by the method of averages and from the equation determined by the method of least squares.

TABLE 79

Observed Values		*Computed Values*			
		By Least Squares		*By Averages*	
e	i	$\log i$	i	$\log i$	i
2	0.0245	$\bar{2}.3878423$	0.0244	$\bar{2}.3882234$	0.0244
4	0.0370	$\bar{2}.5700219$	0.0372	$\bar{2}.5702759$	0.0372
8	0.0570	$\bar{2}.7520148$	0.0565	$\bar{2}.7523285$	0.0565
16	0.0855	$\bar{2}.9343811$	0.0860	$\bar{2}.9343810$	0.0860
32	0.1295	$\bar{1}.1165607$	0.1308	$\bar{1}.1164336$	0.1308
64	0.2000	$\bar{1}.2987403$	0.1990	$\bar{1}.2984862$	0.1988
128	0.3035	$\bar{1}.4809199$	0.3027	$\bar{1}.4805387$	0.3023

EXERCISES

1. Find an equation of the form $Y = aX^b$ for the data:

X	5	7	9	15	20	30	40	50
Y	1	2	3	9	16	37	65	100

2. Find an equation of the form $Y = aX^b$ for the data:

X	4	8	12	16	20	24
Y	2.9	23.0	77.8	184	360	622

3. Find an equation of the form $Y = aX^b$ for the data:

X	10	20	30	40	50	60
Y	11	31	57	88	122	161

4. If Y is the diameter of a tree in inches at age X years, the relation is $Y = aX^b$. For the following data, find the equation of the given type:

X	19	58	114	140	181	229
Y	3	7	13.2	17.9	24.5	33

5. A body in sliding down a plane of length l feet attained a velocity of V feet per second. Find the relation $V = al^b$ for the data given in the table:

l	19.9	45.1	67.5	94.4	109	126
V	10.1	15.2	18.6	22.0	23.6	25.4

6. The quantity of water, Q pounds, discharged per second from a circular orifice in a tank, under a pressure head of h feet, was found by experiment to result in the following data. Find the equation of the type $Q = ah^b$.

h	0.583	0.667	0.750	0.834	0.876	0.958	1.0
Q	7.00	7.60	7.94	8.42	8.68	9.04	9.34

7. At the following draughts, h feet, a particular vessel has the given tonnage, T, in salt water. Find the equation of the type $T = ah^b$.

h	15	12	9	6
T	2100	1510	1020	590

84. THE PARABOLA $Y = aX^2 + bX + c$

Due to the fact that the quadratic parabola possesses a three-constant flexibility, it is very useful in graduating statistical data from many fields. Three constants are to be determined, and this can be done by (1) the method of selected points, (2) the method of averages, (3) the method of least squares, and (4) the method of moments.

To apply the method of selected points, we draw a curve among the plotted points which will pass as near as possible to each of them. If the curve happens to pass through three of the plotted points or through any other three points whose coördinates can be approximately determined, we can substitute their coördinates in the given equation and solve the three resulting equations for a, b, and c. Of course the points so used should be chosen at the extreme and middle portions of the data.

As previously stated, the method of averages assumes that the sum of the residuals is zero. That is:

$$\Sigma(Y - aX^2 - bX - c) = 0$$

or

$$a\Sigma X^2 + b\Sigma X + nc = \Sigma Y \tag{9}$$

In order to obtain three equations which can be solved for the unknowns a, b, and c, we divide our data up into three sets. For each set find n, ΣX, ΣX^2, and ΣY. Substitute in (9) and solve for a, b, and c.

The method of least squares can be used to advantage with this curve. By proceeding as in Section 59 (p. 217), we can find three normal equations which can be solved for a, b, and c. Thus if ρ_i is any residual:

$$\rho_i = Y_i - aX_i^2 - bX_i - c$$

and

$$\Sigma\rho_i^2 = \Sigma(Y_i - aX_i^2 - bX_i - c)^2$$

The expression $\Sigma\rho_i^2$ can be written as a quadratic in a, in b, and in c. By imposing the condition that $\Sigma\rho_i^2$ be a minimum upon each quadratic we find the normal equations: [1]

$$\left.\begin{aligned} a\Sigma X^2 + b\Sigma X + cn &= \Sigma Y \\ a\Sigma X^3 + b\Sigma X^2 + c\Sigma X &= \Sigma XY \\ a\Sigma X^4 + b\Sigma X^3 + c\Sigma X^2 &= \Sigma X^2 Y \end{aligned}\right\} \tag{10}$$

Note that the first equation is merely the summation of the given function; the second is the summation of X multiplied into the given function, and the third is the summation of X^2 multiplied into the given function (see Exercise 16 at the end of this chapter).

If the values of X are in arithmetic progression — that is, if ΔX is constant — we can choose our units in such a manner that ΣX and ΣX^3 are zero. Further we may frequently use the relationships in Exercises 2, page 10, and 20b, page 22, to determine ΣX^2 and ΣX^4. By these artifices, the solution by least squares is not so laborious as it might appear.

A test for the use of the parabola is contained in the general theorem of Section 80 (p. 310). We shall quote here the theorem for our special case.

Theorem: *If, when ΔX is constant, $\Delta^2 Y$ is also constant, the relation between the variables may be expressed by the equation:*

$$Y = aX^2 + bX + c \tag{11}$$

Illustrative Example. The following table gives the modulus of torsion of steel T, in kilograms per square centimeter, at various temperatures θ in degrees Centigrade.

[1] A knowledge of the calculus would enable the student to write out such normal equations very easily. By setting the partial derivatives of $\Sigma\rho_i^2$ with respect to c, b, and a each equal to zero, equations (10) are obtained.

TABLE 80

$\Delta\theta$	θ	T	ΔT	$\Delta^2 T$
	0	8,290		
20			− 37	
	20	8,253		− 1
20			− 38	
	40	8,215		− 1
20			− 39	
	60	8,176		− 1
20			− 40	
	80	8,136		− 0
20			− 40	
	100	8,096		

We note that $\Delta\theta$ is constant ($= 20$) and that $\Delta^2 T$ is nearly constant, hence by the preceding theorem the data follow the law:

$$T = a\theta^2 + b\theta + c \tag{12}$$

We shall use the method of least squares, and in order to shorten the work we shall use the substitutions:

$$X = \frac{\theta - 50}{10} \quad \text{and} \quad Y = T - 8200$$

or

$$\theta = 10X + 50 \quad \text{and} \quad T = Y + 8200$$

Our equation (12) then becomes:

$$Y = AX^2 + BX + C$$

where

$$\left.\begin{aligned} A &= 100a \\ B &= 1000a + 10b = 10A + 10b \\ C &= 2500a + 50b + c - 8200 \\ C &= 5B - 25A + c - 8200 \end{aligned}\right\} \tag{13}$$

To form our normal equations we prepare the following table:

TABLE 81

θ	T	X	Y	X^2	X^3	X^4	XY	X^2Y	Computed T
0	8,290	− 5	90	25	− 125	625	− 450	2,250	8,290.1
20	8,253	− 3	53	9	− 27	81	− 159	477	8,252.9
40	8,215	− 1	15	1	− 1	1	− 15	15	8,214.9
60	8,176	1	− 24	1	1	1	− 24	− 24	8,176.0
80	8,136	3	− 64	9	27	81	− 192	− 576	8,136.3
100	8,096	5	−104	25	125	625	− 520	− 2,600	8,095.8
Total		0	− 34	70	0	1,414	− 1,360	− 458	

Substituting the proper sums in equations (10) we have the following normal equations:

$$6C + 70A = -34$$
$$70B = -1360$$
$$70C + 1414A = -458$$

Solving, we obtain

$$A = -0.10267857 \qquad B = -19.428571 \qquad C = -4.46875$$

from which it follows, using (13), that:

$$a = -0.0010268 \qquad b = -1.8402 \qquad c = 8290.11$$

Hence our equation is:

$$T = -0.0010268\theta^2 - 1.8402\theta + 8290.11$$

Assigning the given values to θ we obtain the computed values of T that are found in the last column of the Table 81.

85. OTHER USEFUL CURVES

In this chapter we have attempted to introduce the student to some of the methods of fitting simple curves to observed data. We have considered in great detail the methods of fitting the straight line, the exponential function, the power function, and the quadratic polynomial.

We shall mention now with less detail a few additional well-known curves that are frequently found useful.

A. The Hyperbola $\qquad Y = a + \frac{b}{X}$ (14)

This equation represents the hyperbola with the lines $Y = a$ and $X = 0$ as asymptotes. It can be written in the form:

$$Y = a + b\left(\frac{1}{X}\right),$$

which is a straight line with slope b in the $\left(\frac{1}{X}, Y\right)$ coördinates. Hence we can state the test: If $\Delta Y/\Delta(1/X)$ is constant, the data obey the law given by (14).

It may be noted that if $a = 0$, we have as a special case the well-known:

$$XY = b$$

B. The Hyperbola

$$Y = \frac{X}{a + bX} \tag{15}$$

This equation represents the hyperbola with the lines $a + bX = 0$ and $bY = 1$ as asymptotes. We can write the equation in the form

$$\frac{X}{Y} = a + bX \tag{16}$$

which is that of a straight line with slope b in the $(X, X/Y)$ coördinates. Hence we can state the test: If $\Delta(X/Y)/\Delta X$ is constant, the data may be represented by (15).

The methods for determining the constants for (14) and (15) should be evident.

C. The Modified Exponential Function $Y = a + bc^X$. The following theorem may be used to determine if the modified exponential law is applicable.

Theorem: *If the values of X are in arithmetic progression and the values of ΔY are in geometric progression, the data follow the law:*

$$Y = a + bc^X \tag{17}$$

TABLE 82

X	Y	ΔY
X_1	Y_1	
		ΔY_1
$X_2 = X_1 + \Delta X$	$Y_2 = Y_1 + \Delta Y_1$	
		$\Delta Y_2 = r\Delta Y_1$
$X_3 = X_1 + 2\Delta X$	$Y_3 = Y_1 + \Delta Y_1 + r\Delta Y_1$	
. . .	. . . . .	. . .
. . .	. . . . .	. . .
		$\Delta Y_{n-1} = r^{n-2}\Delta Y_1$
$X_n = X_1 + (n-1)\Delta X$	$Y_n = Y_1 + \Delta Y_1 + r\Delta Y_1 + \cdots + r^{n-2}\Delta Y_1$	

From the hypothesis we have, since the values of ΔY are in geometric progression:

$$Y_n = Y_1 + \Delta Y_1 + r\Delta Y_1 + r^2\Delta Y_1 + \cdots + r^{n-2}\Delta Y_1$$

Using the formula for the sum of a geometric progression, we have:

$$Y_n = Y_1 + \Delta Y_1\left[\frac{1 - r^{n-1}}{1 - r}\right]$$

or

$$Y_n = Y_1 + \frac{\Delta Y_1}{1 - r} - \frac{\Delta Y_1}{1 - r} \cdot r^{n-1} \tag{18}$$

Further, since the values of X are in arithmetic progression, we have:

$$X_n = X_1 + (n - 1)\Delta X$$

or

$$n - 1 = \frac{X_n - X_1}{\Delta X}$$

Substituting this value of $(n - 1)$ in (18) we have:

$$Y_n = Y_1 + \frac{\Delta Y_1}{1 - r} - \frac{\Delta Y_1}{1 - r} \cdot r^{\frac{X_n - X_1}{\Delta X}}$$

or

$$Y_n = a + bc^{X_n}$$

where

$$a = Y_1 + \frac{\Delta Y_1}{1 - r}, \qquad b = -\frac{\Delta Y_1}{1 - r} r^{\frac{-X_1}{\Delta X}}, \qquad c = r^{\frac{1}{\Delta X}}$$

To determine the constants of this equation we shall employ the method of selected points. We draw the best-fitting curve among the points. We now choose three points on the curve whose coordinates are known — or can be estimated — and *whose abscissas are in arithmetic progression.* We can form three equations by substituting the coördinates of the selected points in (17), and solve for the unknowns.

Exercise. A curve of the type $Y = a + bc^X$ passes through the three points (1, 10), (3, 28), and (5, 100). What is its equation?

For an illustrative example, consider the data of Table 83 on page 336.

TABLE 83

X	Y	ΔY	*Ratio of Any ΔY to Preceding*
4	20.1		
		7.7	
5	27.8		
		7.2	0.93
6	35.0		
		6.5	0.90
7	41.5		
		6.1	0.93
8	47.6		
		5.4	0.88
9	53.0		
		5.1	0.94
10	58.1		
		4.6	0.90
11	62.7		
		4.1	0.89
12	66.8		
		3.8	0.92
13	70.6		

We note that the values of X are in arithmetic progression, that the values of ΔY are approximately in geometric progression, and conclude that our data follow the law:

$$Y = a + bc^X$$

To determine the constants, assume that the curve passes through the points (4, 20.1), (8, 47.6), and (12, 66.8). We then have the equations:

$$a + bc^4 = 20.1$$
$$a + bc^8 = 47.6$$
$$a + bc^{12} = 66.8$$

Then

$$bc^4(c^4 - 1) = 47.6 - 20.1 = 27.5$$
$$bc^8(c^4 - 1) = 66.8 - 47.6 = 19.2$$

and by division we obtain:

$$c^4 = 0.6982$$
$$c = 0.9141$$

By substitution we have:

$$b(0.6982)(0.6982 - 1) = 27.5$$

and

$$b = -130.5$$

Now a is easily found, for:

$$a + (-130.5)(0.6982) = 20.1$$

and

$$a = 111.2$$

Hence our equation is:

$$Y = 111.2 - 130.5(0.9141)^X$$

Other selections of the points will give slightly different values. The computed values and the residuals may now be found.

D. The Modified Power Function $Y = c + aX^b$. A test for the applicability of this law is contained in the

Theorem: *If the values of X form a geometric progression and the values of ΔY also form a geometric progression, then the data obey the law:*

$$Y = c + aX^b \tag{19}$$

TABLE 84

X	Y	ΔY
X_1	Y_1	
		ΔY_1
$X_2 = rX_1$	$Y_2 = Y_1 + \Delta Y_1$	
		$\Delta Y_2 = R\Delta Y_1$
$X_3 = r^2X_1$	$Y_3 = Y_1 + \Delta Y_1 + R\Delta Y_1$	
. . .	. . . . .	. . .
. . .	. . . . .	. . .
. . .	. . . . .	
$X_n = r^{n-1}X_1$	$Y_n = Y_1 + \Delta Y_1 + R\Delta Y_1 + \cdots + R^{n-2}\Delta Y_1$	$\Delta Y_{n-1} = R^{n-2}\Delta Y_1$

From the hypothesis we have:

$$X_n = r^{n-1}X_1 \quad \text{or} \quad n - 1 = \frac{\log X_n - \log X_1}{\log r}$$

and

$$Y_n = Y_1 + \Delta Y_1[1 + R + R^2 + \cdots + R^{n-2}]$$

or

$$Y_n = Y_1 + \Delta Y_1 \left[\frac{1 - R^{n-1}}{1 - R}\right]$$

The remainder of the proof easily follows, and we leave its completion to the reader.

As in the modified exponential, we determine the constants by the method of selected points but in this case the *abscissas should be chosen in geometric progression.*

86. LIMITATIONS OF EMPIRICAL EQUATIONS

In the preceding pages of this chapter we have been concerned with two fundamental questions that relate to empirical equations: first, what type of equation should be selected to describe the data, and, having decided upon the type of equation, the second question is, how can the constants be determined? Having answered the first question, the second presents no great difficulty.

Once the equation for the data has been determined, we have an expression that may be used, within certain limits, to estimate values of the dependent variable and thus to compare values on the curve with observed values. Further, if a criterion of goodness of fit is desired, we may turn to the sum of the squares of the residuals.

To assist in determining the type of equation to be selected we have devised tests to apply to the observations. The illustrative examples that we have solved have enjoyed a singular peculiarity; they have presented data for which the tests were closely satisfied. In general, the data have come from the laboratories of the physical sciences where it is possible to restrict the problem to a study of the variables in question, and to control or eliminate outside influences. There have been internal as well as mathematical reasons for selecting an equation of given type and thus our empirical equations have been "true relations" between the variables in question.

When a physicist is analyzing a set of *distance, time* data of the flight of a projectile, he will know for internal reasons that his curve is a second degree parabola $D = AT^2 + BT + C$. Similarly, a chemist in analyzing *pressure, volume* data would likely choose $P = AV^B$. As a result of slow and painful research, the scientist learns how certain phenomena behave. It frequently occurs that a study of empirical data leads to a formulation and discovery of rela-

tionships that the investigator had not been able to formulate from analytical considerations. A classic example of this method was the discovery and formulation of Kepler's Laws which explain the motions of the planets. These laws were formulated by Johann Kepler (1571–1630) *after* a study of a tremendous quantity of observed data collected over a number of years by the brilliant astronomer, Tycho Brahe (1546–1601). The truths hidden in the data were not revealed to the observer, Brahe, but when Kepler analyzed the data he saw in them relationships that he formulated into what are known as Kepler's Laws. Science is replete with similar examples.

When one moves outside the realm of physical science, he has difficulty in finding an equation that explains and expresses a "true relationship." Internal evidence is lacking. Too many uncontrollable influences are present that cannot be eliminated, and thus our data may not lead to an analytical formulation of an inherent relationship. In biological, educational, economic, and social relationships our knowledge is too limited to enable us to say *why* a relationship exists. The best we can do in these fields is to find a functional relationship between the variables in question for the particular data at hand. Generally, we cannot *explain* the why of the relationship. In such cases the data obviously may not reveal that a *certain type of equation* is indicated. Sometimes experience comes to the assistance of the investigator, otherwise he does what all of us do, namely, *the best he can.*

Usually the purpose of this functional relationship is to estimate sufficiently well the values of one variable from known values of another, and frequently *this purpose can be accomplished by using more than one type of equation.* In fact, we can establish the functional relationship without an equation at all. If to each value of X there is determined one or more values of Y, then Y is a function of X. We may determine the values of Y from a graph, a table of values, and that is all that is really necessary. However, much is gained if we can obtain a summarizing expression in the form of an equation.

We then face the practical problem of finding a functional relationship. If we choose to find an equation, the curve may fit poorly or closely. When the data are such that a careful analysis is warranted, they should be subjected to a careful analysis; however, should they

not warrant a careful analysis, it is the height of absurdity to subject them to such a treatment. The investigator must determine the type of treatment the data merit.

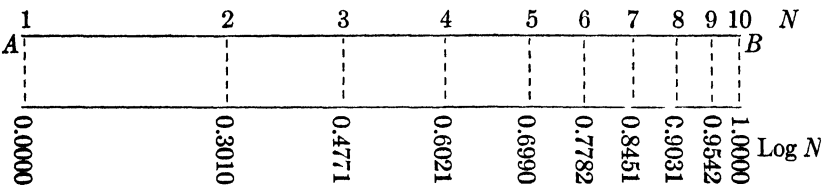
In our previous sections we have discussed methods of dealing with precise measurements in a precise manner. In fact, we have frequently used eight-place logarithms in our computations in order that our results might be the more precise. In the next section we discuss methods of dealing with data that may not merit a careful analysis.

87. GRAPHICAL METHODS IN TREND ANALYSIS

Frequently workers in practical statistics are confronted with data that do not warrant a careful algebraical and numerical analysis. Rough approximations may be sufficiently accurate for the investigator's needs. In such cases he usually resorts to the use of graphical methods. Especially are graphs widely employed in trend analysis. Not only may the graph be used to give a clew to the equation of the curve that may be used to represent the trend; it may even be used to determine the unknown constants that appear in the equation that is selected.

We are familiar with graphs made on the conventional cross-section coördinate paper. On this paper a given distance in any direction, when applied to a given problem, always represents a constant quantity. Such paper may be specifically called "arithmetic paper," and the uniform scale an "arithmetic scale."

We may, however, develop scales on which equal distances do not always represent equal magnitudes. A very common and widely used scale of this kind is the "logarithmic scale" on which equal distances represent equal proportional or percentage changes. In this scale the points correspond to the logarithms of numbers. By placing the natural numbers, N, and their logarithms, $\log N$, into



correspondence, it is noted that the logarithms are spaced uniformly along the line while the integers are spaced non-uniformly.

The scale from 1 to 10 as shown on the line AB constitutes a *cycle*. Any number on the scale, say X, corresponds to log X. That is, *the logarithmic scale serves the purpose of finding the logarithms.* By prolonging the line AB and repeating the scale, we may construct a segment of two cycles.

It is customary to assign a value to the initial point A. It may be any number greater than zero. The value to be assigned is determined by the problem at hand. The value placed at the end of the cycle, B, is 10 times the value assigned to the point, A. Thus, the numbers along the following scale, AB, serve as illustrations.

A								B
1	2	3	4	5	6	. . . .	9	10
2	4	6	8	. . . . .				20
5	10	15	20	. . . . .				50
13	26	39	52	. . . . .				130

A. Arithmetic Paper. As an illustration of the use of the graphical method in determining the straight-line trend we shall consider the data that were given in Exercise 3, page 315.

TABLE 85. ANNUAL EARNINGS OF THE ASSOCIATED GAS AND ELECTRIC SYSTEM, 1920–1928

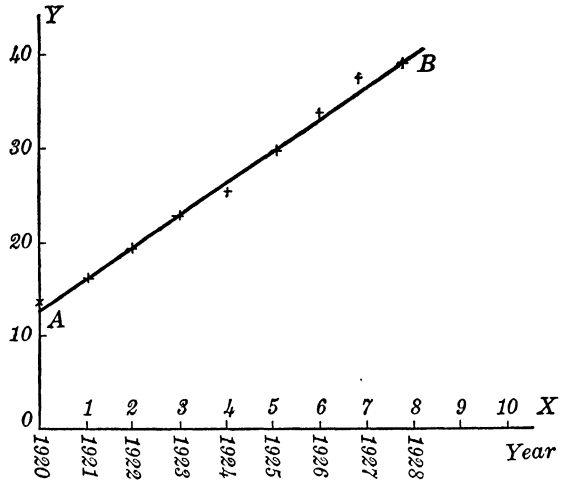
Year	*Net Earnings (millions of dollars)*	*Year*	*Net Earnings (millions of dollars)*
1920	13.4	1925	29.5
1921	16.2	1926	33.5
1922	19.2	1927	37.8
1923	22.7	1928	40.6
1924	25.1	1929	....

We plot the data carefully on arithmetic coördinate paper with $X = 0$ at 1920 [Figure 44]. The observed points are indicated by the small crosses. We then sketch in "by sight" the line of trend. It cuts the Y-axis at 12.5. By using this point and the point (8, 40.5) as two known points on the line, we obtain

$$m = \frac{40.5 - 12.5}{8 - 0} = 3.5$$

Hence we have the equation of the line $Y = 3.5X + 12.5$.

FIGURE 44



In general we proceed as follows: We plot the data carefully on arithmetic paper. Next, we draw in by sight the trend line. Then selecting two widely separated points A and B on the line, we evaluate the ratio of the difference in the ordinates to the difference of the abscissas of the two points. This gives us the slope, m, of the trend line. Using this slope with some point on the line whose coordinates are read from the graph, we can find from the point-slope form

$$Y - Y_1 = m(X - X_1)$$

the equation of the trend. Of course if the Y-intercept can be determined from the graph, we may use the slope-intercept form

$$Y = mX + b$$

and thus find the equation of the trend line.

Obviously this same method may be employed for parabolic, exponential, or other types of trend. We choose points equal in number to the number of constants in the equation, substitute the coördinates in the chosen equation, and solve for the unknowns.

B. Semi-logarithmic Paper. Logarithmic scales may be used on the axes of coördinate paper. If the scale on one of the axes is logarithmic and on the other is arithmetic, the paper is called semi-

logarithmic paper. This type of paper, usually with three cycles, can be purchased at stationery stores. It is used by statisticians in studying the growth of populations, bank clearings — in short, in studying data that may follow the exponential function $Y = ab^X$.

The following theorems contain the gist of the theory.

Theorem 1. The graph of the exponential function $Y = ab^X$ plotted on semi-logarithmic paper is a straight line whose slope is $\log b$ and whose intercept on the non-uniform scale is a.

Proof: From

$$Y = ab^X$$

we have, taking logarithms,

$$\log Y = (\log b)X + \log a$$

which is an equation of the first degree in the $(X, \log Y)$ coördinates, and consequently represents a straight line. The slope is $\log b$ and the vertical intercept is a on the $\log Y$-axis. That is, if the points (X, Y) plotted on uniform coördinate paper fall upon the curve $Y = ab^X$, when plotted on semi-logarithmic paper they fall upon the straight line $\log Y = (\log b)X + \log a$.

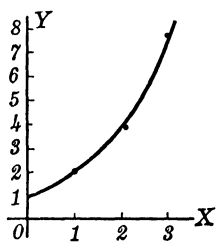
Conversely, if the points (X, Y) when plotted on semi-logarithmic paper is a straight line with slope $\log b$ and with the intercept on the non-uniform scaled vertical axis a, the data follow the exponential law $Y = ab^X$.

We shall leave the proof as an exercise for the reader.

Example 1. Draw the graph of the curve $Y = 2^X$ on arithmetic paper and on semi-logarithmic paper.

We prepare a table of values and plot the points as indicated.

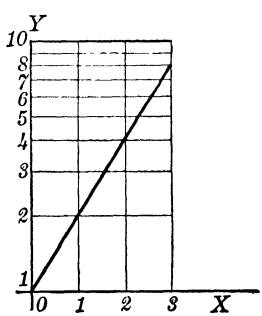
FIGURE 45(a)



$Y = 2^X$

X	Y
0	1
1	2
2	4
3	8

FIGURE 45(b)

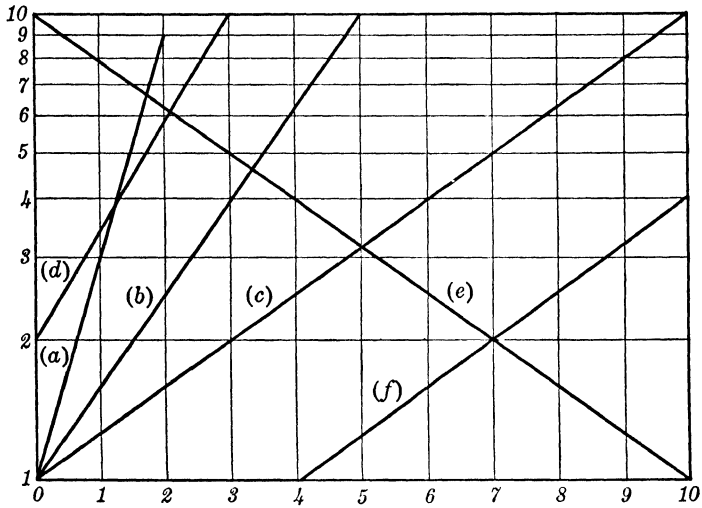


It is noted that the equation plots into a curve on the arithmetic paper and into a straight line on the semi-logarithmic paper. The equation of the straight line in the (X, log Y) coördinates may be written

$$\log Y = (\log 2)X + \log 1$$

in which the slope is log 2 and the vertical intercept is 1.

FIGURE 46



EXERCISES

1. Find the equations of the straight lines in Fig. 46 in semi-logarithmic form. What are the corresponding equations in exponential form?

2. Plot the following equations on semi-logarithmic paper.

(a) $Y = 2(3)^X$ (c) $\log Y = 0.5X + \log 3$

(b) $Y = 2(10)^{2X}$ (d) $Y = 3(10)^{-2X}$

3. If \$10 is invested at 5 per cent compounded annually the amount Y at the end of X years is given by $Y = 10(1.05)^X$. Plot this curve on semi-logarithmic paper.

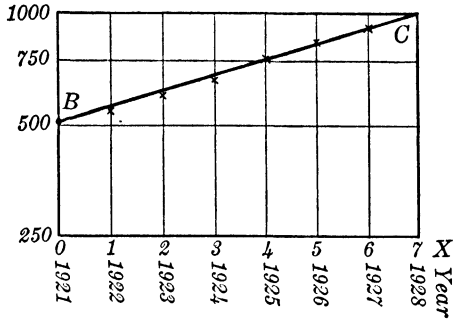
Let us next employ semi-logarithmic paper to determine graphically the approximate exponential equation that obtains for a mass of empirical data. We illustrate the procedure in Example 2.

Example 2. Find graphically the exponential trend of the gross earnings in millions of dollars of all Bell telephone companies in the United States as given in the accompanying table.

TABLE 86

Year	*Earnings*
1921	521
1922	564
1923	623
1924	678
1925	761
1926	845
1927	917
1928	1003

FIGURE 47



We choose $X = 0$ at 1921. We note that the data, when plotted on semi-logarithmic paper, lie along the line BC which we draw by sight. For this line $a = 520$. Taking the point $C(7,1000)$ as a second point on the line we have

$$\text{slope} = \log b = \frac{\log 1000 - \log 520}{7 - 0}$$

$$= \frac{3.0000 - 2.7160}{7} = 0.0406$$

$$b = 1.1 \text{ approximately}$$

The equation of the straight line in semi-logarithmic coördinates is therefore

$$\log Y = 0.0406X + \log 520$$

and the corresponding exponential equation is

$$Y = 520(1.1)^X$$

EXERCISES

1. The registration (in millions) of motor vehicles in the United States in the given years is shown by the following table. [*Statistical Abstract of the U.S.*, 1930, p. 385.] Using semi-logarithmic paper, find an exponential function that will approximately fit the data.

Year	*Registration*	*Year*	*Registration*
1917	5.0	1922	12.2
1918	6.1	1923	15.1
1919	7.6	1924	17.6
1920	9.2	1925	19.9
1921	10.5	1926	22.0

2. Use semi-logarithmic paper to fit an exponential curve to the following data which give the average number of shares (in millions) sold on the New York Stock Exchange from 1919 to 1929 inclusive.

Year	*Sales*	*Year*	*Sales*
1919	26.07	1925	37.69
1920	18.73	1926	37.42
1921	14.30	1927	48.08
1922	21.73	1928	76.71
1923	19.77	1929	93.75
1924	23.50		

3. Use semi-logarithmic paper to fit an exponential curve to the following data which give the production (in millions of barrels) of petroleum in the United States 1920–1929. [*Statistical Abstract of the United States*, 1936, p. 723.]

Year	*Production*	*Year*	*Production*
1920	443.0	1925	763.7
1921	472.2	1926	770.9
1922	557.5	1927	901.1
1923	732.4	1928	901.5
1924	713.9	1929	1007.3

C. Logarithmic Paper. Thus far we have used two types of coördinate paper in our work, arithmetic and semi-logarithmic. In the arithmetic paper, the scale along both axes is the natural scale. The semi-logarithmic paper has the natural scale along the axis of abscissas and a logarithmic scale along the axis of ordinates.

Another useful type of paper is *logarithmic* paper. This paper is ruled with logarithmic scales both horizontally and vertically. It is frequently called *double logarithmic* and *log-log* paper. When a point (X, Y) is plotted on log-log paper, its actual distances from the reference lines are proportional to $\log X$ and $\log Y$. In other words, in graphing pairs of numbers on logarithmic paper we really graph the logarithms of the numbers. The logarithmic paper serves the purpose of finding the logarithms of the numbers. The effect of this is to tone down the contrasts. For examples,

$$\log 1000 \text{ is only } 3, \quad \text{and} \quad \log 0.0001 \text{ is } -4.$$

Double logarithmic paper is very useful in studying the power function

$$Y = aX^b$$

where a and b are constants. This is due to the fact that the graph of the power function on logarithmic paper is a straight line. For we have the

Theorem: The graph of the power function

$$Y = aX^b$$

plotted on logarithmic paper is the straight line whose slope is b and whose intercept on the Y-axis is a.

Proof: Taking logarithms of the above equation, we have

$$\log Y = b \log X + \log a$$

which is an equation, in logarithmic coördinates, of a straight line with slope b and Y-intercept a.

Conversely, if the (X, Y) data when plotted on logarithmic paper give a straight line with slope b and Y-intercept a, the data follow the law

$$Y = aX^b$$

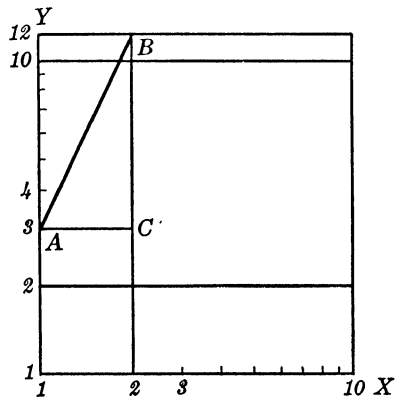
Proof: The $(\log X, \log Y)$ relation is linear. Hence

$$\log Y = b \log X + \log a$$

which can be immediately reduced to

$$Y = aX^b$$

FIGURE 48



Example 1. Draw the graph of $Y = 3X^2$ on logarithmic paper.

Since the graph is a straight line, we need but two points say (1, 3) and (2, 12) to determine the constants. These two points determine the line AB of Figure 48.

The slope b is given by

$$b = \frac{\log 12 - \log 3}{\log 2 - \log 1}$$

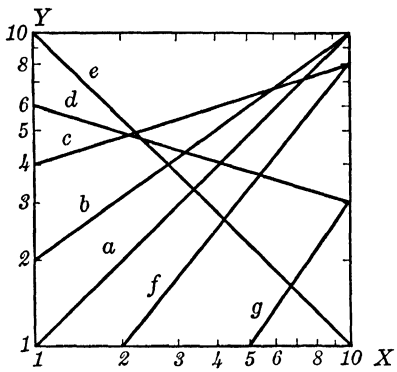
$$= \frac{\log 4}{\log 2} = \frac{2 \log 2}{\log 2}$$

$$b = 2$$

From the figure $a = 3$, hence the log-log equation is

$$\log Y = 2 \log X + \log 3$$

FIGURE 49



Example 2. Find the equation of the curve that graphs into the line marked c of Figure 49.

We have

$$\text{slope} = \frac{\log 8 - \log 4}{\log 10 - \log 1}$$

$$= \frac{\log 2}{1} = \log 2$$

$$b = \log 2$$

From the figure $a = 4$, and the log-log equation is

$$\log Y = (\log 2) \log X + \log 4$$

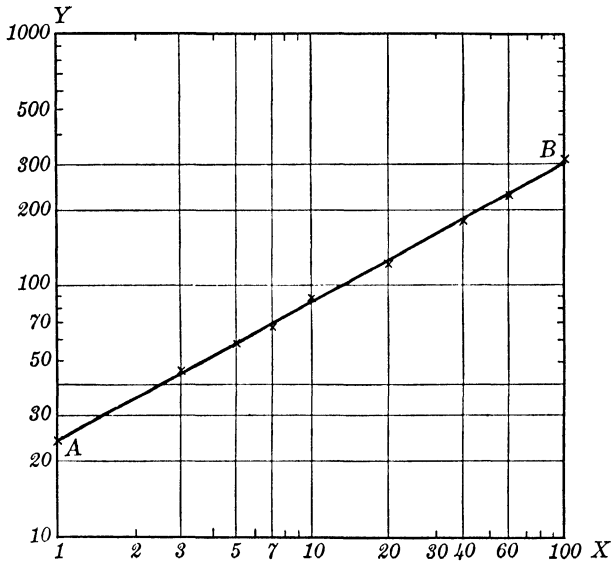
Using the properties of logarithms we obtain

$$Y = 4X^{\log 2} = 4X^{.3010}$$

Example 3. Using log-log paper find the equation that approximately fits the data:

X	1	3	5	7	10	20	40	60	100
Y	25	45	60	70	90	130	190	240	300

FIGURE 50



In solving this problem we find it necessary to use two cycle log-log paper. We indicate the points by small crosses. Since the points lie approximately upon the straight line AB, the data may be approximately represented by $Y = aX^b$. Assume that the line passes through the points $A(1,25)$ and $B(100,300)$.

$$b = \text{slope} = \frac{\log 300 - \log 25}{\log 100 - \log 1} = \frac{\log 12}{2} = \frac{1.0792}{2} = 0.54$$

From Figure 50 the Y-intercept $= a = 25$. Hence the log-log equation is

$$\log Y = 0.54 \log X + \log 25$$

from which we immediately obtain

$$Y = 25X^{0.54}$$

EXERCISES

1. Find the log X, log Y and the X, Y equations of the lines a, b, d, e, f, and g of Figure 49.

2. Use log-log paper to determine the log X, log Y and the X, Y equations for the data given in the table:

X	4	8	12	16	20	24
Y	2.9	23.0	77.8	184.0	360.0	622.1

3. Use log-log paper to determine the approximate X, Y relation for the data given in the table:

X	10	20	30	40	50	60
Y	11	31	57	88	122	161

4. Plot the following data on two-cycle log-log paper and determine the log X, log Y and the X, Y relations.

X	5	7	9	15	20	30	40	50
Y	1	2	3	9	16	37	65	100

5. Solve Exercise 4 above using the method of averages for the (log X, log Y) straight line.

6. Use semi-log paper to find the X, log Y and the X, Y equations for the data:

X	1.6	3.1	4.7	6.3	7.9	9.4	11.0
Y	5.4	7.2	9.6	12.8	17.1	22.9	30.8

7. The number N of bacteria in a given culture t hours after they were first observed was found to be that given by the table. Using semi-log paper find N in terms of t.

t	0	1	2	3	4	5	6
N	125	209	340	561	924	1525	2512

8. The number N of bacteria in a culture at the end of t hours is shown by the following table. Use semi-log paper to find N in terms of t.

t	0	1	2	3	4	5	6
N	100	162	265	450	742	1230	2020

9. The annual expenditure of the United States Government (in millions of dollars) has increased as in the table. Use semi-log paper to determine the appropriate law. Would you advise using this law to extrapolate for the expenditure in 1918?

Year	*Expenditure*	*Year*	*Expenditure*
1840	24	1880	265
1850	41	1890	298
1860	63	1900	488
1870	294	1910	660

10. The total assets (in billions of dollars) of Building and Loan Associations in the United States for the given years are shown in the following table. Use semi-log paper to find the X, log Y and the X, Y equations.

Year	X	*Assets* Y	*Year*	X	*Assets* Y
1920	0	2.52	1925	5	5.51
1921	1	2.89	1926	6	6.33
1922	2	3.34	1927	7	7.18
1923	3	3.94	1928	8	8.02
1924	4	4.77	1929	9	8.70

11. The following table gives the average monthly imports of wood pulp (millions of short tons) into the United States for the given years. Choose $X = 0$ at 1926 and find the straight-line equation by selected points. Extrapolate for the years 1931, 1932, and 1933. The actual imports these years were 133.0, 123.5, and 161.8 short tons.

Year	*Imports*	*Year*	*Imports*
1922	105	1927	140
1923	115	1928	147
1924	127	1929	157
1925	139	1930	152
1926	145		

12. The following table gives the production of women's shoes (in millions of pairs) for the given years. Plot the data on semi-logarithmic paper and determine the X, log Y and the X, Y relations using $X = 0$ at 1931. Find the extrapolated value for 1940. The actual value was 12.5 million pairs.

Year	*X*	*Production*	*Year*	*X*	*Production*
1931	0	9.4	1936	5	13.5
1932	1	9.5	1937	6	12.5
1933	2	10.9	1938	7	12.3
1934	3	11.1	1939	8	14.0
1935	4	12.1	1940	9	

13. The following table gives (in millions of pounds) the domestic consumption of rayon in the United States from 1920 to 1936. Plot on semi-logarithmic paper with $X = 0$ at 1920, and find graphically the X, log Y and the X, Y equations. Find the extrapolated value for 1937. The actual value was 261.2 million pounds.

Year	*X*	*Consumption*	*Year*	*X*	*Consumption*
1920	0	9	1929	9	131
1921	1	20	1930	10	118
1922	2	25	1931	11	157
1923	3	33	1932	12	152
1924	4	42	1933	13	212
1925	5	58	1934	14	195
1926	6	61	1935	15	253
1927	7	100	1936	16	298
1928	8	100	1937	17	

14. Plot the data of Exercise 13 above on arithmetic paper and use the method of selected points to find the equation of the parabola $Y = AX^2 + BX + C$ that will approximately fit the data. Choose $X = 0$ at 1920. Extrapolate for 1937.

15. The following table gives the annual production of cigarettes (billions) in the United States in the given years. Use semi-logarithmic paper to find the X, log Y and the X, Y equations. Choose $X = 0$ at 1920, and let Y = Production. Find the extrapolated value for 1930. The actual value for 1930 was 123.8 billions.

Year	*Annual Production (billions)*	*Year*	*Annual Production (billions)*
1920	47.4	1925	82.2
1921	52.1	1926	92.1
1922	55.8	1927	99.8
1923	66.7	1928	108.7
1924	72.7	1929	122.3

16. Find the trend line for the changing price of beef as described in the data of Table 12 (p. 47).

17. In the following table the unit is 1,000,000 barrels of 42 gallons.

ANNUAL PRODUCTION OF PETROLEUM IN THE UNITED STATES, 1900–1913

Year	*Production*	*Year*	*Production*	*Year*	*Production*
1900	63.6	1905	134.7	1910	209.6
1901	69.4	1906	126.5	1911	220.4
1902	88.8	1907	166.1	1912	222.9
1903	100.5	1908	178.5	1913	248.4
1904	117.1	1909	183.2		

Find the equation of the trend line, the computed values of the production for the given years, and the residuals. Find the predicted values for the years 1915 and 1920 and compare your results with those given in *Commerce Yearbook*, 1930, page 293, which are as follows: 1915, production, 281.1; 1920, production, 442.9. What can you say for the trend line for purposes of prediction?

18. Find the equation of the trend line (a) excluding the years 1916, 1917, and 1918, and (b) including these years. Find the computed production and the residuals in each case.

AVERAGE MONTHLY PRODUCTION OF PIG IRON IN THE UNITED STATES, 1903–1918 [1]

Year	*Production (1000 long tons)*	*Year*	*Production (1000 long tons)*	*Year*	*Production (1000 long tons)*
1903	1,452	1909	2,116	1914	1,921
1904	1,344	1910	2,237	1915	2,472
1905	1,882	1911	1,944	1916	3,252
1906	2,066	1912	2,448	1917	3,182
1907	2,109	1913	2,560	1918	3,209
1908	1,302				

19. Find the equation of the trend line, the computed values of the production, and the residuals.

[1] The data are taken from *Review of Economic Statistics*, Vol. I, p. 66; United States Department of Commerce, *Survey of Current Business*, No. 42, p. 44.

TOTAL PRODUCTION OF CRUDE STEEL, 1900–1929 [1]

Year	*Production (millions of long tons)*	*Year*	*Production (millions of long tons)*	*Year*	*Production (millions of long tons)*	*Year*	*Production (millions of long tons)*
1900	10.6	1908	14.0	1916	42.8	1923	44.9
1901	13.5	1909	24.0	1917	45.1	1924	37.9
1902	14.9	1910	26.1	1918	44.5	1925	45.4
1903	13.9	1911	23.7	1919	34.7	1926	48.3
1904	13.9	1912	31.3	1920	42.1	1927	44.9
1905	20.0	1913	31.3	1921	19.8	1928	51.5
1906	23.4	1914	23.5	1922	35.6	1929	56.4
1907	23.4	1915	32.2				

88. GOODNESS OF FIT OF CURVES TO OBSERVED DATA: NONLINEAR CORRELATION

A. Goodness of Fit. The investigator who takes the time to derive an empirical formula for a set of observed data is naturally interested in knowing how well the curve fits the observations. He therefore will always find the computed values of the dependent variable by his formula, and usually the Y-residuals if Y is the dependent variable.

Any Y-residual, it will be recalled, is given by ρ_i where

$$\rho_i = \text{the observed } Y_i - \text{the computed } Y_i$$

The variation in the residuals may be measured by their mean deviation or by their standard error. That is by:

$$M.D. \text{ of } \rho = \frac{\Sigma |\rho|}{n}$$

or by

$$S_y = \sqrt{\frac{\Sigma \rho^2}{n}}$$

If the constants have been found by the method of selected points or by the method of averages, the mean deviation is adequate, but

[1] The data are taken from *Statistical Abstract of the United States*, 1918, p. 251; *ibid.*, 1930, p. 756.

if the constants have been determined by the method of least squares, S_y is the natural measure. In either case the results will be expressed in the given Y unit.

In Section 63 (p. 238) while evaluating S_y for the line $Y = mX + b$, which has been fitted by least squares, we found

$$S_y = \sigma_Y\sqrt{1 - r^2}$$

where $r = \dfrac{\Sigma xy}{n\sigma_X\sigma_Y}$ is the cross-product formula for measuring linear correlation. We also found r to be an excellent measure of the goodness of fit of the points to the derived line. If the formula above is solved for r, we have

$$\text{correlation based upon the straight line} = r = \sqrt{1 - \frac{S_y^2}{\sigma_Y^2}} \qquad (20)$$

where S_y is the standard error of estimate based upon the straight line.

B. Nonlinear Correlation. The process of arriving at a coefficient of correlation based upon curvilinear regression is comparatively simple in principle but often becomes very complex in practice. To emphasize the evident simplicity of the process let us proceed exactly as we did in Section 63 and find a coefficient of correlation based upon the parabola

$$y = ax^2$$

where x and y are deviations of X and Y from their respective means, M_X and M_Y. Since any y-residual is given by

$$\rho_i = y_i - ax_i^2$$

we have

$$\Sigma\rho_i^2 = a^2\Sigma x_i^4 - 2a\Sigma x_i^2y_i + \Sigma y_i^2$$

which is a quadratic in a. Now $\Sigma\rho_i^2$ is a minimum when

$$a = \frac{-(-2\Sigma x^2y)}{2\Sigma x^4} = \frac{\Sigma x^2y}{\Sigma x^4}$$

Hence the best-fitting curve is given by

$$y = \left(\frac{\Sigma x^2y}{\Sigma x^4}\right)x^2$$

where a is computed, of course, from the observed values.

For this value of a, the sum of the squares of the residuals becomes:

$$\Sigma\rho^2 = \Sigma x^4\left[\frac{\Sigma x^2y}{\Sigma x^4}\right]^2 - 2\Sigma x^2y\left[\frac{\Sigma x^2y}{\Sigma x^4}\right] + \Sigma y^2$$

and $S_y^2 = \dfrac{\Sigma\rho^2}{n}$ becomes:

$$S_y^2 = \sigma_Y^2\left[1 - \left(\frac{\Sigma x^2y}{\sqrt{\Sigma y^2\Sigma x^4}}\right)^2\right]$$

Now evidently a coefficient of correlation based upon the parabola $y = ax^2$ is the expression:

$$\frac{\Sigma x^2y}{\sqrt{\Sigma y^2\Sigma x^4}}$$

Note that in this case

$$\text{the coefficient of correlation based on the parabola} = \sqrt{1 - \frac{S_y^2}{\sigma_Y^2}}$$

where S_y is the standard error of estimate for the parabola.

In order to emphasize that this simple method will not always work, let the student undertake its application to the curves:

$$y = x^a \quad \text{and} \quad y = a^x$$

He will soon discover that "the method is simple in principle but very complex in practice."

However, for any curve which can be fitted to observed data we can always find S_y by the definition:

$$S_y = \sqrt{\frac{\Sigma[Y \text{ observed} - Y \text{ computed}]^2}{\text{the number of observations}}}$$

and we can find σ_Y by the methods of Chapter 4. We can therefore define as a measure of correlation based upon any such curve the function [1]

$$\text{measure of correlation based upon any curve} = \sqrt{1 - \frac{S_y^2}{\sigma_Y^2}}$$

where S_y is the standard error of estimate for the curve, and σ_Y is the standard deviation of the given Y measures. This measure of correlation has been called the *index of correlation*, and is denoted by:

$$\rho_{XY}$$

[1] For a test of linearity of regression, see Rietz and others, *op. cit.*, p. 131.

The limits of ρ_{XY} are 0 and 1, a value of 0 indicating no relationship based upon the given function and a value of 1 denoting perfect relationship. In general:

No positive or negative sign should be attached to ρ_{XY}, for the relationship might be positive over part of the range and negative over other parts.[1]

If the given curve is a straight line, then

$$\rho_{XY} = r_{XY}$$

It seems hardly necessary to state that if correlation is measured by ρ_{XY}, the curve to which it applies should always be stated. In the case of r no statement is necessary for it is generally understood that r is based upon linear regression.

EXERCISES

1. Fit a straight line to the data of the following table.

PATIENTS IN NEW YORK STATE HOSPITALS FOR THE INSANE, 1910–1931 [2]

Year	*Number of Patients per 1,000,000 Population*	*Year*	*Number of Patients per 1,000,000 Population*
1910	35.6	1922	40.2
1913	36.9	1925	41.6
1916	38.1	1928	43.3
1919	38.8	1931	45.0

2. In a certain gas-pressure experiment the following results, in which V is the volume corresponding to the pressure p, were obtained. Fit an appropriate curve to the data.

p	33.13	40.44	50.48	59.30	67.08	74.36
V	12.20	9.45	7.55	6.47	5.65	5.07

3. The following table gives the number of divorces per 1,000 marriages during the given years. Fit a curve of the type $Y = aX^2 + bX + c$ to the data. (Choose $X = 0$ at 1910.)

[1] F. C. Mills, *Statistical Methods*, Revised, p. 408.

[2] The data are from *World Almanac*, 1932, p. 534.

DIVORCES IN THE UNITED STATES, 1890–1930 [1]

Year	*Number of Divorces per 1,000 Marriages*	*Year*	*Number of Divorces per 1,000 Marriages*
1890	62	1915	104
1895	67	1920	134
1900	81	1925	148
1905	84	1930	170
1910	88		

4. The following table gives the number of divorces per 1,000 population during the given years. Fit a curve of the type $Y = ab^X$ to these data. What are the computed values for the years 1915 and 1928? The actual values were 1.05 and 1.66.

DIVORCES IN THE UNITED STATES, 1870–1930 [2]

Year	*Number of Divorces per 1,000 Population*	*Year*	*Number of Divorces per 1,000 Population*
1870	0.28	1910	0.90
1880	0.39	1920	1.60
1890	0.53	1930	1.56
1900	0.73		

5. The following table gives the number of grams S of anhydrous ammonium chloride which, dissolved in 100 grams of water, makes a saturated solution of $\theta°$ absolute temperature. Fit an appropriate curve to the data.

θ	273	283	288	293	313	333	353	373
S	29.4	33.3	35.2	37.2	45.8	55.2	65.6	77.3

6. The velocity of water in feet per second in the Mississippi river was measured at various depths, and the ratios, D, of the measured depth to the depth of the river were computed. Fit a curve of the type

[1] The data are from *Statistical Abstract of the United States*, 1932, p. 87.
[2] The data are from *World Almanac*, 1932, p. 444.

$V = aD^2 + bD + c$. Find the computed V when $D = 0.9$. The observed value was $V = 2.9759$.

D	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
V	3.1950	3.2299	3.2532	3.2611	3.2516	3.2282	3.1807	3.1266	3.0594

7. The following table gives the temperature θ of a vessel of cooling water at the end of t minutes. Show that the data may be appropriately fitted to $\theta = c + ab^t$. Find the values of a, b, and c and the computed values of θ.

t	0	1	2	3	5	7	10	15	20
θ	92.0	85.3	79.5	74.5	67.0	60.5	53.5	45.0	39.5

8. For the data of the following table find the exponential curve which appropriately describes the trend. Find the amount in force in 1930 computed by the trend and compare the result with 107.9, which was the actual value.

LIFE INSURANCE IN FORCE IN THE UNITED STATES, 1880–1928 [1]

Year	*Total Amount (billions)*	*Year*	*Total Amount (billions)*
1880	1.6	1915	22.8
1890	4.0	1920	42.3
1900	8.6	1925	71.7
1905	13.4	1928	95.2
1910	16.4	1930	....

9. The indicated horse-power, I, required to drive a ship of displacement D tons at a ten-knot speed is given by the following data. Justify the use of the curve $I = aD^b$. Fit this curve to the data.

D	1,720	2,300	3,200	4,100
I	655	789	1,000	1,164

10. For the data of the following table fit a parabola $Y = aX^3 + bX^2 + cX + d$. (Choose $X = 0$ at 1920.) Use the derived formula to predict the number of failures in 1931, and compare with the actual number, 28.3.

[1] The data are from *Statistical Abstract of the United States*, 1932, p. 283.

COMMERCIAL FAILURES IN THE UNITED STATES, 1910–1930 [1]

Year	*Number of Failures (thousands)*	*Year*	*Number of Failures (thousands)*
1910	12.6	1921	19.7
1911	13.4	1922	23.7
1912	15.5	1923	18.7
1913	16.0	1924	20.6
1914	18.3	1925	21.2
1915	22.2	1926	21.8
1916	17.0	1927	23.1
1917	13.9	1928	23.8
1918	10.0	1929	22.9
1919	6.5	1930	26.4
1920	8.9	1931	...

11. Using the method of Section 63 (p. 237), show that a coefficient of correlation based upon the parabola $Y = a\sqrt{X}$ is $\dfrac{\Sigma Y\sqrt{X}}{\sqrt{\Sigma X \Sigma Y^2}}$.

12. Show that a coefficient of correlation based upon the equilateral hyperbola $xy = a$ is $\dfrac{\Sigma\frac{y}{x}}{\sqrt{\Sigma\frac{1}{x^2}\Sigma y^2}}$.

13. Find the correlation coefficient based upon $xy = a$ for the data of Table 55 (p. 242). Compare your result with the correlation based upon linear regression.

14. Fit an appropriate curve to the data of Exercise 18 (p. 106).

15. What law will satisfactorily represent the following data? Find the values of the constants for the curve selected.

x	y	x	y
2	12.83	8	19.95
3	13.48	9	22.31
4	14.28	10	25.24
5	15.28	11	28.87
6	16.52	12	33.37
7	18.05	13	38.44

[1] The data are from *Statistical Abstract of the United States*, 1932, p. 295.

16. Show that if $Y = aX^2 + bX + c$ is selected to represent a mass of observed data, the equations for the determination of the constants by the method of moments (see Section 59B, p. 219) are those given by (10).

17. The curve $Y = c + aX^b$ passes through the points (2, 11.5), (4, 18.8), and (8, 39.7). Determine a, b, c. Find Y when $X = 5$.

18. The curve $Y = c + ab^X$ passes through the points (2, 5.3), (4, 12.8), and (6, 30.2). Determine a, b, c. Find Y when $X = 3$.

19. Does the point (M_X, M_Y) lie on the parabola $Y = AX^2 + BX + C$ if it is fitted by least squares?

Chapter 11

PERMUTATIONS, COMBINATIONS, AND PROBABILITY

89. INTRODUCTION

In Section 2 of this text we indicated that the solution of a general statistical problem may be divided into four parts: (1) the collection of the data, (2) its organization, (3) its analysis, and (4) the interpretation of the results of the analysis. The earlier chapters have been devoted primarily to the steps of organization and analysis. Given masses of numerical data, we have learned to present them in suitable tabular form, to represent them with graphic devices which emphasize some of the significant features, and to effect numerical analyses the results of which — *when properly interpreted* — present numerical descriptions of the groups.

In our previous discussion we have analyzed a large number of frequency distributions that were derived from several fields: biology, education, sociology, economics, psychology, engineering. Each distribution has presented a specific problem and has been analyzed as a specific problem. We have thus far made but little attempt at generalization. Our method has been the method of science: observation, classification, analysis. We now approach the final step, generalization.

In order to extend our method beyond the analysis of a specific group of data, we are now about to enter upon a study of problems that are rather theoretical. It must not be assumed that because the problems are theoretical they are impractical. We shall find that they are decidedly practical. The first theoretical problem to which we shall give attention will be the development of some general laws to describe frequency distributions, the point binomial and the normal curve, that are usually spoken of as laws of chance. We shall then be in a position to compare theory with observation and to determine whether the differences between theory and ob-

servation are such as may be accounted for by causes other than chance.

The reader has doubtless noted that statistical measurements when gathered in fairly large numbers, although possessing considerable variation, show a quality of orderliness that is at times amazing. As we pass along the scale of measurement of a variable, from the smallest magnitude to the largest, *we find orderliness in the change in the frequency.* Most commonly the frequency, relatively small at the lower end of the range, increases regularly until a maximum is reached in the central portion of the range then diminishes regularly toward zero at the upper end of the range.

This behavior in variation in observed phenomena was first appreciated by the mathematical astronomer, Pierre Simon Laplace, (1749–1827) to the degree that he expressed the behavior by a mathematical function known as the *normal law.* The law had been previously discovered by the mathematician, Abraham de Moivre, (1667–1754) in 1733 as an adventure in pure mathematics to explain the probabilities of games of chance. Carl Friedrich Gauss (1777–1855) made use of it and thus gave it the approval of a very great mathematician. The application of this function to biological variations was soon appreciated by the Belgian scientist, Adolphe Quetelet (1796–1874). The normal law has thus become a foundation stone in the modern statistical structure. That it would someday be used in the solution of biological, social, and economic problems and be invoked in countless investigations of the sciences was of course never dreamed or imagined by its discoverer.

The second theoretical problem, one to which we have alluded several times in the text and to which we shall devote further attention, is what may be termed *the problem of sampling.* We have seen that we may describe a mass of quantitative data as precisely as we please by computing for the data certain statistical constants. These constants give a condensed description of the group in terms of the group's characteristics. Among the tremendous gains realized by this summary, not the least important is this: the summary makes possible the comparison of the characteristics of the individual with the characteristics of the group of which he is a part.

This group that is measured and analyzed is usually a *sample,* a small part of a larger *universe or parent population* that is impossible

or impracticable to measure. Generally, we desire to use the results of the study of the sample to make estimates of the constants that statistically describe the universe. This process is called *statistical inference* or *statistical induction.* It is the problem of *inferring the characteristics of the universe from the characteristics of the sample,* and measuring the reliability of the inferences. This problem may be stated as a question: To what extent is M, σ, or any other constant computed from a sample of N observations randomly made from a universe trustworthy as the mean, standard deviation, or other value of the universe? The answer to this question constitutes what we term *the interpretation of statistical results.*[1]

Statistical induction is literally permeated with questions that relate to the theory of probability, and in order to understand enough of this science to appreciate its widespread applications we shall now introduce the student to the simplest ideas of the theory.

In the present chapter we shall consider certain elementary notions of probability. These notions we shall approach along the avenue of permutations and combinations. We shall undertake to give a thorough and much needed drill in a number of important algebraic concepts which will find repeated application in the chapters that follow. Permutations and combinations will lead us to the point binomial, which in turn will serve to introduce us to the normal probability curve. Thus we start with the notion of a permutation.

90. PERMUTATIONS

A *permutation* is an order or an arrangement of all or a part of a number of things.

Thus, the permutations of the three letters a, b, c, taken all at a time are: $a\,b\,c$, $a\,c\,b$, $b\,a\,c$, $b\,c\,a$, $c\,a\,b$, $c\,b\,a$.

It is seen that 3 objects can be arranged linearly in $3 \cdot 2 = 6$ different ways. We might reason in the following manner. There are 3 places to be filled. The first place can be filled in 3 ways, and with each of these the second place can be filled in 2 ways. Hence the 2 places can be filled in 6 ways. With each of these 6 ways of

[1] So important is this aspect of our study that some writers devote practically their entire treatments to it. For examples, see the texts by R. A. Fisher and by Alan E. Treloar which are listed in Appendix A.

filling the first 2 places there is 1 way of filling the last place, hence $3 \cdot 2 \cdot 1$ ways in all.

This example illustrates the following:

Fundamental Principle. *If one thing can be done in* m *ways, and if, after this is done in one of these ways, a second thing can be done in* n *ways, then the two together can be done in* mn *ways.*

The foregoing principle may be extended into the

Theorem. *If one thing can be done in* m_1 *ways, a second in* m_2 *ways, a third in* m_3 *ways, and so on, the number of different ways in which they can be done when taken all together in the order stated is* $m_1 m_2 m_3 \cdot \cdot \cdot$.

Example 1. How many three-digit numbers can be formed from the digits 1, 2, 3, 4, 5 if each digit is used only once?

The first place can be filled in 5 ways, and after that is done the second place can be filled in 4 ways, and then the third place in 3 ways. Hence, we can form $5 \cdot 4 \cdot 3 = 60$ different numbers of the specified kind.

Example 2. How many three-digit numbers can be formed from the digits 1, 2, 3, 4, 5 if each digit can be repeated?

The first place can be filled in 5 ways, and after that is done the second place can be filled in 5 ways, and then the third place in 5 ways. Hence, we can form $5 \cdot 5 \cdot 5 = 125$ different numbers of the specified kind.

Example 3. How many three-digit *even* numbers can be formed from the digits 1, 2, 3, 4, 5 if each digit is used only once?

The unit's place can be filled in two ways (either with the 2 or 4). The ten's place can then be filled in 4 ways and the hundred's place in 3 ways. In all, there are $3 \cdot 4 \cdot 2 = 24$ numbers of the specified kind.

Example 4. In an introductory course in statistical analysis there are four lecture sections, A, B, C, D, and three laboratory sections, X, Y, Z. In how many ways may a student choose a section in each?

He may choose the lecture section in 4 ways and the laboratory section in 3 ways. He may choose both in $4 \cdot 3 = 12$ ways.

Question. In an election there are three candidates for mayor and four candidates for treasurer. In how many ways can a ballot be marked for both of these offices?

EXERCISES

1. If 2 coins are tossed, in how many ways can they fall?
2. If 3 coins are tossed, in how many ways can they fall?

3. If 2 dice are thrown, in how many ways can they fall?

4. If 2 dice and 3 coins are tossed, in how many ways can they fall?

5. How many signals can be made by hoisting 3 flags if there are 9 different flags from which to choose?

6. In how many different ways can 3 positions be filled by selections from 15 different people?

7. How many four-digit numbers can be formed from the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9?

91. NUMBER OF PERMUTATIONS

In the preceding section we wrote the permutations of the three letters *a*, *b*, *c*, taken all at a time. We may also write the permutations of the same three letters taken two at a time. They are *ab*, *ac*, *ba*, *bc*, *ca*, *cb*.

Now let us consider the general problem: the number of permutations of n things taken r at a time ($r \leqq n$). The number of permutations of n things taken r at a time is denoted by ${}_nP_r$ and is given by the formula:[1]

$$ {}_nP_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} \qquad (1) $$

There are r places to fill and n things from which to choose. The first place may be filled in n ways, the second in $(n-1)$ ways, the third place in $(n-2)$ ways, and so on. The rth place may be filled in $(n-r+1)$ ways. Applying the theorem of Section 90, we immediately have (1).

If all n things are taken n at a time, $n = r$, and we have:

$$ {}_nP_n = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 = n! \qquad (2) $$

Since ${}_nP_n = n! = \dfrac{n!}{(n-n)!} = \dfrac{n!}{0!}$, the use of the second form of (1) when $n = r$ requires that we define $0!$ to equal unity.

It frequently happens that some restrictions are imposed upon the number of permutations we are seeking. Whenever any restriction exists, it is important to consider the restricted groups first. The method is illustrated by the following:

Example. How many six-place numbers can be found from the digits 1, 2, 3, 4, 5, 6, if 3 and 4 are always to occupy the middle two places?

The two digits, 3 and 4, can be arranged in $2!$ ways. The other four digits can be arranged in $4!$ ways. Hence in all $2!\,4! = 48$ numbers.

[1] $n! = 1\cdot 2\cdot 3\cdots n$ is called *factorial n*.

EXERCISES

1. How many different numbers less than 1,000 can be formed from the digits 1, 2, 3, 4, 5, 6?

2. Five persons enter a car in which 8 seats are vacant. In how many ways can they be seated?

3. In how many ways can 10 boys stand in a row when:

(a) a given boy is at a given end? (b) a given boy is at an end? (c) two given boys are always together? (d) two given boys are never together?

4. In how many ways can 3 different algebras and 4 different geometries be arranged on a shelf so that the algebras are always together?

5. Find the number of permutations, P, of the letters $a\ a\ b\ b\ b$ taken 5 at a time. Hint: $P \cdot 2! \cdot 3! = 5!$

6. If P represents the number of distinct permutations of n things, taken all at a time, when, of the n things, there are n_1 alike, n_2 others alike, n_3 others alike, etc., then:

$$P = \frac{n!}{n_1!\,n_2!\,n_3!\,\ldots}$$

7. How many distinct permutations can be made of the letters of the word *attention* taken all at a time?

8. How many distinct permutations of the letters of the word *Mississippi* can be formed taking the letters all at a time?

9. How many ways can ten balls be arranged in a line if 3 are white, 5 are red, and 2 are blue?

92. COMBINATIONS

A group of things or elements without reference to the order of the individuals in the group is called a *combination.*

Thus, the combinations of $a\ b\ c\ d$ taken 3 at a time are $a\ b\ c$, $a\ b\ d$, $a\ c\ d$, $b\ c\ d$. From each combination we can form 3! different permutations, and hence from the 4 combinations we can form $(3!) \cdot 4 = 24$ permutations of 4 letters 3 at a time.

A combination is frequently called a *selection,* whereas a permutation is an *arrangement.*

The number of combinations of n things taken r at a time is denoted by ${}_nC_r$, and is given by the formula:

$${}_nC_r = \frac{{}_nP_r}{r!} \tag{3}$$

For $r!$ permutations can be formed from each combination of r elements; and hence the total number of permutations must be $r!$

times the number of combinations, ${}_nC_r$. That is $r! \cdot {}_nC_r = {}_nP_r$ from which (3) immediately follows.

By applying (1):

$$ {}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} \tag{4} $$

From (4) it follows immediately that:

$$ {}_nC_r = {}_nC_{n-r} \tag{5} $$

The binomial theorem, which is usually written in the form

$$ (a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \cdots + \frac{n(n-1)\cdots(n-r+1)}{r!}a^{n-r}b^r + \cdots + b^n, $$

may be conveniently written

$$ (a+b)^n = a^n + {}_nC_1a^{n-1}b + {}_nC_2a^{n-2}b^2 + \cdots + {}_nC_ra^{n-r}b^r + \cdots + b^n \tag{6} $$

$$ = \sum_{r=0}^{r=n} {}_nC_ra^{n-r}b^r \tag{7} $$

if we define ${}_nC_0$ to be 1.

We shall now illustrate these remarks with a few examples.

Example 1. In how many ways can a committee of 9 be selected from 12 people?

This is evidently a problem of selection, not of arrangement, and the result is evidently:

$$ {}_{12}C_9 = {}_{12}C_3 = \frac{12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3} = 220 $$

Example 2. From 6 men and 5 women, in how many ways can we select a group of 4 men and 3 women?

a. We can select the 4 men from 6 men in ${}_6C_4$ ways.
b. We can select the 3 women from 5 women in ${}_5C_3$ ways.

By the fundamental principle we can do a. and b. in ${}_6C_4 \cdot {}_5C_3 = 150$ ways.

Example 3. From 6 men and 5 women, how many committees of 8 each can be formed when the committee contains at least 3 women?

The conditions of the problem are satisfied if the committee contains:

a. 5 men and 3 women
b. 4 men and 4 women
c. 3 men and 5 women

Therefore the number of possible committees is

$$ {}_6C_5 \cdot {}_5C_3 + {}_6C_4 \cdot {}_5C_4 + {}_6C_3 \cdot {}_5C_5 = 155 $$

It frequently happens that the problem involves both a selection and an arrangement with a limitation upon either. In such problems it is best to consider the two steps separately. A safe procedure is to deal first with the question of the selections (combinations) and then with the arrangements (permutations).

Example 4. How many line-ups are possible in choosing a baseball nine of 5 seniors and 4 juniors from a squad of 8 seniors and 7 juniors, if any man can be used in any position?

The 5 seniors can be selected in ${}_8C_5$ ways, the 4 juniors in ${}_7C_4$ ways. Hence the set of players can be selected in ${}_8C_5 \cdot {}_7C_4$ ways.

Any one set of 9 men can be arranged in 9! ways. Hence the total number of possible line-ups is ${}_8C_5 \cdot {}_7C_4 \cdot 9!$.

EXERCISES

1. Compute ${}_{10}C_2$; ${}_{10}C_8$; ${}_{100}C_{98}$.

2. How many squads of 6 men each can be selected from a squad of 60 men?

3. In how many ways can a committee of 3 teachers and 2 students be selected from 8 teachers and 15 students?

4. How many straight lines are determined from 10 points, no 3 of which are in the same straight line?

5. How many different sums can be made from a cent, a nickel, a dime, a quarter, a half-dollar, and a dollar?

6. From 10 books, in how many ways can a selection of 6 be made: (a) when a specified book is always included? (b) when a specified book is always excluded?

7. Prove that ${}_nC_r + {}_nC_{r-1} = {}_{n+1}C_r$.

8. Out of 6 different consonants and 4 different vowels, how many linear arrangements of letters, each containing 4 consonants and 3 vowels, can be formed?

9. A lodge has 50 members of whom 6 are physicians. In how many ways can a committee of 10 be chosen so as to contain at least 3 physicians?

10. In equation (6) make $a = b = 1$, and show that

$$ {}_nC_1 + {}_nC_2 + \cdots + {}_nC_n = 2^n - 1 $$

11. Solve Exercise 5 above, using Exercise 10.

12. In how many ways can 7 men stand in line so that 2 particular men will not be together?

13. A committee of 7 is to be chosen from 8 Englishmen and 5 Americans. In how many ways can a committee be chosen if it is to contain: (a) just 4 Englishmen? (b) at least 4 Englishmen?

14. Prove: ${}_{n+2}C_{r+1} = {}_nC_{r+1} + 2 \cdot {}_nC_r + {}_nC_{r-1}$.
15. If ${}_nP_r = 110$ and ${}_nC_r = 55$, find n and r.
16. If ${}_nC_4 = {}_nC_2$, find n.
17. If ${}_nC_3 = 10/21({}_nC_5)$, find n.
18. If ${}_{2n}C_{n-1} = 91/24({}_{2n-2}C_n)$, find n.
19. Prove: ${}_nC_1 + 2 \cdot {}_nC_2 + 3 \cdot {}_nC_3 + \cdots + n \cdot {}_nC_n = n(2)^{n-1}$.

93. RELATIVE FREQUENCY: EMPIRICAL PROBABILITY

A box contains 2 white and 3 black balls alike except in color. A ball is drawn at random, the color of it is noted, and then it is replaced in the box. The drawing of the ball and replacing it is called a *trial.* Suppose we make 100 such drawings, mixing the balls thoroughly after each trial, and note that in this sample of 100 drawings we have obtained 38 white and 62 black balls. Then we say 38/100 is the *relative frequency* of white balls and 62/100 is the *relative frequency* of black balls in this set of trials. Suppose that this experiment is repeated and that in the next 100 trials we obtain 42 white balls and 58 black balls. In the second sample of 100 trials the relative frequency of white balls is 42/100 and that of the black balls is 58/100. If the results of the two sample sets are combined, we will then have obtained in the 200 drawings 80 white balls and 120 black balls, and the resulting relative frequencies of white balls and black balls are 80/200 = 2/5 and 120/200 = 3/5 respectively.

In performing experiments of the type described in the preceding paragraph the happening of the event in question is frequently called a *success,* and the nonhappening of the event a *failure.* In the experiments described the drawing of a white ball may be counted a success and that of the black ball a failure. It may be noted that the sum of the relative frequencies of white balls and black balls in every sample drawing is equal to unity. In general if we make $s + f = n$ trials resulting in s successes and f failures we say that:

$$\frac{s}{n} = \text{the relative frequency of the successes}$$

and

$$\frac{f}{n} = \text{the relative frequency of the failures}$$

The sum of the relative frequencies of successes and of failures in any set of trials is equal to:

$$\frac{s}{n} + \frac{f}{n} = \frac{s+f}{n} = \frac{n}{n} = 1$$

The fraction s/n, which we have called the relative frequency of successes in n trials, may be considered an approximate probability derived from observation. If n is large, then, until further knowledge is obtained, s/n may be taken as a good estimate of the probability of success in a given trial. Our confidence in this estimate increases as the number, n, of observed cases increases. If, as n increases indefinitely, the ratio s/n approaches a limiting value, p, this limiting value is called *the probability of a success in one trial.* Hence:

$$p = \lim_{n \to \infty} \frac{s}{n}$$

Thus, if we continue indefinitely the drawing of a ball from a box 2/5 of the contents of which are white balls, we may assume that the relative frequency of white balls would approach 2/5, and we say 2/5 is the probability of obtaining a white ball in a single trial.

The probability that we have thus far discussed as coming from observation is frequently called *empirical probability.*

The empirical method of determining probability is widely used in statistics, pension systems, life insurance, fire insurance, etc. In using the experimental method we shall simply idealize actual experience and assume that the limit of s/n exists, and that, if n is large, s/n is a good estimate of the limit.

EXERCISES

1. In a certain experiment of coin-tossing heads appeared 2,048 times in 4,040 throws. What is the relative frequency of heads? of tails?

2. In an experiment in coin tossing, 7 dimes were thrown 128 times with the following results:

Number of Heads X	*Number of Times* X *Heads Appeared* $f(x)$
0	2
1	8
2	16
3	38
4	43
5	16
6	2
7	3
Total	128

Find the relative frequency of 0 heads; 1 head; etc.

Compute M and σ for this distribution.

3. Among 10,000 people aged 30, 85 deaths occurred in a year. What was the relative frequency of deaths?

4. Out of 1,000 children born in a city in a given year, 514 were boys and 486 were girls. What is the relative frequency of boys among the children that year?

5. As a coöperative exercise for the class, make 1,000 tosses of a coin and keep a record of the number of heads in (a) 10 trials, (b) 100 trials, (c) 250 trials, (d) 1,000 trials. In each case compare the observed relative frequency with the expected relative frequency, 1/2.

94. THEORETICAL RELATIVE FREQUENCY: A PRIORI PROBABILITY

In certain cases, such as games of chance or drawing balls from a bag, the probability may be obtained without collecting statistical data on frequencies. In these cases we make use of certain assumptions that will give us the probability without actually making the trials. For example, if a coin is tossed we assume that it is so constructed and tossed that tails are just as likely to come up as heads, and hence:

$$\text{the probability of heads} = \text{the probability of tails} = \tfrac{1}{2}$$

Similarly, if a bag contains 4 white balls and 6 black balls alike except as to color, and thoroughly mixed, and a ball is drawn at random, the probability of drawing a white ball is 4/10 and the probability of drawing a black ball is 6/10. These illustrations are simple applications of the following:

Definition. *If all the successes and failures can be analyzed into $s + f$ possible ways, each of which is equally likely, and if s of these ways give successes and f of them failures, the probability of success in a single trial is defined as $p = s/(s + f)$ and the probability of failure is defined as $q = f/(s + f)$.*

Example 1. A bag contains 8 black balls and 3 white balls, and a ball is drawn at random. What is the probability of drawing a white ball? a black ball?

If the probability of drawing a white ball is counted a success, we have $s = 3, f = 8, s + f = 11$, and hence $p = \frac{3}{11}$ and $q = \frac{8}{11}$.

Returning to the foregoing definition, we may note that if $p = 0$, the event in question cannot happen or is impossible. If $p = 1$, the event is certain to happen.

Example 2. From a bag containing 8 white balls and 3 black balls, 5 balls are drawn at random. What is the probability that 3 are white and 2 are black?

The total number of balls in the bag is 11. Hence the number of ways of selecting 5 balls from 11 balls is ${}_{11}C_5$. The 3 white balls can be selected from 8 white balls in ${}_8C_3$ ways, and the 2 black balls can be selected from 3 black balls in ${}_3C_2$ ways. Hence $s = {}_8C_3 \cdot {}_3C_2$, and the probability of drawing 3 white and 2 black balls is:

$$p = \frac{{}_8C_3 \cdot {}_3C_2}{{}_{11}C_5} = \frac{4}{11}$$

Example 3. If 5 coins are tossed, what is the probability of obtaining 2 heads and 3 tails?

Five coins may fall in $2^5 = 32$ ways. Two heads may be selected from the 5 in ${}_5C_2 = 10$ ways. Hence the probability is $\frac{10}{32}$.

EXERCISES

1. If a die is thrown, what is the probability that a six will appear? that either a five or a six will appear? that a four, five, or six will appear?

2. If 2 dice are thrown, what is the probability of obtaining a double six?

3. If 2 dice are thrown, what is the probability of obtaining a sum of 11? a sum of 7? What is the most probable sum in a throw of 2 dice?

4. A deck of 52 cards is well shuffled and a card is drawn. What is the probability that it is a queen? an ace or a queen? a heart? a red card?

5. What is the chance of throwing one and only one five in one throw with 2 dice?

6. If 2 dice are thrown, what is the chance of throwing at least one five?

7. If 2 coins are tossed, what is the probability of obtaining 2 heads? 2 tails? 1 head and 1 tail?

8. If 3 coins are tossed, what is the probability of getting 3 heads? 3 tails? 2 heads and 1 tail?

9. What is more likely to happen, a throw of four with 1 die or a throw of six with 2 dice?

10. What is the probability of throwing 2 sixes and 1 five in a single throw with 3 dice?

11. If 12 men stand in line, what is the chance that A and B are next to each other?

12. From a pack of 52 cards, 3 cards are drawn at random. What is the chance that they are all clubs?

13. Prove:

$$_{2n}C_{n+r+1} = {}_{2n}C_{n+r}\left(\frac{n-r}{n+r+1}\right)$$

95. EXPECTATION

The *expected number of occurrences* of an event in n trials is defined as np where p is the probability of occurrence of the event in a single trial.

Thus, if 100 coins are thrown or if 1 coin is thrown 100 times, theoretically we expect 50 heads and 50 tails, for $n = 100$ and $p = q = \frac{1}{2}$.

If a die is rolled 36 times, theoretically we expect an ace to turn up 6 times, for $n = 36$ and $p = \frac{1}{6}$.

If .008 is the probability of death within a year of a man aged 30, the expected number of deaths within a year among 10,000 men of this age would be 80, for $n = 10{,}000$ and $p = .008$.

Question: Two coins are thrown 100 times. What is the expected number of 2 heads? 2 tails? 1 head and 1 tail?

If p is the probability that a person will win a sum of money m, we define his *expectation* as pm.

Thus, if a person is to receive \$32 in case he tosses 4 coins and they all fall heads, the value of his expectation is \$2, for $m = \$32$ and $p = \frac{1}{16}$.

Question: A stake of \$24 is made contingent upon getting a sum greater than 10 in a single throw with 2 dice. What is the value of the expectation?

96. SOME ELEMENTARY THEOREMS

A. Mutually Exclusive Events. Two or more events are said to be mutually exclusive when the occurrence of any one of them excludes the occurrence of any other. Thus, in the toss of a coin the appearance of heads and the appearance of tails are mutually exclusive. Also, if a bag contains white and black balls and a ball is drawn, the drawing of a white ball and the drawing of a black ball are mutually exclusive events.

Theorem. *If $p_1, p_2, \ldots, p_r$ are the separate probabilities of r mutually exclusive events, the probability P, that one of these events will happen in a single trial is the sum of the probabilities of the separate events. That is:*

$$P = p_1 + p_2 + \cdots + p_r \tag{8}$$

By the definition in the preceding section, out of n trials in which all of the events are in question, the r events are expected to occur $p_1n, p_2n, \ldots, p_rn$ times respectively. Since only one of these events can occur on a given trial, it follows that out of n trials one or another of the r events will occur $(p_1n + p_2n + \cdots + p_rn)$ or $(p_1 + p_2 + \cdots + p_r)n$ times. That is, the total probability that one of the events will occur on a given trial is:

$$P = \frac{(p_1 + p_2 + \cdots + p_r)n}{n} = p_1 + p_2 + \cdots + p_r$$

When two mutually exclusive events are in question, the probabilities are frequently called *either or* probabilities. Thus, if a die is thrown, the probability of *either* an ace *or* a deuce is $\frac{1}{6} + \frac{1}{6}$ or $\frac{1}{3}$.

B. Independent Events. Two or more events are *dependent* or *independent* according as the occurrence of any one of them does or does not affect the occurrence of the others. Thus, if A tosses a coin and B throws a die, the tossing of heads by A and the throwing of a deuce by B are independent events. However, if a bag contains a mixture of white and black balls and a ball is drawn and not returned to the bag, the probabilities in a second drawing will be dependent upon the first event.

Theorem. *If $p_1, p_2, \ldots, p_r$ are the separate probabilities of r independent events, the probability P, that they all occur on a given trial when all of them are in question, is the product of their separate probabilities. That is:*

$$P = p_1p_2p_3 \ldots p_r \tag{9}$$

By the definition of the preceding section, out of n trials in which all of the events are in question, the first event is expected to occur p_1n times. Out of this number, p_1n, the second event is expected to occur $p_2(p_1n) = np_1p_2$ times. That is, both are expected to occur np_1p_2 times in n trials. Continuing this process, it is seen that out

of n trials all of the r events are expected to occur $np_1p_2p_3 \cdot \cdot \cdot p_r$ times. Hence:

$$P = \frac{np_1p_2p_3 \cdot \cdot \cdot p_r}{n} = p_1p_2p_3 \cdot \cdot \cdot p_r$$

Example 1. If A tosses a coin and B throws a die, what is the probability that A will toss heads and B will throw a deuce?

The probability that A will toss heads is $\frac{1}{2}$ and the probability that B will throw a deuce is $\frac{1}{6}$. Since the two events are independent, the probability that both events will occur is $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$.

Example 2. If a coin is tossed 3 times, what is the probability of heads every time?

The probability of heads on any throw is $\frac{1}{2}$. Hence for the 3 throws, since they are independent, $P = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

When two independent events are in question, the probabilities are frequently called *both and* probabilities. Thus in Example 1 if the tossing of heads by A is event E_1 and the throwing a deuce by B is event E_2, then the probability that *both* E_1 *and* E_2 occur is $\frac{1}{12}$.

In Example 1, what is the probability that *either* A will toss heads *or* B will throw a deuce?

Corollary. *If* $p_1, p_2, \ldots, p_r$ *are the separate probabilities of* r *independent events, the probability that they will all fail on a given occasion is*

$$(1 - p_1)(1 - p_2) \cdot \cdot \cdot (1 - p_r) \tag{10}$$

and the probability that the first k *events will occur and the remainder fail is:*

$$p_1 \cdot p_2 \cdot \cdot \cdot p_k(1 - p_{k+1}) \cdot \cdot \cdot (1 - p_r) \tag{11}$$

C. Dependent Events. The following theorem for dependent events may be proved by an analogous method to that used for independent events.

Theorem. *If the probability of a first of* r *events is* p_1, *and if, after this has occurred, the probability of a second event is* p_2, *and if, after the first and second events have occurred, the probability of a third event is* p_3, *and so on, then the probability* P, *that the events will occur in the order specified is:*

$$P = p_1p_2p_3 \cdot \cdot \cdot p_r \tag{12}$$

EXERCISES

1. If 5 balls are drawn from a bag containing 6 red and 9 white balls, what is the probability: (a) that all will be red? (b) that 3 will be red and 2 white?

2. A draws 3 cards from a well-shuffled pack and simultaneously B tosses a coin. What is the probability of 3 aces and 1 head?

3. If 4 coins are tossed, what is the probability that all will fall tails?

4. A, B, and C go bird-hunting. A has a record of 1 bird out of 2, B gets 2 out of 3, and C gets 3 out of 4. What is the probability that they will kill a bird at which all shoot simultaneously? Hint: What is the probability that all 3 miss?

5. If the probability that A will die within a year is $\frac{2}{10}$ and the probability that B will die within a year is $\frac{3}{10}$, what is the probability that: (a) both A and B will die within a year? (b) both A and B will live a year? (c) one life will fail within a year?

6. The probability that A will solve a problem is $\frac{1}{3}$ and that B will solve it is $\frac{2}{3}$. What is the probability that if A and B try the problem it will be solved?

7. In a single throw of 2 dice what is the chance that neither doublets nor seven will appear?

97. REPEATED TRIALS

As we proceed into the text the observing student will be amazed at the importance of the theory of the probability of repeated trials in the theory of statistics. This is, of course, due primarily to the fact that much of statistical data is a kind of repeated measurement.

In order to familiarize ourselves with the method of proof of the general theorem of this section, let us consider something simple.

Example. What is the probability of throwing 2 aces in 4 throws of a die?

The conditions of the problem are met if in the first 2 throws we obtain aces and in the next 2 throws not-aces; or if in the first throw we get ace, the second throw not-ace, the third throw ace, and the fourth throw not-ace; and so on. We shall illustrate the possibilities symbolically as follows:

$$A_1A_2 - -, A_1 - A_3 -, A_1 - - A_4, - A_2A_3 - , - A_2 - A_4, - - A_3A_4$$

Considering the first case, the probability of throwing an ace on any throw is $\frac{1}{6}$. The probability of not throwing an ace on any throw is $\frac{5}{6}$. Hence the probability of throwing an ace on the first and second throws and not throwing an ace on the two remaining throws is $(\frac{1}{6})^2(\frac{5}{6})^2$.

In the second case, the probability of events occurring as the symbol above indicates is $(\frac{1}{6})(\frac{5}{6})(\frac{1}{6})(\frac{5}{6}) = (\frac{1}{6})^2(\frac{5}{6})^2$.

The remaining cases may be treated in a similar manner, and in each instance the result for any specified set is $(\frac{1}{6})^2(\frac{5}{6})^2$. Now it is evident that the 2 aces can be selected from the 4 possible aces in ${}_4C_2 = 6$ ways. Since the 6 cases are mutually exclusive, the chance that one or the other of the specified cases occurs is $6(\frac{1}{6})^2(\frac{5}{6})^2 = \frac{150}{1296}$.

Let us now consider an important theorem.

Theorem of Repeated Trials. *If p is the probability of the success of an event in a single trial and q is the probability of its failure, $(p + q = 1)$, then the probability P_r that the event will succeed exactly r times in n trials is:*[1]

$$P_r = {}_nC_r p^r q^{n-r} \tag{13}$$

For the probability that the event will succeed in each of r specified trials and will fail in the remaining $(n - r)$ trials is, by (11), $p^r q^{n-r}$. Further, it is possible for the r successes to occur out of n trials in ${}_nC_r$ different ways. These ways being mutually exclusive, by (8) the probability in question is $P_r = {}_nC_r p^r q^{n-r}$.

The various probabilities are indicated in the following table:

TABLE 87. VALUES OF P_r FOR VARIOUS VALUES OF r

r	P_r	*The Probability That in n Trials There Will Be*
n	p^n	n successes, 0 failures
$n-1$	${}_nC_1 p^{n-1}q$	$n-1$ " , 1 "
$n-2$	${}_nC_2 p^{n-2}q^2$	$n-2$ " , 2 "
.....	.....	....., ...
.....	.....	....., ...
$n-r$	${}_nC_r p^{n-r}q^r$	$n-r$ successes, r failures
.....	.....	..... " , ... "
r	${}_nC_r p^r q^{n-r}$	r " , $n-r$ "
.....	.....	..... " , ... "
2	${}_nC_2 p^2 q^{n-2}$	2 " , $n-2$ "
1	${}_nC_1 pq^{n-1}$	1 " , $n-1$ "
0	q^n	0 " , n "
Total	$(p+q)^n = 1$	

From Table 87 we have at once the following:

Corollary. *The probability that an event will succeed at least r times in n trials is $P_r + P_{r+1} + \cdots + P_n$, that is:*

$$\sum_r^n P_r = p^n + {}_nC_1 p^{n-1}q + {}_nC_2 p^{n-2}q^2 + \cdots + {}_nC_r p^r q^{n-r} \tag{14}$$

It will be noted that (14) consists of the first $(n - r + 1)$ terms of the expansion $(p + q)^n$.

[1] It will be noted that (13) is the $(n - r + 1)$th term of the expansion $(p + q)^n$ and the $(r + 1)$th term of the expansion $(q + p)^n$.

Example. An urn contains 12 white and 24 black balls. What is the probability that, in 10 drawings with replacements, exactly 6 white balls are drawn?

We have:
$$p = \frac{12}{36} = \frac{1}{3} \qquad q = \frac{24}{36} = \frac{2}{3},$$
$$n = 10, r = 6 \qquad n - r = 4,$$

hence:
$$P_6 = {}_{10}C_6\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^4 = \frac{3360}{3^{10}}$$

Since the computation of P_r in (13) involves the computation of ${}_nC_r = \dfrac{n!}{r!(n-r)!}$, we may naturally wonder what can be done when n and r are so large that the labor of evaluating $n!$, $r!$, and $(n-r)!$ becomes tedious if not prohibitive. At present we can recommend two alternatives. If tables of the logarithms of factorial n are at hand,[1] then P_r can be conveniently computed by logarithms. If such tables are not at hand, approximate results can be found by applying Stirling's formula, namely:

$$n! = e^{-n}n^n\sqrt{2\pi n} \tag{15}$$

The derivation of this formula depends upon the calculus and is therefore beyond the scope of this text.[2] For large values of n, it gives satisfactory results.

Consider the following:

Example. An urn contains 2 white and 3 black balls. What is the probability that, in 500 drawings with replacements, exactly 200 white balls will be drawn?

Solution:
$$n = 500, \; p = \frac{2}{5}, \; q = \frac{3}{5}$$
$$P_{200} = {}_{500}C_{200}\left(\frac{2}{5}\right)^{200}\left(\frac{3}{5}\right)^{300}$$
$$= \frac{500!}{200!\,300!}\left(\frac{2}{5}\right)^{200}\left(\frac{3}{5}\right)^{300}$$

[1] An excellent set of tables is J. W. Glover, *Tables of Applied Mathematics*, George Wahr, Ann Arbor, Michigan, 1923.

[2] See J. L. Coolidge, *An Introduction to Mathematical Probability*, p. 38, for a derivation.

Applying Stirling's formula:

$$P_{200} = \frac{500^{500}e^{-500}\sqrt{2\pi \cdot 500}}{200^{200}e^{-200}\sqrt{2\pi \cdot 200}\,300^{300}e^{-300}\sqrt{2\pi \cdot 300}}\left(\frac{2}{5}\right)^{200}\left(\frac{3}{5}\right)^{300}$$

$$= \frac{500^{500}\sqrt{5}}{200^{200}300^{300}10\sqrt{12\pi}}\left(\frac{2}{5}\right)^{200}\left(\frac{3}{5}\right)^{300}$$

$$= \frac{5^{500}100^{500}\sqrt{5}\,2^{200}3^{300}}{2^{200}100^{200}3^{300}100^{300}10\sqrt{12\pi}\,5^{500}}$$

$$= \frac{\sqrt{5}}{10\sqrt{12\pi}} = .036.$$

If Glover's *Tables* are used with logarithms the result is:

$$P_{200} = .041$$

EXERCISES

1. A coin is tossed 7 times, or 7 coins are tossed one time. Find the probability of exactly: (a) no heads, (b) 1 head, (c) 2 heads, etc. to 7 heads.

2. Seven coins are tossed 128 times. Using the Definition in Section 95 (p. 374), and the probabilities of the last exercise (1), find the expected number of times of 0 heads, 1 head, 2 heads, etc. to 7 heads. Compare the results with those of Exercise 2 (p. 371).

3. If a die is thrown 6 times or if 6 dice are thrown 1 time, what is the probability of obtaining: (a) exactly 2 aces? (b) at least 3 aces?

4. Find the probability of throwing with a single die a deuce at least once in 5 trials.

5. Prove that the probability that an event will succeed at least once in n trials is $(1 - q^n)$.

6. In tossing 10 coins, what is the probability of obtaining at least 8 heads?

7. A man whose batting average is $\frac{3}{10}$ will bat 4 times in a game. What is the probability that he will get (a) exactly 2 hits? (b) at least 2 hits?

8. According to the American Experience Table of Mortality, out of 100,000 persons living at the age of 10 years, 91,914 are living at the age of 21 years. Each of 7 boys is now 10 years old. What is the probability that exactly 5 of them will live to be 21?

9. A bag contains 4 white and 2 black balls. Five balls are drawn with replacements. What is the probability: (a) that exactly 3 are white? (b) that at least 3 are black?

10. What is the probability of throwing at least 3 sevens in 5 throws with a pair of dice?

11. How many throws with 2 dice will be required in order that the probability of obtaining a double six at least once will have the value $\frac{1}{2}$?

Hint: If $\frac{1}{2} = 1 - (\frac{35}{36})^n$, find n.

12. At an old men's home are 5 seventy-year old men. Find the probability that (a) exactly 2 of them will die within a year, (b) that a specified 2 of them will die within a year, (c) that at least 2 of them will die within a year. The probability that a man aged 70 lives a year is $p_{70} = 0.94$. Hence $q_{70} = 0.06$.

13. Hospital records show that 5 per cent of cases of a certain disease are fatal. Five patients are admitted with this disease. Find the probability (a) that all will recover, (b) that exactly 3 will die, (c) that at least 3 will die.

14. A marksman is able, on the average, to hit a target 950 times out of 1,000. Find the probability that he will obtain (a) exactly 9 hits in 10 shots, (b) exactly 10 hits in 10 shots, (c) either 9 or 10 hits in 10 shots, (d) at least 5 hits in 10 shots. Express symbolically.

15. The registrar's records show that 10 per cent of the students fail a certain course. The present enrolment in the course is 25. What is the probability that 5 will fail?

16. In the long run 3 vessels out of every 100 are sunk. If 10 vessels are out, what is the probability (a) that exactly 6 will arrive safely? (b) that at least 6 will arrive safely? Express symbolically.

17. A batch of 1,000 electric bulbs was tested and found to be 5 per cent bad. If another batch of 100 lamps is manufactured under similar conditions, what is the probability that not more than 10 per cent will be defective? Give the result symbolically.

18. The American Experience Mortality Table states that for an individual aged 25 the probability of survival another year is $p = 0.992$. What probabilities are expressed by the following:

(a) ${}_{1000}C_{200}(.992)^{800}(.008)^{200}$? (b) $\sum_{r=700}^{900} {}_{1000}C_r(.992)^{1000-r}(.008)^r$?

19. A, B, and C are three marksmen. A's record is 4 hits in 5 shots, B's record is 3 hits in 4 shots, and C's record is 2 hits in 3 shots. They fire simultaneously. What is the probability that at least 2 shots hit?

20. Of 7 dates picked at random, what is the probability that (a) exactly 5 are Sundays, (b) at least 5 are Sundays, (c) the first 5 but no others are Sundays?

21. A can hit a target 4 times in 5 shots; B, three times in four shots. They fire a volley. What is the probability (a) that at least two shots hit? (b) that at least one shot hits?

22. A student takes a true-false test consisting of 10 questions and guesses at the answers. Assuming he is equally likely to answer incorrectly as correctly on each question, find the probability (a) that he will answer all the questions correctly, (b) that he will answer half of them correctly, (c) that he will answer 80 per cent or more of them correctly.

23. In the long run a child under one year of age who is attacked by whooping cough has about a fifty-fifty chance of recovery. If 10 children under one year of age are attacked by this disease,

(a) what is the expected number of deaths?
(b) what is the probability that the expected number will die?
(c) what is the probability that 8 or more recover?

24. In how many throws with a single die will it be an even chance that "1" turns up at least once?

25. If 2 dice are thrown, what is the probability of obtaining a total of 7?

Hint: The number of ways of obtaining a total of 7 is the coefficient of x^7 in $(x + x^2 + x^3 + x^4 + x^5 + x^6)^2$, or of x^5 in $(1 + x + x^2 + x^3 + x^4 + x^5)^2$, or of x^5 in $\left(\frac{1 - x^6}{1 - x}\right)^2 = (1 - x^6)^2(1 - x)^{-2}$.

26. If three dice are thrown what is the probability of obtaining a total of 10?

Hint: The number of ways of obtaining a total of 10 is the coefficient of x^{10} in $(x + x^2 + x^3 + x^4 + x^5 + x^6)^3$.

27. If three dice are thrown what is the probability of obtaining a total of 8?

28. Three dice are thrown. Show that the probability of obtaining a total of 4 is equal to the probability of obtaining a total of 17.

Chapter 12

THE POINT BINOMIAL AND THE NORMAL CURVE

98. INTRODUCTION

In the preceding chapter considerable emphasis was placed upon what is essentially the

Theorem. *If* p *is the probability of the success of an event in a single trial and* q *is the probability of its failure* $(p + q = 1)$, *then the successive terms of the binomial expansion*

$$(q+p)^n = q^n + {}_nC_1 q^{n-1}p + {}_nC_2 q^{n-2}p^2 + \cdots + {}_nC_X q^{n-X}p^X + \cdots + p^n \qquad (1)$$

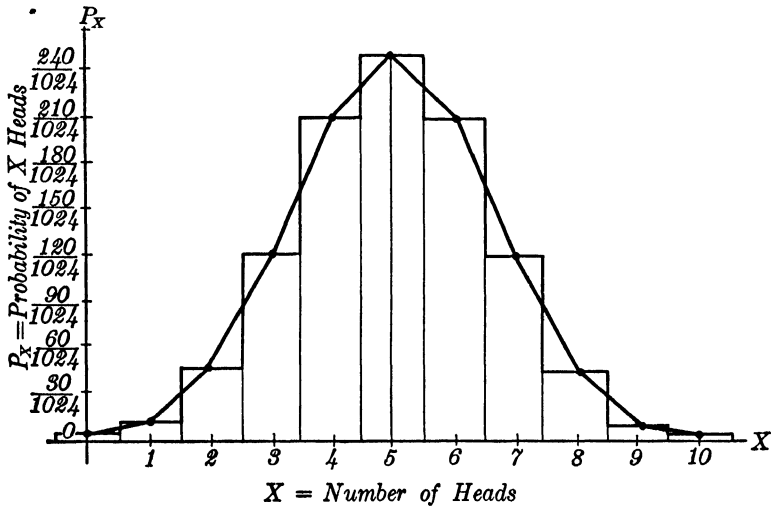
give the respective probabilities that in n *trials the event will succeed in* $0, 1, 2, \ldots, X, \ldots, n$ *times.*

It should be especially noted that the general term

$$P_X = {}_nC_X q^{n-X}p^X$$

gives the probability that the event will succeed *exactly* X times in n trials.

FIGURE 51



Example 1. If a coin is tossed 10 times (or if 10 coins are tossed 1 time), the successive terms of the expansion

$$(\tfrac{1}{2}+\tfrac{1}{2})^{10}=\tfrac{1}{1024}[1+10+45+120+210+252+210+120+45+10+1]$$

give the probabilities of 0 heads; 1 head, 9 tails; 2 heads, 8 tails; etc.

If the terms of this expansion $(\frac{1}{2}+\frac{1}{2})^{10}$ be plotted as ordinates at unit distances along the horizontal axis, it will be noted that the points are *symmetrically* distributed about the vertical through $X = 5$ (Fig. 51). It will be shown later that the symmetry is due to the fact that $p = q = \frac{1}{2}$.

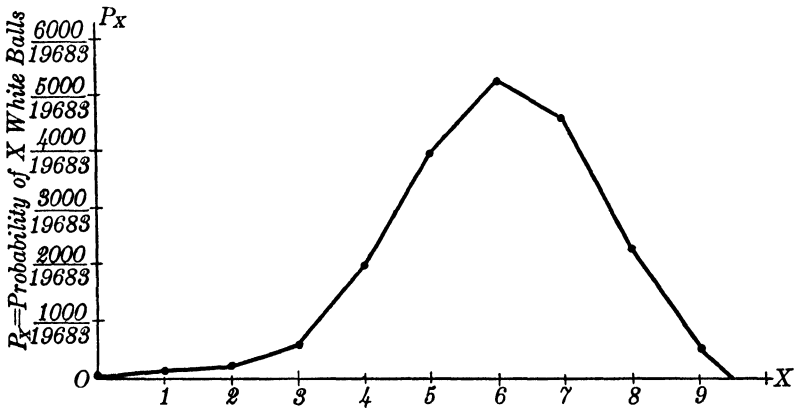
Example 2. Nine balls are drawn singly, with replacements, from a bag containing white and black balls in the ratio of 2 to 1. If the probability of drawing a white ball is counted a success, $p = \frac{2}{3}$, $q = \frac{1}{3}$, and the successive terms of the expansion

$$(\tfrac{1}{3}+\tfrac{2}{3})^{9}=\tfrac{1}{19683}[1+18+144+672+2016+4032+5376+4608+2304+512]$$

give the probabilities of drawing 0 white balls; 1 white, 8 black balls; 2 white, 7 black balls, etc.

If these probabilities be plotted as ordinates, as the figure below indicates, it is noted that the points are not symmetrically distributed. That the skewness here is due to the inequality of p and q will be shown in the succeeding section.

FIGURE 52



X = Number of White Balls

99. CHARACTERISTICS OF THE POINT BINOMIAL

It has been observed in the preceding section that the binomial distributions possess certain geometrical similarities to the observed

distributions studied in Chapters 3, 4, and 5; namely, they are relatively low at the extremes and rise to a single mode near the center. These similarities are so striking that we shall use the binomial distribution as a theoretical or approximate distribution with which to compare distributions of observed data. That is, we shall use the point binomial as the first approximation to distributions of observed data.

We shall need to characterize the binomial distribution, as we have other distributions, by computing measures of central tendency, dispersion, skewness, etcetera. Having computed these constants for the theoretical distribution, we shall apply the results to distributions of observed data for purposes of comparison and generalization.

A. The Mode. Since the sum of the terms of $(q + p)^n$ equals unity and the extreme terms are usually smaller than those near the center, it would seem that for a determinate value of X, say $X = \alpha$,

$$P_X = {}_nC_X q^{n-X} p^X$$

will have a maximum.

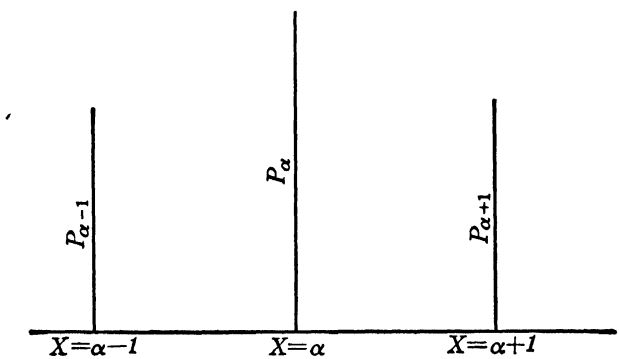
In order for P_X to be a maximum for $X = \alpha$, we must have

a. $P_{\alpha-1} \leqq P_\alpha$
b. $P_\alpha \geqq P_{\alpha+1}$

that is, we must have

a. ${}_nC_{\alpha-1} q^{n-\alpha+1} p^{\alpha-1} \leqq {}_nC_\alpha q^{n-\alpha} p^\alpha$
b. ${}_nC_\alpha q^{n-\alpha} p^\alpha \geqq {}_nC_{\alpha+1} q^{n-\alpha-1} p^{\alpha+1}$

FIGURE 53



Using the relation

$$ {}_nC_\alpha = \frac{n!}{\alpha!(n-\alpha)!}, $$

the first inequality reduces to

$$ \alpha \leqq np + p = (n+1)p $$

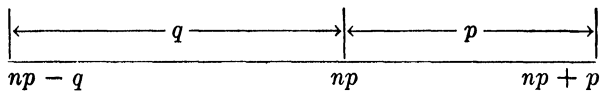
and the second reduces to:

$$ \alpha \geqq np - q $$

That is, α satisfies the double inequality:

$$ np - q \leqq \alpha \leqq np + p \tag{2} $$

DIAGRAM 13



If $np + p$ is an integer, so is $np - q$ the next lower integer. In this case two values of α satisfy (2) since α is necessarily integral. They are $\alpha = np + p$ and $\alpha = np - q$. [See Exercise 6, p. 390.] Thus:

$$ (\tfrac{1}{2} + \tfrac{1}{2})^3 = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8} $$

has two equal terms which are larger than any other terms, one at $\alpha = \frac{3}{2} + \frac{1}{2} = 2$, and the other at $\alpha = \frac{3}{2} - \frac{1}{2} = 1$. Recalling that P_α is the $(\alpha + 1)$th term of (1), the second and the third terms are two equal terms which are larger than any other terms. Similarly,

$$ (\tfrac{1}{3} + \tfrac{2}{3})^5 = \tfrac{1}{243}[1 + 10 + 40 + 80 + 80 + 32] $$

has two equal terms which are larger than any other terms, since $np + p = 5(\frac{2}{3}) + \frac{2}{3}$ is an integer. They are the fourth and the fifth terms.

If $np + p$ or $(n + 1)p$ is fractional, so is $np - q$ since $np - q = np - (1 - p) = (n + 1)p - 1$. By (2), α must be the integer lying between them. Since there is only one such integer, it must be α. Thus $(\frac{1}{3} + \frac{2}{3})^6$ has only one maximum term. For in this case $np + p = 6(\frac{2}{3}) + \frac{2}{3} = 4\frac{2}{3}$, and $np - q = 3\frac{2}{3}$. Hence $\alpha = 4$, and the fifth term

$$ P_4 = {}_6C_4(\tfrac{1}{3})^2(\tfrac{2}{3})^4 = \tfrac{240}{729} $$

is the maximum term. The entire expansion is:

$$ (\tfrac{1}{3} + \tfrac{2}{3})^6 = \tfrac{1}{729}[1 + 12 + 60 + 160 + 240 + 192 + 64] $$

We may summarize these results into the following:

Theorem. *If $np + p$ or $np - q$ is fractional, P_X has one maximum term for X equal to the greatest integer in $np + p$. If $np + p$ or $np - q$ is integral, P_X has two equal terms which are larger than any other terms. They occur when X equals $np + p$ and $np - q$.*

If n is large and np relatively large when compared with p and q, np closely approximates $np + p$ and $np - q$. In this case we call np the *expected number of successes* and $n - np = n(1 - p) = nq$ the *expected number of failures* of the event in n trials.

The probability of np successes is given by $P_{np} = {}_nC_{np}p^{np}q^{nq}$ which, upon applying Stirling's formula (p. 379), reduces to

$$P_{np} = \frac{1}{\sqrt{2\pi npq}},$$

a very small number. That is, the probability of obtaining the expected number of successes (or failures) is a very improbable event.

B. The Mean, the Dispersion, the Skewness. The computation of M, σ, α_3, and α_4 is greatly facilitated by the preparation of Table 88 in which $f(X) = {}_nC_Xq^{n-X}p^X$ indicates the ordinate corresponding to the given abscissa, X.

TABLE 88

X (1)	$f(X) = {}_nC_Xq^{n-X}p^X$ (2)	$Xf(X)$ (3)	$X(X-1)f(X)$ (4)	$X(X-1)(X-2)f(X)$ (5)
0	q^n	0	0	0
1	$nq^{n-1}p$	$nq^{n-1}p$	0	0
2	$\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2$	$n(n-1)q^{n-2}p^2$	$n(n-1)q^{n-2}p^2$	0
3	$\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}q^{n-3}p^3$	$\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$	$\frac{n(n-1)(n-2)}{1}q^{n-3}p^3$	$n(n-1)(n-2)q^{n-3}p^3$
	. . .	. . .	. . .	. . .
$n-1$	nqp^{n-1}	$n(n-1)qp^{n-1}$	$n(n-1)(n-2)qp^{n-1}$	$n(n-1)(n-2)(n-3)qp^{n-1}$
n	p^n	np^n	$n(n-1)p^n$	$n(n-1)(n-2)p^n$
Total	$(q+p)^n$	$np(q+p)^{n-1} = np$	$n(n-1)p^2(q+p)^{n-2} = n(n-1)p^2$	$n(n-1)(n-2)p^3(q+p)^{n-3} = n(n-1)(n-2)p^3$

The total of column (2) of the table, $\Sigma f(X)$, is obviously unity since:

$$\Sigma f(X) = q^n + nq^{n-1}p + \frac{n(n-1)}{1\cdot 2}q^{n-2}p^2 + \cdots + p^n = (q+p)^n = 1$$

The total of column (3), $\Sigma Xf(X)$, is easily recognized if one takes the common factor, np, out of every term. Thus:

$$\Sigma Xf(X) = np\left[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{1 \cdot 2}q^{n-3}p^2 + \cdots + p^{n-1}\right]$$
$$= np(q + p)^{n-1} = np$$

Likewise columns (4) and (5) may be reduced to:

$$\Sigma X(X-1)f(X) = n(n-1)p^2(q+p)^{n-2} = n(n-1)p^2$$
$$\Sigma X(X-1)(X-2)f(X) = n(n-1)(n-2)p^3(q+p)^{n-3}$$
$$= n(n-1)(n-2)p^3$$

We therefore have:

$$M = \frac{\Sigma Xf(X)}{\Sigma f(X)} = \frac{np}{1} = np$$

Since

$$\Sigma X(X-1)f(X) = \Sigma X^2 f(X) - \Sigma Xf(X) = n(n-1)p^2$$

(from Table 88), we have:

$$\Sigma X^2 f(X) = n(n-1)p^2 + \Sigma Xf(X) = n(n-1)p^2 + np$$
$$= n^2p^2 + np(1-p) = n^2p^2 + npq = M^2 + npq$$

But:

$$\sigma^2 = \frac{\Sigma X^2 f(X)}{\Sigma f(X)} - M^2$$

Hence:

$$\sigma^2 = \frac{M^2 + npq}{1} - M^2 = npq$$

or

$$\sigma = \sqrt{npq}$$

Similarly:

$$\Sigma X(X-1)(X-2)f(X) = \Sigma X^3 f(X) - 3\Sigma X^2 f(X) + 2\Sigma Xf(X)$$
$$= \Sigma X^3 f(X) - 3(n^2p^2 + npq) + 2np$$
$$= \Sigma X^3 f(X) - 3n^2p^2 - 3npq + 2np$$
$$= n(n-1)(n-2)p^3 \quad \text{(from Table 88)}$$

Therefore:

$$\Sigma X^3 f(X) = n^3p^3 + 3n^2p^2(1-p) + 3npq + 2np^3 - 2np$$
$$= n^3p^3 + 3n^2p^2q + 3npq + 2np^3 - 2np$$

Using the formula for ν_3 given on page 162 for the case in which $w = 1$, $h = 0$, $N = \Sigma f(X) = 1$, we have:

$$\nu_3 = \Sigma X^3 f(X) - 3\Sigma X^2 f(X)M + 2M^3$$

Substituting the values given above:

$$\nu_3 = 3npq + 2np^3 - 2np = np(3q + 2p^2 - 2)$$
$$= np(3 - 3p + 2p^2 - 2) = np(1-p)(1-2p)$$

and finally

$$\nu_3 \text{ or } \mu_3 = npq(1 - 2p) = npq(q - p)$$

Hence:

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{npq(q - p)}{(npq)^{\frac{3}{2}}} = \frac{q - p}{\sigma}$$

Collecting these results we have

$$\left.\begin{aligned} M &= np \\ \sigma &= \sqrt{npq} \\ \alpha_3 &= \frac{q - p}{\sigma} \end{aligned}\right\} \tag{3}$$

where the positive direction is that of increasing X.

The equation $M = np$ shows that for the point binomial, $(q + p)^n$, the mean value is equal to the expected value. The value of α_3 shows that the skewness is positive when p is less than q, is negative when p is greater than q, and is zero when p equals q.

In the next list of exercises we ask the student to show that when n becomes infinite in the point binomial $(q + p)^n$, the skewness α_3 approaches zero, and the kurtosis $(\alpha_4 - 3)$ also approaches zero. We have stated in Chapter 5 that, for a *normal distribution*, α_3 equals zero and α_4 equals 3. Thus we see that as n increases the moments (of order less than 5) of the point binomial approach the same moments of the normal distribution.

VALUES OF α_3 AND α_4 FOR $(.98 + .02)^n$

n	α_3	α_4
100	.68	3.45
200	.48	3.23
300	.40	3.15
400	.34	3.11
500	.31	3.09
600	.28	3.075
700	.26	3.06
800	.24	3.06
900	.23	3.05
1000	.21	3.045

The rapidity with which α_3 approaches zero and α_4 approaches 3 as n increases, even for the case where p is extremely small, is shown by the accompanying table.

EXERCISES

1. Plot the histograms and the frequency polygons for the binomials following. Find for each binomial the M_o, M, σ, and α_3.

a. $(\frac{1}{6} + \frac{5}{6})^6$ b. $(\frac{5}{6} + \frac{1}{6})^6$ c. $(\frac{1}{10} + \frac{9}{10})^4$ d. $(\frac{2}{5} + \frac{3}{5})^4$

2. By extending Table 88 show that:

$$\Sigma X(X-1)(X-2)(X-3)f(X) = n(n-1)(n-2)(n-3)p^4$$

3. Using the value of ν_4 given on page 162, show that for the point binomial:

$$\mu_4 = \nu_4 = npq[1 + 3pq(n-2)]$$

4. Show that α_4 for the point binomial is given by:

$$\alpha_4 = 3 + \frac{1}{\sigma^2} - \frac{6}{n}$$

5. Show that

$$P_{np} = \frac{n!}{(np)!\,(nq)!} q^{nq} p^{np}$$

reduces to

$$P_{np} = \frac{1}{\sqrt{2\pi npq}} = \frac{1}{\sigma\sqrt{2\pi}}$$

when Stirling's formula is applied.

6. Show that if $np - q$ is an integer

$$P_{np-q} = P_{np+p}$$

Hint. (1) Let $np - q = k$, then $np + p = k + 1$ from which obtain $(n-k)/(k+1) = q/p$. (2) Show that $P_{np+p} = P_{k+1} = (n-k)/(k+1) \cdot p/q \cdot P_k$. (3) Combine the results of (1) and (2).

7. Show that as n becomes infinite, α_3 equals zero and α_4 equals 3.

8. Verify the values of the table for $(.9 + .1)^n$.

n	α_3	α_4
100	.2667	3.0511
200	.1886	3.0256
1000	.0843	3.0051

The following exercises are for students of the calculus.

9. Show that

$$\frac{d^i}{dx^i}(q + px)^n \Big]_{x=1}, \quad (i = 0, 1, 2, 3)$$

give the totals of columns 2, 3, 4, and 5 of Table 88.

10. Show that the moments of $(q + p)^n$ given by

$$\mu_i = \frac{d^i}{dx^i}(pe^{qx} + qe^{-px})^n \Big]_{x=0}, \quad (i = 1, 2, 3, 4)$$

are the same as we have given in the text. This relationship was given by Karl Pearson in *Biometrika*, Vol. XII, p. 270.

11. The moments of $(q + p)^n$ can be obtained from

$$\mu_{i+1} = pq\left[ni\mu_{i-1} - \frac{d\mu_i}{dq}\right]$$

recalling that $\mu_0 = 1$ and $\mu_1 = 0$. Use this relation [1] to establish the values given in the text.

100. THE POINT BINOMIAL APPLIED TO FREQUENCY DISTRIBUTIONS

It should be emphasized that the terms of (1) represent probabilities and that their sum is unity. By Section 95 (p. 374), if the terms of (1) are multiplied by some suitable number, the several terms will then represent frequencies. Thus, if 10 coins are thrown 1,024 times, the terms of the expansion

$$1{,}024(\tfrac{1}{2} + \tfrac{1}{2})^{10} = 1 + 10 + 45 + \cdots + 252 + \cdots + 10 + 1$$

represent the expected number of times that we should obtain 0, 1, 2, ..., 5, ..., 9, 10 heads, that is,

$$\textit{expected frequency of} \quad X = (1024)_{10}C_X(\tfrac{1}{2})^{10-X}(\tfrac{1}{2})^X$$

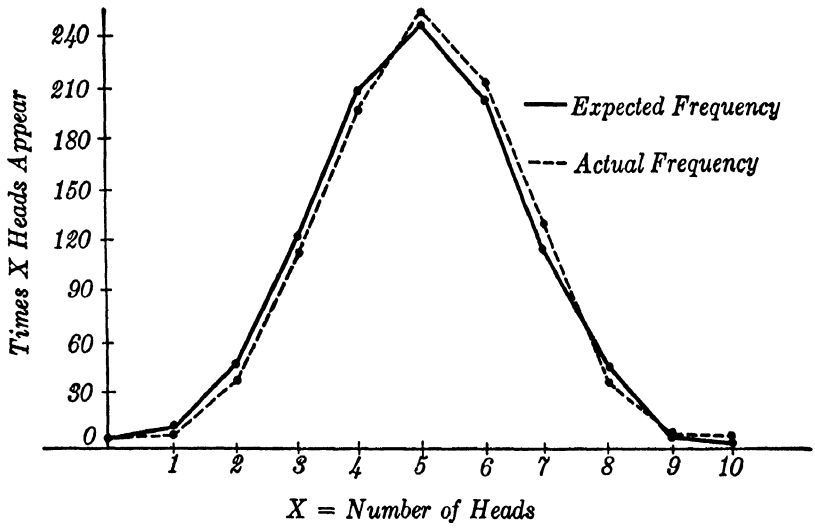
An experiment in which 10 coins were thrown 1,024 times was performed, and the actual results together with the theoretical or expected results are shown in Table 89.

[1] See article by A. T. Craig, *Bulletin of the American Mathematical Society*, Vol. 40, p. 262.

TABLE 89. ACTUAL AND EXPECTED RESULTS IN TOSSING 10 COINS 1,024 TIMES

Number Heads Up X	*Actual Frequency* $f(X)$	*Expected Frequency* $f'(X)$
0	2	1
1	10	10
2	40	45
3	116	120
4	205	210
5	257	252
6	216	210
7	126	120
8	42	45
9	8	10
10	2	1
Total	1,024	1,024

FIGURE 54



When the data of Table 89 are plotted as in the Figure 54, and the frequency polygons are drawn, the differences between the observed and the expected frequencies are seen to be slight. These differences may be the result of many causes, such as the lack of homogeneity of the coins, the faulty methods of tossing them, and what are usually known as variations due to chance.

The statistical constants for the observed and the theoretical distributions of Table 89 are given in Table 90. The constants for the theoretical distribution were computed by (3) and Exercise 4 on page 390, whereas those for the distribution of observed values were computed by the methods of Section 44 (p. 164).

TABLE 90

	Distribution of Theoretical Values	*Distribution of Observed Values*
M	5.0000	5.0283
σ	1.581	1.567
α_3	0.0000	− 0.0499
α_4	2.8000	2.9246

In a similar manner any distribution of observed values can be more or less approximately reproduced by multiplying the terms of the expansion of $(q + p)^n$ by the total frequency N. If the distribution is nearly symmetrical, we take $p = q = \frac{1}{2}$ and n such a number that the $(n + 1)$ terms of the expansion when multiplied by N will give $(n + 1)$ theoretical frequencies.

Thus, let us consider the following distribution of the heights of 750 college men. Since distributions of heights of men are known to be closely symmetrical, we choose $p = q = \frac{1}{2}$. Also, since there are 14 classes of heights ranging from 61 inches to 74 inches inclusive, we choose $n = 13$. Hence the terms of the expansion $750(\frac{1}{2} + \frac{1}{2})^{13}$ give 14 theoretical frequencies. The following table exhibits the frequency distributions of theoretical and observed values. The theoretical frequency, for a given X, is $750_{13}C_X(\frac{1}{2})^{13-X}(\frac{1}{2})^X$.

Exercise. Compute values of M, σ, α_3, and α_4 for the two distributions of Table 91 and thus make a comparison of their moments.

TABLE 91. OBSERVED AND BINOMIAL FREQUENCIES OF THE HEIGHTS OF 750 COLLEGE MEN

Height	X	*Observed* $f(X)$	*Binomial* $f(X)$
61	0	2	0
62	1	4	1
63	2	10	7
64	3	32	26
65	4	63	66*
66	5	103	118
67	6	146	157
68	7	143	157
69	8	111	118
70	9	75	66*
71	10	35	26
72	11	12	7
73	12	3	1
74	13	1	0
	Total	750	750

* This value was 65.5.

Comparing the observed with the theoretical frequency it is of course noted that, for a given value of X, the observed frequency differs from the theoretical frequency. Even the most scrupulous among us are not surprised at these differences. However, the student may properly inquire as to just how large such differences may be. This is one of the fundamental questions to which we shall give attention in Chapter 13 when we consider the problem of sampling.

For a given N and n, the theoretical distribution $N(q + p)^n$ obviously depends upon the value of p or q. The value of p may be determined *a priori* as in dice-throwing or coin-tossing experiments, or it may be determined empirically from experiment or observation as in the probabilities of life and death. When p is determined empirically, it is influenced by sampling errors. Other samples of the same size chosen from the same universe will not yield the same values of p, and consequently the goodness of the theoretical distribution $N(q + p)^n$ for graduation purposes will depend upon the accuracy of p.

The binomial distribution $(q + p)^n$ was the first theoretical distribution to be established. It was first discussed in *Ars Conjectandi* (published posthumously in 1713) by James Bernoulli and thus any discrete distribution with frequencies proportional to the terms of the expansion is frequently called a Bernoulli Distribution. In fact, what we have called the *Repeated Trials Theorem* is frequently called *The Bernoulli Theorem.*

EXERCISES

1. Table A below gives the I.Q.'s of 905 school children. Table B gives the weights of 1000 school children. Graduate Table A by the expansion $905(\frac{1}{2} + \frac{1}{2})^8$ and Table B by $1000(\frac{1}{2} + \frac{1}{2})^9$.

A

TABLE

X	$f(x)$
60.5	3
70.5	21
80.5	78
90.5	182
100.5	305
110.5	209
120.5	81
130.5	21
140.5	5
Total	905

$M = 100.95$
$\sigma = 13.0$

B

TABLE

X	$f(x)$
29.5	1
33.5	14
37.5	56
41.5	172
45.5	245
49.5	263
53.5	156
57.5	67
61.5	23
65.5	3
Total	1000

$M = 47.71$ pounds
$\sigma = 5.88$ pounds

101. THE NORMAL CURVE: INTRODUCTORY REMARKS

In preceding chapters we have described frequency distributions by three methods: the graphical method, the method of moments, and the point binomial. The graphical method is a mere pictorial representation of the tabulated data and is inadequate statistically because it is only a picture. The method of moments is a refined method which is adequate for many purposes, especially for purposes of comparison, when M, σ, α_3, and α_4 are computed. The binomial

distribution is still a step forward. It gives us an *equation* for writing down the theoretical frequency for a given integral value of X, and the estimated sum of such frequencies between certain specified limits. Thus, theoretically at least, the point binomial provides all the advantages that accrue from an equation.

Practically, the point binomial is unsatisfactory for two important reasons. First, it is a discontinuous function, being strictly defined only for integral values of X. Second, when n is large, its use in answering many questions entails so much labor as to render it unfit for practical usage. We seek, therefore, a continuous function having approximately the same ordinates as the binomial series and which is so well tabulated that important questions in probability can be answered by its use without the tedium of undue labor. The simplest continuous function that meets our needs is the *normal* or Gaussian function, whose general equation is:

$$y = Ce^{-h^2x^2}$$

Here e is the base of the natural or Napierian system of logarithms whose value is 2.71828.... The constant C determines the maximum height of the curve and the constant h its spread.

As was stated in Section 89, the normal or Gaussian curve was first established by De Moivre. A proof was also given by Laplace at a later date and hence the curve is sometimes called the Laplacean curve. Gauss approved the law, used it, and gave an original proof of it. Thus, the normal law began its early life with a rare hereditary background. No wonder the lesser lights of the first half of the nineteenth century claimed for it a value that was undeserved, considered it to be "the ideal curve," and demanded an explanation if a distribution did not obey it.

The writers in the latter half of the nineteenth century seem to have been more careful that their enthusiasms did not outrun the facts, for as data from many fields accumulated it became general knowledge that the normal curve is but one of a number of types of curves which are used to describe frequency distributions. So we must not assume that a non-normal distribution is "abnormal" in the usual sense of the word.

The normal curve, however, is by far the most important type; further, its importance seems to have increased within recent years,

and the history of the theory of statistics may date from its discovery by De Moivre in 1733. There are good reasons why this is so.

First, it is a continuous function.

Second, the normal curve lends itself well to mathematical treatment. That is, it possesses properties that are mathematically elegant, comparatively simple to derive, and expressible in simple forms.

Third, a large number of distributions, mound-shaped in appearance, are approximately of the normal form and may be subjected to normal curve analysis as a first approximation.

Fourth, many sampling distributions, such as distributions of means, distributions of standard deviations, and others are of the normal form exactly or to a satisfactory degree of approximation. Thus, the formulas for determining the reliability of a statistical function "lean heavily upon this law."

Fifth, of two well-known systems of generalized frequency curves, one of them, that developed by Gram, Thiele, Charlier (known as the Scandinavian school), is based upon the normal curve as a generating function.

A development of the theory of generalized frequency functions, though an important and attractive study, is so severe in the mathematical background required to comprehend it that its inclusion in our elementary study would seem inappropriate. However, a derivation of the normal curve and a study of its properties are so essential to the study of elementary statistical analysis that their inclusion in our text seems mandatory.

102. DERIVATION OF THE EQUATION TO THE NORMAL CURVE

Figure 51 (p. 383) shows the frequency polygon for the point binomial $(\frac{1}{2} + \frac{1}{2})^{10}$. The eleven points are symmetrically distributed about the vertical line through $X = M = np = 5$. In like manner if $(\frac{1}{2} + \frac{1}{2})^n$ be plotted for any n, the points will be symmetrically distributed about the vertical line through $X = M = np = n/2$ since $p = q$ and $\alpha_3 = 0$. Now if n be allowed to increase indefinitely the polygon of $(n + 1)$ vertices and $(n + 2)$ sides will approach a smooth curve,[1] the normal curve, symmetrical to the vertical line through

[1] As n increases, it becomes necessary to reduce the X-scale to keep the diagram within reasonable dimensions. We are interested in confining the range to an interval of three or four standard deviations from the mean. Consequently, we assume that n increases and (Δx) decreases in such a way that $n(\Delta x)^2$ always equals a constant $2\sigma^2$.

$X = M$. In other words, *the normal curve is the limit of the point binomial* $(\frac{1}{2} + \frac{1}{2})^n$ *as* n *becomes infinite.*

The proof of the statement above is facilitated by assuming that n is even and by employing the

Lemma. If the several terms of the expansion $(\frac{1}{2} + \frac{1}{2})^{2n}$ be plotted as ordinates at intervals of ΔX along the X-axis, ${}_{2n}C_0/2^{2n}$ being taken at the origin, so that the abscissas of ${}_{2n}C_1/2^{2n}$, ${}_{2n}C_2/2^{2n}$, . . . , ${}_{2n}C_n/2^{2n}$, . . . , ${}_{2n}C_{2n}/2^{2n}$ are ΔX, $2\Delta X$, . . . , $n\Delta X$, . . . , $2n\Delta X$, then:

$$M = n\Delta X \quad \text{and} \quad \sigma^2 = \frac{n}{2}\overline{\Delta X}^2$$

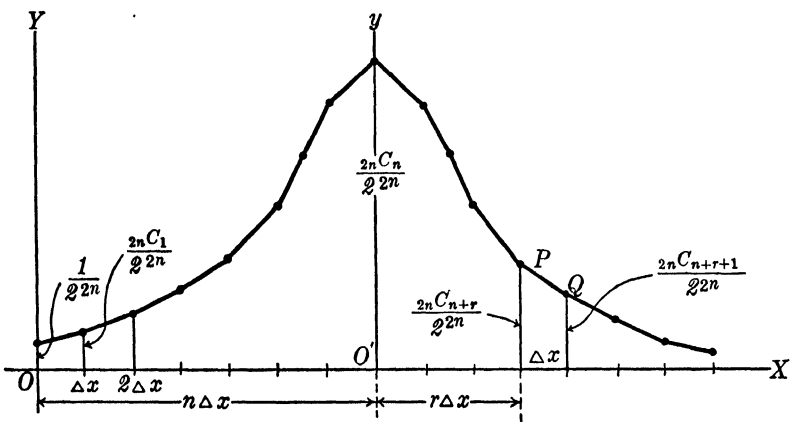
The proof of this lemma is identical in method to that used in Section 99B (p. 387), hence its derivation will be left as an exercise for the student.

Let us consider then the expansion:

$$\left(\frac{1}{2}+\frac{1}{2}\right)^{2n} = \frac{1}{2^{2n}}[1 + {}_{2n}C_1 + {}_{2n}C_2 + \cdots + {}_{2n}C_n + \cdots + {}_{2n}C_{n+r} + \cdots + 1]$$

Let us plot the terms of this expansion as ordinates at equal intervals ΔX along the X-axis beginning with *the first term at the origin.* The maximum term is evidently $\frac{{}_{2n}C_n}{2^{2n}}$ which we erect at the

FIGURE 55



mean, O'. We plot the other terms with respect to this new origin. Evidently $\Delta x = \Delta X$. Let $P(x, y)$ and $Q(x + \Delta x, y + \Delta y)$ be the successive vertices of the polygon which are determined by the rth and the $(r + 1)$th terms from the middle term of the above expansion. Then the ordinates of the points are:

$$y = \frac{{}_{2n}C_{n+r}}{2^{2n}} \quad \text{and} \quad y + \Delta y = \frac{{}_{2n}C_{n+r+1}}{2^{2n}}$$

Since

$${}_{2n}C_{n+r+1} = {}_{2n}C_{n+r}\left(\frac{n - r}{n + r + 1}\right) \quad \text{(see Exercise 13 on p. 374)}$$

we have:

$$\frac{\Delta y}{\Delta x} = \frac{y}{\Delta x}\left(\frac{-2r - 1}{n + r + 1}\right)$$

The abscissa of P is $x = r\Delta x$; hence:

$$r = \frac{x}{\Delta x}$$

Consequently:

$$\frac{\Delta y}{\Delta x} = -y\left(\frac{2x + \Delta x}{n\overline{\Delta x}^2 + x\Delta x + \overline{\Delta x}^2}\right)$$

From the lemma above we have:

$$n\overline{\Delta x}^2 = 2\sigma^2, \text{ a constant}$$

Therefore:

$$\frac{\Delta y}{\Delta x} = -y\left(\frac{2x + \Delta x}{2\sigma^2 + x\Delta x + \overline{\Delta x}^2}\right)$$

Now let n become infinite and Δx approach zero.[1] We then have

$$\frac{dy}{dx} = -\frac{xy}{\sigma^2}$$

which, upon integration, reduces to:

$$y = Ce^{-\frac{x^2}{2\sigma^2}} = Ce^{-h^2x^2}$$

where $h\left(= \frac{1}{\sigma\sqrt{2}}\right)$ is called the *index of precision.*

[1] See footnote page 397.

In order to make this curve statistically useful, we shall assume that the area under the curve is equal to the area of the histogram, Nw, where w is the class width and N is the total frequency. That is, we assume

$$\int_{-\infty}^{\infty} y dx = Nw$$

from which it follows, using the well-known relation: [Ex. 4, p. 404]

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

$$C = \frac{Nw}{\sigma\sqrt{2\pi}}$$

Substituting this value, we have the equation to the *normal frequency curve:*

$$y = \frac{Nw}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{4}$$

It must be emphasized that in equation (4) x is the deviation of the frequency y or $f(x)$ from the *mean.* By replacing x by its equal $X - M$ we may express the equation in the form:

$$Y = \frac{Nw}{\sigma\sqrt{2\pi}} e^{-\frac{(X-M)^2}{2\sigma^2}} \tag{5}$$

If in (4) we make the area under the curve equal to unity, the equation reduces to the *normal probability curve:*

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{6}$$

which gives the probability of any deviation x.

It is customary, due to the simplicity of application, to express the deviations in standard units, that is, to make σ the unit for measuring deviations. If in (4) and (5) we place

$$t = \frac{x}{\sigma} = \frac{X - M}{\sigma}$$

we obtain:

$$y = \frac{Nw}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{7}$$

Finally we write:

$$y = \frac{Nw}{\sigma_x}\phi(t) \tag{8}$$

where

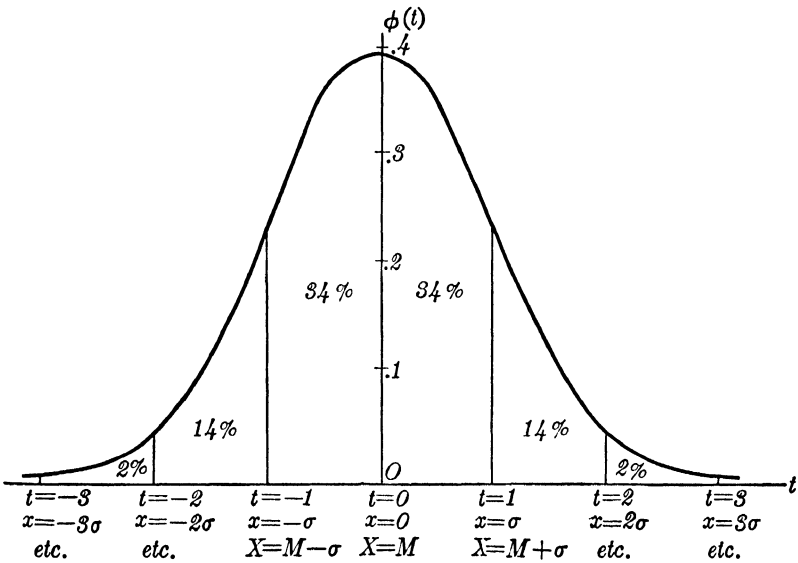
$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{9}$$

(Read: phi of tee.)

103. SOME PROPERTIES OF $\phi(t)$[1]

Values of $\phi(t)$ and of the areas bounded by $\phi(t)$, the t-axis, and certain ordinates are tabulated in Appendix B. The graph of $\phi(t)$ is shown in the accompanying figure which is drawn from the values in Table 92.

FIGURE 56



Since $-t$ yields the same value to $\phi(t)$ as $+t$, that is, since $\phi(-t) = \phi(t)$, the curve is symmetrical with respect to the vertical

[1] Several of these properties require the calculus for *proofs*.

line through $t = 0$. It is therefore not necessary to tabulate negative values of t. Since the total area under $\phi(t)$ is 1.0000, the area on either side of the vertical line of symmetry is 0.5000. Therefore the median coincides with the mean. The largest value of $\phi(t)$ is that for which $t = 0$, therefore the mode coincides with the mean. There is no finite value of t for which $\phi(t) = 0$, but $\phi(t)$ is relatively small for values of t outside of $t = \pm 3$. It is because of the last-mentioned fact that the normal curve can be used to represent finite distributions. As a matter of fact the combined area of the two tails beyond $t = -3$ and $t = +3$ is only 0.0026, and the combined area of the two tails beyond $t = -4$ and $t = +4$ is 0.000,064. The curve crosses its tangent at $t = \pm 1$, $\phi(t) = .2420$. These are called *inflection points.*

TABLE 92

t	$\phi(t)$
0	.3989
0.5	.3521
1.0	.2420
1.5	.1295
2.0	.0540
2.5	.0175
3.0	.0044

The areas of certain portions of $\phi(t)$ are so important in statistical analysis that we must not fail to emphasize them. We shall use the symbol $A_{\phi}\rceil_{t=a}^{t=b}$ or, more briefly, $A_{\phi}\rceil_{a}^{b}$ to mean "the area under $\phi(t)$ from $t = a$ to $t = b$." Thus, we have from the table $A_{\phi}\rceil_{0}^{1} = .3413$, $A_{\phi}\rceil_{0}^{2} = .4773$, $A_{\phi}\rceil_{0}^{3} = .4987$. By the simple addition and subtraction of areas we also have

$$A_{\phi}\rceil_{1}^{2} = .1360, \quad A_{\phi}\rceil_{2}^{3} = .0214, \quad A_{\phi}\rceil_{-1}^{0} = .3413, \quad A_{\phi}\rceil_{-1}^{2} = .8186.$$

The statement $A_{\phi}\rceil_{0}^{1} = .3413$ means that between the ordinates erected at $t = 0$ and $t = 1$ is included 34.13 per cent of the total area under the curve. More broadly interpreted, it means that for a *normal frequency distribution* about one-third of the total frequency is found between the mean and $x = \sigma$ (see p. 135). In the language of probability, the statement means that the chance is approximately 1/3 that a measure selected at random from a given distribution of variates normally distributed will fall within the interval between $t = 0$ and $t = 1$, or between $x = 0$ and $x = \sigma$ or between $X = M$ and $X = M + \sigma$.

It will be left as exercises for the student to interpret the other areas illustrated above.

The value of t that satisfies one of the equations

$$A_\phi\Big]_0^t = .2500 \qquad A_\phi\Big]_{-t}^{+t} = .5000 \tag{10}$$

defines one of the most important concepts found in statistics. The value of t defined by either of the given equations (10) is called the *probable error*, E, of a single observation. The probable error, E, is that distance which, when laid off on either side of the mean of a normal curve, defines an interval such that, if ordinates are erected at its end points, the area included by the ordinates, the curve, and the base line is one-half the total area under the curve. Stated somewhat differently, the probable error of a distribution of variates normally distributed may be defined as that deviation on either side of the mean within which exactly half the variates lie. Since half the total frequency lies within the interval $M - E$ to $M + E$, if any one variate be selected at random from the N given variates there is an even chance that the selected variate falls within the given interval $M - E$ to $M + E$ or without it.

For an approximate solution of equation (10) let us interpolate between $t = .67$ and $t = .68$. The solution is:

$$.01\left[\begin{array}{l} z\left[\begin{array}{l} A_\phi\Big]_0^{.67} = .2486 \\ A_\phi\Big]_0^{t} = .2500 \end{array}\right. \leftarrow .0014 \\ \qquad A_\phi\Big]_0^{.68} = .2518 \end{array}\right] \leftarrow .0032$$

$$\frac{z}{.01} = \frac{.0014}{.0032}$$

$$z = .0044$$

and $t = .67 + z = .6744$. More extended tables lead to the more accurate value

$$t = \frac{x}{\sigma} = .6745 \text{ (approximately)}$$

and therefore

$$x = .6745\sigma$$

that is:

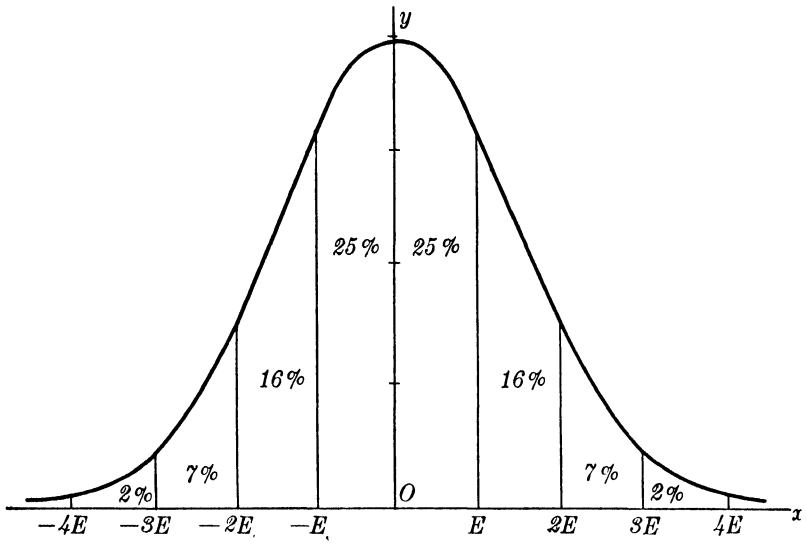
$$E_X = .6745\sigma_X \tag{11}$$

If a distribution is not normal, its probable error is estimated by equation (11).

Figure 57 will assist in clarifying the concept of probable error.

The values of the moments of the normal curve are given in Exercise 8, page 405.

FIGURE 57



EXERCISES

1. Find the portions of the area under $\phi(t)$ indicated, and draw a figure in each case.

a. $A_\phi\Big]_{-\infty}^{2}$ d. $A_\phi\Big]_{2}^{\infty}$ g. $A_\phi\Big]_{-\infty}^{-2}$

b. $A_\phi\Big]_{-2.4}^{2.4}$ e. $A_\phi\Big]_{-2.4}^{\infty}$ h. $A_\phi\Big]_{-2}^{8}$

c. $A_\phi\Big]_{-2}^{2.389}$ f. $A_\phi\Big]_{2.389}^{\infty}$ i. $A_\phi\Big]_{-3.468}^{-2.746}$

2. Find t in the following equations:

a. $A_\phi\Big]_{0}^{t} = .4838$ c. $A_\phi\Big]_{0}^{t} = .4510$

b. $A_\phi\Big]_{-t}^{t} = .4844$ d. $A_\phi\Big]_{-t}^{t} = .4878$

3. Verify the percentages of Figure 57, in which E is taken as the x-unit.

The following exercises are for students of calculus.

4. Prove: $$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$$

Hint: Let $$I = \int_{-\infty}^{\infty} e^{-x^2}dx = \int_{-\infty}^{\infty} e^{-y^2}dy$$

or $$I^2 = \int_{-\infty}^{\infty} e^{-x^2}dx \int_{-\infty}^{\infty} e^{-y^2}dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2-y^2}dy\,dx$$

which is the volume under the surface $z = e^{-x^2-y^2}$. Change to polar coordinates. Then $I^2 = 4\int_0^{\frac{\pi}{2}}\int_0^{\infty} e^{-r^2} r\,dr\,d\theta = \pi$.

5. Show that y in (4) has a maximum at $x = 0$.

6. Show that y in (4) has inflection points at $x = \pm\sigma$.

7. Consider equation (4). Show that the mean deviation about the mean $= \frac{1}{Nw}\int_{-\infty}^{\infty} |x| y\,dx = \frac{2}{Nw}\int_0^{\infty} xy\,dx = \sqrt{\frac{2}{\pi}}\sigma = 0.79788\cdots\sigma$.

8. Evaluate the moments of the normal curve (4), where $\mu_i = \frac{1}{Nw}\int_{-\infty}^{\infty} x^i y\,dx$. That is, show that

$$\mu_0 = 1, \quad \mu_1 = 0, \quad \mu_2 = \sigma^2, \quad \mu_3 = 0, \quad \mu_4 = 3\mu_2^2 = 3\sigma^4$$
$$\alpha_0 = 1, \quad \alpha_1 = 0, \quad \alpha_2 = 1, \quad \alpha_3 = 0, \quad \alpha_4 = 3$$
$$\alpha_{2n} = 1\cdot 3\cdot 5\cdot\cdots\cdot(2n-1) = \frac{(2n)!}{2^n(n!)}$$
$$\alpha_{2n+1} = 0$$

9. Show that for the normal curve

Mean Deviation about M = 1.183 Probable Error
Probable Error = 0.8454 Mean Deviation

104. ILLUSTRATIVE EXAMPLES

Example 1. Given a normal distribution with $N = 1{,}000$, $w = 2$, $M = 16$, and $\sigma = 4$: a. How many variates fall between $X = 12$ and $X = 20$? b. How many lie above $X = 26$? c. How many lie below $X = 10$?

FIGURE 58

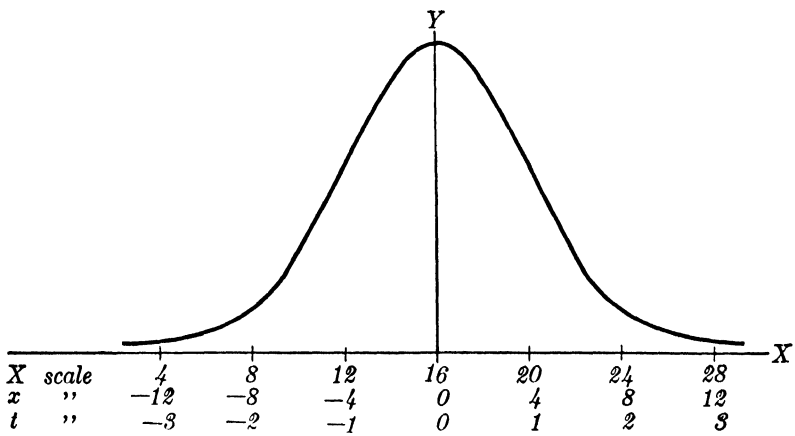


Figure 58 shows a normal curve with $M = 16$, $\sigma = 4$, area $= (1000)2$. Since our tables are expressed for values of t, we must transform our data into t units. We have shown three scales on the base line. If $X = 12$, $x = X - M = 12 - 16 = -4$ and $t = x/\sigma = -4/4 = -1$. Similarly, if $X = 20$, $t = 1$; if $X = 26$, $t = 2.5$, and if $X = 10$, $t = -1.5$.

a. Now $$A_\phi\Big]_{-1}^{1} = .6826$$

This means that 68.26 per cent of the total area under the curve lies between $t = -1$ and $t = 1$, or between $X = 12$ and $X = 20$. By means of the calculus it can be shown that the area under Y from X_1 to X_2 or under y from x_1 to x_2 is $Nw \times$ the area under $\phi(t)$ between t_1 and t_2, that is: $A_Y\Big]_{X_1}^{X_2} = A_y\Big]_{x_1}^{x_2} = Nw \cdot A_\phi\Big]_{t_1}^{t_2}$. See equations (4), (5), and (8). Therefore:

$$A_Y\Big]_{X=12}^{X=20} = .6826(1000)2 = (682.6)2$$

Since

2,000 units of area represent 1,000 variates,
(682.6)2 units of area represent 682.6 variates.

That is, 682.6 variates fall between $X = 12$ and $X = 20$.

In short, since $A_\phi\Big]_{-1}^{1} = .6826$
we may say that 68.26 per cent of N or

$$.6826(1{,}000) = 682.6$$

variates fall between $X = 12$ and $X = 20$.

b. Similarly, since $A_\phi\Big]_{2.5}^{\infty} = .5000 - .4938 = .0062$,

$$.0062(1{,}000) = 6.2$$

variates are beyond $X = 26$.

c. Since $A_\phi\Big]_{-\infty}^{-1.5} = .5000 - .4332 = .0668$,

$$.0668(1{,}000) = 66.8$$

variates are below $X = 10$.

Example 2. For the distribution described in Example 1, compute Y when $X = 4, 8, 12, 16, 20, 24, 28$.

Using (5), the equation of the curve is:

$$Y = \frac{(1000)2}{4\sqrt{2\pi}}\, e^{-\frac{(X-16)^2}{32}}$$

Let:

$$t = \frac{x}{\sigma} = \frac{X - M}{\sigma} = \frac{X - 16}{4}$$

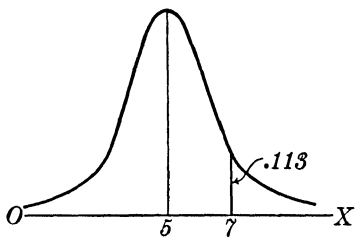
Then:

$$Y = \frac{(1000)2}{4} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} = 500\phi(t)$$

Recalling that $\phi(-t) = \phi(t)$, we have the following table of values.

X	t	$\phi(t)$	Y
4	− 3	.0044	2.2
8	− 2	.0540	27.0
12	− 1	.2420	121.0
16	0	.3989	199.4
20	1	.2420	121.0
24	2	.0540	27.0
28	3	.0044	2.2

Example 3. If 10 coins are thrown, use the normal probability function to find the approximate probability of obtaining exactly 7 heads.



The various probabilities are given by the terms of $(\frac{1}{2} + \frac{1}{2})^{10}$.

The **exact** probability of obtaining 7 heads is given by:

$$P_7 = {}_{10}C_7(\tfrac{1}{2})^3(\tfrac{1}{2})^7 = .117$$

We may apply the normal curve to obtain an **approximate** value of P_7. We have:

$$M = np = 10(\tfrac{1}{2}) = 5$$

$$\sigma = \sqrt{npq} = \sqrt{10(\tfrac{1}{2})(\tfrac{1}{2})} = 1.581$$

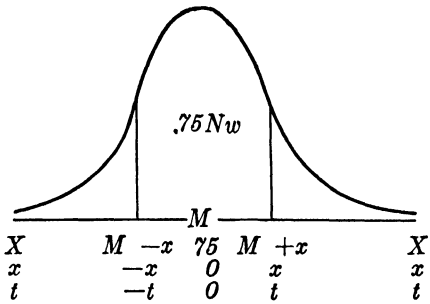
$y = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} = \dfrac{1}{\sigma}\phi(t)$ gives the probability of any deviation x.

We seek y for $X = 7$. But if $X = 7$, $x = X - M = 7 - 5 = 2$ and $t = \dfrac{x}{\sigma} = \dfrac{2}{1.581} = 1.265$. Since $\phi(1.265) = .1792$, we have therefore

$$y = \frac{1}{\sigma}\phi(1.265) = \frac{.1792}{1.581} = .113$$

The slight discrepancy in the two results is an evidence that the point of the given binomial is near the normal curve.

Example 4. Given a normal distribution with $M = 75$ and $\sigma = 8$, what limits will include the middle 75 per cent of the total frequency?



We must solve the equation:

$$A_v\Big]_{-x}^{x} = .75Nw$$

or the equation

$$A_\phi\Big]_{-t}^{t} = .75$$

Since

$$A_\phi\Big]_{-t}^{t} = 2[A_\phi]_0^t = .75,$$

we have:

$$A_\phi\Big]_0^t = .375$$

From the tables

$$t = \frac{x}{\sigma} = 1.15$$

and therefore:

$$x = 1.15\sigma = (1.15)8 = 9.20$$

Hence the limits are $M \pm x = 75 \pm 9.20 = 65.80$ and 84.20.

In approximating a *sum* of the successive terms of the point binomial by the normal curve, we must find the area under the appropriate part of the curve. The sum of the successive terms of the binomial equals the sum of the areas of the corresponding rectangles of the histogram. We must then replace the rectangles of the histogram by corresponding areas of the curve and *this requires that we use whole rectangles*, not half rectangles at the ends.

It is evident that the normal curve will give a close approximation to the sum of the terms of a binomial only when p and q are nearly equal, and n is fairly large. Certainly if there is considerable skewness, the approximation by the normal curve may not be satisfactory, especially near the ends of the distribution. We cannot make definite statements as to when the normal curve may be used as an approximation to the binomial. Whether the approximation is satisfactory or not depends upon the accuracy of the results desired and how the approximation is to be used.

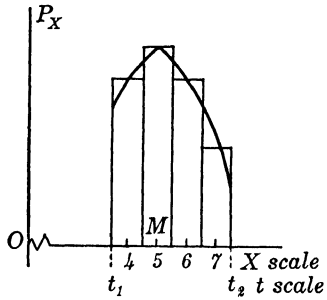
Exercise 5. If 10 coins are tossed, what is the probability of getting 4, 5, 6 or 7 heads? (a) Use the theorem of repeated trials for an accurate

result correct to two decimals, and (b) use the normal curve to find an approximate result.

Solution to (a). By the theorem of repeated trials the required probability is the sum $\sum_{4}^{7} {}_{10}C_x(\frac{1}{2})^x(\frac{1}{2})^{10-x}$. This sum is

$$P = \frac{210 + 252 + 210 + 120}{1024} = \frac{792}{1024} = .77$$

Solution to (b).



$$n = 10$$
$$p = q = \tfrac{1}{2}$$
$$M = np = 5$$
$$\sigma = \sqrt{npq} = 1.58$$
$$x_1 = 3.5 - 5 = -1.5$$
$$x_2 = 7.5 - 5 = 2.5$$
$$t_1 = \frac{x_1}{\sigma} = \frac{-1.5}{1.58} = -.95$$
$$t_2 = \frac{x_2}{\sigma} = \frac{2.5}{1.58} = 1.58$$

Approximate $P = A_\phi\Big]_{t_1}^{t_2} = .3289 + .4430 = .7719.$

105. ON THE SIGNIFICANCE OF RESULTS

It has been observed that for a normal or a moderately skewed, mound-shaped distribution the total range seldom exceeds six times the standard deviation. If, then, a distribution is approximately normal, it is not expected that a measure chosen at random will show a variation of more than three times the standard deviation, on either side, from the mean. A divergence of more than $\pm 3\sigma$ (about $\pm 4.5E$) may be called *significant;* that is, other forces than mere chance have most probably operated to bring about *abnormal* results. Thus if 400 coins are tossed (or if one coin is tossed 400 times) what is the allowable variation in the number of heads? We have:

$$M = np = 400(\tfrac{1}{2}) = 200 = \text{the } \textit{expected} \text{ number of heads}$$

and

$$\sigma = \sqrt{npq} = \sqrt{400(\tfrac{1}{2})(\tfrac{1}{2})} = 10$$
$$3\sigma = 30$$

It is very improbable then that less than 170 ($= 200 - 30$) and more than 230 ($= 200 + 30$) heads will appear. In fact we can meas-

ure the probability in question. Since $A_\phi]_{-3}^{3} = .9974$, if 400 coins are tossed, the probability of obtaining between 170 and 230 heads is 9,974/10,000. That is, the probability of obtaining more than 230 and less than 170 heads is 26/10,000. In other words, the odds in favor of obtaining between 200 ± 30 heads are 9,974 to 26 or 383.6 to 1.

In general, we may state that the probability of a measure's lying within the range $M \pm 3\sigma$ or $M \pm 4.5E$ is 9,974/10,000 and that the odds favoring a measure's lying within this range are nearly 385 to 1.

Another type of language has become fashionable when speaking of certain t or x values in connection with the normal curve. It is seen from our tables that $A_\phi]_{-1.96}^{1.96} = .95$ and thus 5 per cent of the area lies outside the limits $t = \pm 1.96$ or $x = \pm 1.96\sigma$. Consequently, there is 1 chance in 20 that x may lie outside $\pm 1.96\sigma$. This value 1.96σ is called the *5 per cent level of significance.* Similarly, $A_\phi]_{-2.576}^{2.576} = .99$ and thus 1 per cent of the area lies outside the limits $t = \pm 2.576$ or $x = \pm 2.576\sigma$. Consequently, there is 1 chance in 100 that x may lie outside $\pm 2.576\sigma$. This value 2.576σ is called the *1 per cent level of significance.* These values may be called *confidence limits*, the probability giving a measure of confidence that an item falls within the stated limits.

The question, "At what probability level does a deviation become significant?" is one that cannot be answered with scrupulous exactness. Statisticians differ in their credulity. Any level that is set is arbitrary. Conceivably, a deviation x may be any amount. However, the occurrence of the deviation may be so unlikely that it can hardly be looked upon as due to chance. Some authorities state that if x is outside the 5 per cent level it is *significant;* if it is outside the 1 per cent level, it is *highly significant.* A safe procedure for the student is that he be prepared to state in terms of probability, or as a percentage, *the level of significance* for any deviation.

Questions. What are the values of t and x for the 10 per cent level of significance?

What are the values of t and x for the 25 per cent level of significance?

What is the per cent level of significance of a deviation $t = \pm 3$ or $x = \pm 3\sigma$?

EXERCISES

1. In a coin-tossing experiment in which a coin was tossed 400 times, 250 heads appeared. Do you believe that the experiment was honestly performed?

2. Suppose that the mortality statistics for a large group of cities show the average death rate from tuberculosis to be 196.5 per 100,000 population, and $\sigma = 14$. A particular city showed a death rate from tuberculosis of 110.3 per 100,000. Is this surprising? Another city (a haven for tuberculosis patients) showed a death rate of 245 per 100,000 for the same disease. Is this surprising from the point of view of mere chance?

3. A coin was tossed 100 times. Find, using the normal curve, the probability of obtaining exactly 60 heads.

4. In a college the 12 grades A+, A, A−; B+, B, B−; C+, C, C−; D, E, and F are given. On the assumption that ability in mathematics is normally distributed, how many in a group of 1,000 grades should receive each grade mentioned? Assume that the total range is $M \pm 3.6\sigma$.

5. (*Thurstone*) Construct three frequency curves on the same sheet according to the following specifications. Indicate an ordinate at the midpoint of each class interval.

Curve	σ	M	N	w
A	15	50	400	10
B	15	50	800	10
C	15	50	1,200	10

6. Construct three frequency curves on the same sheet according to the following specifications. Compute ordinates for each half-sigma.

Curve	σ	M	N	w
A	5	50	1,000	10
B	10	50	1,000	10
C	15	50	1,000	10

7. Draw a normal curve $\phi(t)$ and divide the base line into five parts such that when ordinates are erected at the points of division the five areas will be equal.

8. A normal distribution has the following constants: $N = 1{,}000$; $w = 5$; $M = 73.64$; $\sigma = 8.3$. How many variates are between $X = 61$ and $X = 94$?

9. Determine whether it is expected that one will obtain:

a. 2,048 heads in 4,040 throws of a coin.

b. 3,300 heads in 6,400 throws of a coin.

c. 38,024 appearances of a four, a five, or a six in 78,000 throws of a single die.

10. Compute the ordinates for the point binomial $(\frac{1}{2} + \frac{1}{2})^{16}$ and compare them with the ordinates of a superimposed normal curve.

11. If a baseball player has a batting average of 0.300, what is the probability that he will hit safely at least 25 times out of 100 times at bat? Estimate by the normal curve. Note that α_3 is small.

12. If 16 coins are tossed, what is the probability of getting 5, 6, 7, 8, 9, 10, 11, or 12 heads? (a) Use the theorem of repeated trials for a result correct to two decimals, and (b) the normal curve for an approximate result to two decimals.

13. The probability of a man of age 56 dying within a year is 0.02. If an insurance company has 10,000 policies in force on men of this age, find the probability of the company's having to pay less than 180 death claims; more than 220 death claims. Estimate by the normal curve. Note that α_3 is small.

14. A large number of students were measured as to height and for them we found $M = 67.5$ inches. We found that 40 per cent of the students were between 66.2 inches and 68.8 inches in height. What is the standard deviation of the heights?

15. In the United States in 1930, 12 per cent of the marriageable men were widowers. Assume this situation normal. A city has 6,000 men who are marriageable (single men 15 years old and over). (a) How many would you expect to be widowers? Note that α_3 is small. (b) Estimate the probability that there will be as few as 600 widowers. (c) As many as 750 widowers.

16. The experience of a manufacturing concern has been that in the past they have had to discard 5 per cent of the units inspected as defective. A sample of 1,000 units is up for inspection. (a) How many defective units would you expect? (b) What are the values at the 5 per cent level of significance?

17. In 1930, about 9 per cent of the people of the United States were "20 and under 25" years of age. In a typical city of the United States of population 10,000, how many would you expect to find between 20 and 25 years of age? Adopting $\pm 3\sigma$ as the limits of reasonable chance occurrence, would you be surprised to find as few as 800? As many as 1000?

18. (*Waugh*) In an epidemic of infantile paralysis which took place in the eastern part of the United States in the fall of 1931, we have records on 927 children who contracted the disease. Of these, 408 received no serum and 104 of the 408 became paralyzed, while the other 304 recovered without paralysis. If the serum had no effect, how many cases would you have expected among the 519 who were given serum? (Assume 3σ marks the limit of reasonable chance occurrence.) Actually 166 of the children receiving serum were paralyzed. What do you conclude as to the efficacy of the serum? What other factors might influence the result besides the effect of the serum?

19. A group of 1,000 students took an objective and standardized test. The distribution was closely normal with $M = 60$ and $\sigma = 10$. What are the values of Q_1, Q_3, Q, $M.D.$, α_3, α_4, and the 87th percentile?

20. It has been established that of children under one year of age who are afflicted with whooping cough about 50.5 per cent recover. A hospital has 27 children less than a year old who are afflicted with this disease. Establish the 5 per cent level of significance as to the number of recoveries and state carefully what you have found.

21. In the registration area of the United States in 1931, 51 per cent of the births were males. In a certain city in 1931, 100 babies were born. (a) What is the probability of as few as 45 females? (b) As many as 60 females? (c) What is the probability of exactly 45 females? (d) What is the probability of exactly 60 females?

106. GRADUATION OF A DISTRIBUTION BY THE NORMAL CURVE

In this book we have frequently called attention to the fact that the distributions of observed data that we have analyzed are samples of a larger population or universe. It has been pointed out that the irregularities of the distributions may be due to a paucity of the data or to fluctuations in sampling. The frequency curve is assumed to represent generalized experience of data of a given type on the assumptions (1) that N has been greatly increased and (2) that the class intervals have been indefinitely diminished. By fitting a curve to the observed data we have opportunity to compare observation with idealization and to note the variations due to sampling.

If a mound-shaped frequency distribution is reasonably symmetrical, the normal curve may approximately represent it. Of course if a distribution is decidedly skew, a normal curve is not expected to fit the data. Our problem in this section is to explain the steps in determining the theoretical frequencies of a distribution, assuming that they follow a normal curve. As was implied in the derivation of the normal curve we assume that:

1. The mean and the standard deviation of the curve are equal to M and $\sigma_{adj.}$ of the observed data.
2. The area under the curve equals the area of the histogram.

It follows from the first assumption that *the first step* in fitting a normal curve to a distribution of observed data is to *compute* M and $\sigma_{adj.}$

A. Graduation by Ordinates. The following table of the graduation of the distribution of the heights of colored soldiers (see p. 168)

TABLE 93. GRADUATION BY THE NORMAL CURVE: ORDINATES

X	*Observed* $f(x)$	$x = X - M$	$t = \frac{x}{\sigma}$	$\phi(t)$	*Theoretical* $f(x) = 1894.6346\phi(t)$
(1)	(2)	(3)	(4)	(5)	(6)
148.5	2	− 23.39	− 3.44	.0011	2.1
150.5	9	21.39	3.15	.0028	5.3
152.5	13	19.39	2.85	.0069	13.1
154.5	23	17.39	2.56	.0151	28.6
156.5	56	15.39	2.26	.0310	58.7
158.5	88	13.39	1.97	.0573	108.6
160.5	162	11.39	1.68	.0973	184.4
162.5	318	9.39	1.38	.1540	291.8
164.5	468	7.39	1.09	.2203	417.4
166.5	564	5.39	0.79	.2920	553.2
168.5	665	3.39	0.50	.3521	667.1
170.5	708	− 1.39	− 0.20	.3910	740.8
172.5	749	+ 0.61	+ 0.09	.3973	752.7
174.5	747	2.61	0.38	.3712	703.3
176.5	586	4.61	0.68	.3166	599.8
178.5	469	6.61	0.97	.2492	472.1
180.5	314	8.61	1.27	.1781	337.4
182.5	207	10.61	1.56	.1182	224.0
184.5	133	12.61	1.85	.0721	136.6
186.5	70	14.61	2.15	.0396	75.0
188.5	38	16.61	2.44	.0203	38.5
190.5	22	18.61	2.74	.0094	17.8
192.5	15	20.61	3.03	.0041	7.8
194.5	10	22.61	3.33	.0016	3.0
196.5	3	24.61	3.62	.0006	1.1
198.5	2	• 26.61	3.91	.0002	0.4
Total	6,441				6,440.6

will show the steps in the process. For the distribution in question we have previously computed $M = 171.89$, $\sigma_{adj.} = (3.3996)2$. Applying equation (8), the theoretical frequencies are given by:

$$y = \frac{(6441)2}{2(3.3996)}\phi(t) = 1894.6346\phi(t)$$

The values of t which correspond to the given values of X are most easily found by multiplying x by $1/\sigma_{adj.}$, and in this case:

$$\frac{1}{\sigma_{adj.}} = 0.147076$$

The following steps are recommended as the proper procedure in fitting a normal curve by ordinates.

1. Compute M, $\sigma_{adj.}$, and $1/\sigma_{adj.}$.
2. Using equation (8), write the equation of the theoretical frequencies.
3. Write down columns (1) and (2), giving class marks and frequencies, of the table upon which the computations are to be carried out.
4. Compute values of x for column (3).
5. Compute values of t for column (4).
6. Write down values of $\phi(t)$ from the table in Appendix B.
7. Compute the theoretical frequencies from the equation found in step 2.

B. Graduation by Areas. The graduation of a distribution by areas depends upon a few notions that we have not yet sufficiently clarified. Since [see page 406]

$$A_Y\Big]_{X_1}^{X_2} = A_y\Big]_{x_1}^{x_2} = Nw \cdot A_\phi\Big]_{t_1}^{t_2}$$

and further, since

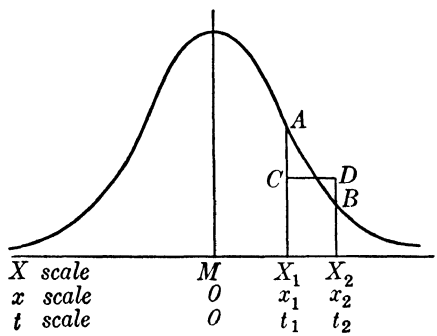
Nw units of area

represent N variates,

then

$Nw \cdot A_\phi\Big]_{t_1}^{t_2}$ units of area

represent $N \cdot A_\phi\Big]_{t_1}^{t_2}$ variates.



We shall indicate the *increment of area* under $\phi(t)$ between t_1 and t_2 by ΔA. The theoretical frequencies will then be computed by $N \cdot \Delta A$.

By this means we are able to find the theoretical frequencies of the various classes (to which the incremental areas under the curve correspond) and compare them to the observed frequencies (to which the rectangular areas of the histogram correspond). That is, we compare, for example, the areas X_1ABX_2 and X_1CDX_2, or the frequencies which they represent.

TABLE 94. GRADUATION BY THE NORMAL CURVE: AREAS

Class lower limit: l_x (1)	*Observed* $f(x)$ (2)	$l_x - M$ (3)	$t = \frac{l_x - M}{\sigma_{adj.}}$ (4)	$A\phi\rfloor_{-\infty}^{t}$ (5)	ΔA (6)	*Theoretical* $f(x) = N \cdot \Delta A$ (7)
147.5	2	− 24.39	− 3.59	.0002	.0003	1.9
149.5	9	22.39	3.29	.0005	.0008	5.2
151.5	13	20.39	3.00	.0013	.0022	14.2
153.5	23	18.39	2.70	.0035	.0045	29.0
155.5	56	16.39	2.41	.0080	.0090	58.0
157.5	88	14.39	2.12	.0170	.0174	112.1
159.5	162	12.39	1.82	.0344	.0286	184.2
161.5	318	10.39	1.53	.0630	.0463	298.2
163.5	468	8.39	1.23	.1093	.0643	414.2
165.5	564	6.39	0.94	.1736	.0842	542.3
167.5	665	4.39	0.65	.2578	.1054	678.9
169.5	708	2.39	0.35	.3632	.1129	727.2
171.5	749	− 0.39	− 0.06	.4761	.1187	764.5
173.5	747	+ 1.61	+ 0.24	.5948	.1071	689.8
175.5	586	3.61	0.53	.7019	.0948	610.6
177.5	469	5.61	0.83	.7967	.0719	463.1
179.5	314	7.61	1.12	.8686	.0521	335.6
181.5	207	9.61	1.41	.9207	.0357	229.9
183.5	133	11.61	1.71	.9564	.0209	134.6
185.5	70	13.61	2.00	.9773	.0120	77.3
187.5	38	15.61	2.30	.9893	.0059	38.0
189.5	22	17.61	2.59	.9952	.0028	18.0
191.5	15	19.61	2.88	.9980	.0013	8.4
193.5	10	21.61	3.18	.9993	.0004	2.6
195.5	3	23.61	3.47	.9997	.0002	1.3
197.5	2	25.61	3.77	.9999	.00008	0.5
199.5	0	27.61	4.06	.99998	.00000	0.0
Total	6,441					6,439.6

We shall illustrate the procedure by graduating the distribution of the heights of colored soldiers (see Table 94) for which we have found:

$$M = 171.89, \quad \sigma_{adj.} = (3.3996)2, \quad \text{and} \quad \frac{1}{\sigma_{adj.}} = .147076.^{1}$$

[1] The question that naturally presents itself to the thoughtful student at this point is: What is the criterion to determine the goodness of fit of a theoretical curve to an observed distribution? We regret that the answer to this important question takes us beyond the scope of this text. We can refer the reader to page 78 of Rietz and others, *Handbook of Mathematical Statistics*, and to Karl Pearson's *Tables for Statisticians*, Pt. I. These references will give a brief discussion of Pearson's Chi-square test. For fuller information we refer the reader to Pearson's original paper in *Philosophical Magazine*, Vol. 50, ser. 5 (1900), pp. 157–75.

In the graduation of a distribution by the normal curve, using areas, we shall find it convenient to follow the following steps.

1. Compute M, $\sigma_{adj.}$, and $1/\sigma_{adj.}$.
2. Write down columns (1) and (2) of the table giving lower class-limits and frequencies. Note that the classes are defined by their lower limits, l_x.
3. Express the lower limits as deviations from M: $l_x - M$. This gives column (3) of the table.
4. Express the deviations from M in units of t: $t = \dfrac{l_x - M}{\sigma_{adj.}}$. This gives column (4) of the table.
5. Using table of $\phi(t)$ in Appendix B, prepare column (5) of the table:

$$A_\phi\Big]_{-\infty}^{t}$$

It will be noted that the desired areas are found by subtracting the values in the table from 0.5000 for $t < 0$, and by adding the values in the table to 0.5000 for $t > 0$.

6. By subtracting each area in column (5) from the area immediately beneath it we compute ΔA. This gives column (6).
7. Compute the theoretical frequencies, $N \cdot \Delta A$.

EXERCISES

1. Graduate by ordinates and by areas the distribution of chest measurements which is given in Exercise 10, page 168.

2. Graduate the distribution of the heights of college men given in (a) of Exercise 1, page 54. Use areas.

3. Graduate by areas the distribution of the head breadths given in Exercise 2, page 54.

4. Find the equation of the distribution of pulse beats which is found in Table 29 (p. 165), assuming normality.

5. Plot the normal curve and the frequency polygon for the theoretical and the observed distributions given in Table 93. Do the same for the distributions in Table 94.

MISCELLANEOUS EXERCISES

1. If $Y_X = \dfrac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{(X-MX)^2}{2\sigma_X^2}}$, show that $Y_{AX} = \dfrac{1}{A}Y_X$.

2. If Y_X has the value given in Exercise 1, find Y_{AX+B} in terms of Y_X.

3. Three per cent of all children are left-handed. In a group of 1,000 children what is the probability that as few as 20 will be left-handed? That as many as 40 will be left-handed? Establish the number of children at the 5 per cent level of significance.

4. Based upon the Mendelian hypothesis, it is expected that, on crossing a certain type of pea, 25 per cent of the seeds will be green. An experiment on this type of pea gave 4,960 yellow and 1,840 green seeds. Is the divergence within the 5 per cent level of significance?

5. Show that the frequency curve

$$y = y_o\left(1 - \frac{x^2}{a^2}\right)^{pa}$$

is symmetrical.

6. Plot on the same axes frequency curves of the form given in Exercise 5 when (1) $a = 5$, $p = 2$; (2) $a = 5$, $p = 10$; (3) $a = 5$, $p = 100$. Assume $y_o = 100$ in each case.

7. Show that as a and p increase without limit but in such a way that a/p is constant and equal to h^2, the curve given in Exercise 5 approaches the normal form

$$y = y_o e^{-x^2/h^2}$$

8. We replace the single constant a in Exercise 5 after factoring by a_1 and a_2 thus obtaining

$$y = y_o\left(1 + \frac{x}{a_1}\right)^{pa_1}\left(1 - \frac{x}{a_2}\right)^{pa_2}$$

which is skew. Plot on the same axes this curve when (1) $a_1 = 4$, $a_2 = 5$, $p = 10$; (2) $a_1 = 10$, $a_2 = 5$, $p = 0.3$. Assume $y_o = 100$.

9. Show that as a_2 increases without limit, a_1 remaining constant, the formula in Exercise 8 approaches the form

$$y = y_o\left(1 + \frac{x}{a_1}\right)^{pa_1} e^{-px}$$

10. Draw the curve in Exercise 9 when $y_o = 25$, $a_1 = 12$, $p = 1.3$.

Chapter 13

THE THEORY OF SAMPLING: MEASURES OF RELIABILITY

107. INTRODUCTION

We may regard the numerical description of any mass of statistical data from two points of view. We may regard the description as an end in itself, a mere summary of our measurements, or we may regard it as a *sample* drawn from a larger group which we call the *parent population* or the *universe.*

Usually, the larger point of view obtains, that of forming judgments of the universe from a study of the sample. In some cases it is impossible to measure the entire universe, and in other cases it is impracticable to do so. Even if such a goal as measuring the entire universe was possible of attainment, the added expense in time and labor would be *an unnecessary luxury.* For, by carefully selecting a sample, excellent estimates of the statistical parameters of the universe can be obtained.

The statistician is, therefore, generally forced to work with samples. We compute the mean of the sample and use this mean as a basis for *estimating* the mean of the universe. Similarly, we use the dispersion of the sample as a basis for estimating the dispersion of the universe; and so on. Naturally, we must then attempt to state the degree of confidence we can attach to our estimates. This we do in terms of probability.

It is obvious that in order to make a good estimate of the universe, we must have a good sample, a representative sample. Securing such a sample is not always an easy task, but generally it can be done. The procedures employed in securing such samples are beyond the scope of this book. In what follows, when we use the term sample, we mean a *statistical sample* wherein any one individual in the parent population is just as likely to be included as any other. Such a sample is often called a *random sample.*

This process of generalizing statistical results, of making inferences regarding the universe from the study of the sample, is called *statistical induction.* Obviously, it is a problem of supreme importance. Karl Pearson has called it "the fundamental problem of practical statistics."

108. THE PROBLEM OF THIS CHAPTER

We have spent no little time in the preceding chapters with questions relating to the numerical description of a mass of data as an end in itself. We have seen that it is possible to describe succinctly a mass of numerical data. The essence of the data may be condensed to four measures: (1) the mean, (2) the dispersion, (3) the skewness, and (4) the excess. For example, given the measurements of the heights of 1,000 men, we are able to give a numerical description of the 1,000 measurements. They may show an arithmetic mean of 67.5 inches, a standard deviation of 2.5 inches, a coefficient of skewness, α_3, of 0.036, and an excess, $\alpha_4 - 3$, of 0.123. If our problem is limited to a characterization of the 1,000 measurements, our problem is fairly completely solved. In characterizing a mass of data by means of a few statistical constants, we are able to comprehend the significant facts of the mass which might not otherwise be possible.

If we adopt the second and broader point of view and consider the 1,000 measurements as a representative sample and are concerned with using the properties of the sample in order to make inferences about the parent population from which the sample is chosen, it is clear that we cannot speak with meticulous certainty concerning the computed statistical constants and, as a consequence, our language should be modified. Another sample of 1,000 measurements of the heights of men chosen in a similar manner will probably yield at least slightly different statistical constants. In other words, *these so-called statistical constants show variation as we move from sample to sample.*

While the statistical constants computed from successive samples show variation, it must not be inferred that the variation is unlimited. As a matter of fact the statistical constants computed from moderately large random samples selected from a larger group show an uncanny stability. It is due to this remarkable and *measurable*

stability of the statistical constants computed from sample to sample that we may make inferences from a relatively small set of observed data. *A measure of the stability of a statistical constant is often called a measure of its reliability.*

The so-called statistical constant derived from the analysis of a sample is frequently called by some writers, following R. A. Fisher, a *statistic*, and the corresponding quantity belonging to the universe a *parameter*. A *statistic* is thus an estimate of a *parameter*. For a given universe a parameter is fixed but the statistic may vary from sample to sample.

It will be the problem of this chapter to define a range of variation about the statistical *parameter* of the universe within which fluctuations of the *statistics*, due to pure chance, may be expected to occur according to definite probabilities. It must be borne in mind that the variations due to a multiplicity of factors other than pure chance can in no way be accounted for by the sampling formulas that we shall discuss. The variations we are considering "are the resultant effect of a complex of forces which cannot be traced, still less measured, and which have been happily described as that 'mass of floating causes generally known as chance.'" If the variations are greater than can be accounted for by chance, the significance of the variation should, if possible, be accounted for and explained by the observer.

We may meet problems that fall into two broad categories. In the first category the parent universe may be known and we may wish to establish whether or not a statistic of a sample falls within a predetermined range of variation. (In this case the parent universe is generally finite.) For example, a manufacturer of some article may have examined a large number of a given type of product, and thus may have been able to adopt rather rigid specifications for the product. A sample is selected for a test. Does the sample fall within the tolerance limits demanded by the universe?

In the second category the parent universe is unknown and we wish to estimate its parameters by finding the statistics of the sample, and to measure the reliability (or degree of confidence) we may place in the estimates. By far, most problems that occur in the applications of the theory of sampling belong in this category. In this case the universe is generally considered as infinite.

In most cases that arise, whether the universe be known or un-

known, stated in rather general terms the question is: How well does the sample describe the universe? More precisely: How much shall we allow the values of the statistical constants obtained from the sample to vary to describe the parent universe?

109. THE STANDARD DEVIATION IN CLASS FREQUENCIES

TABLE 95A

(Parent Population)

Class	$F(x)$
1	3,000
2	6,000
3	13,000
4	18,000
5	20,000
6	19,000
7	12,000
8	7,000
9	2,000
Total	100,000

TABLE 95B

(Sample of 1,000 Theoretical Frequencies)

Class	$f(x)$
1	30
2	60
3	130
4	180
5	200
6	190
7	120
8	70
9	20
Total	1,000

Suppose the frequency distribution of some single characteristic is given by Table 95A. The relative frequencies of the several classes are 3/100, 6/100, 13/100, etc. We choose from this homogeneous population a sample of 1,000. The "expected" distribution of the sample, by Section 95, would be that given by Table 95B. We know of course from experience that the theoretically "expected" frequencies would differ from those that would result from experiment just as I know that if I toss a coin 100 times I "expect" 50 heads and 50 tails whereas I may actually get 48 heads and 52 tails. And from my experience with coin-tossing experiments I am not shocked by this result.

Suppose that we should obtain a large number of samples of 1,000 observations, each taken under the same essential conditions. A class frequency, say that of Class 3, will vary from sample to sample. These values will form a frequency distribution. The variations, called "variations due to sampling" or "variations due to sampling errors," can frequently be accounted for and explained. Such a

question as, "What is the variation that would occur in Class 3 if we obtained a large number of samples of 1,000 observations from the population in Table 95A?" we can answer approximately.

To answer this question we consider any observation as a *trial*, and a *success* if an observation falls in the class. Thus the probability of an observation falling in Class 3 is $p = 13/100$, and the probability of an observation not falling in the class is $q = 87/100$. And we have the standard deviation of the frequency of this class to be theoretically $\sqrt{1{,}000(.13)(.87)} = 10.6$. So that we should expect $Np \pm 3\sqrt{Npq}$ or 130 ± 32 observations as setting the limits of the frequency of Class 3 of the sample of 1,000.

If the probable error rather than the standard deviation is taken as the measure of the variation, then the probable error of the frequency of Class 3 is $0.6745\sqrt{Npq}$ or $0.6745(10.6) = 7.1$. Hence, if many random samples of 1,000 observations were taken from the population of Table A, we should expect theoretically the frequency of Class 3 of the sample to fall within 130 ± 7 about half the time.

If plus and minus three times the standard deviation of the expected frequency be taken as the variation in the frequency that may be allowed due to sampling, then if many samples of 1,000 ob-

TABLE 96C

Class	$f(x)$
1	$30 \pm 3(5.4)$
2	$60 \pm 3(7.5)$
3	$130 \pm 3(10.6)$
4	$180 \pm 3(12.1)$
5	$200 \pm 3(12.6)$
6	$190 \pm 3(12.4)$
7	$120 \pm 3(10.3)$
8	$70 \pm 3(8.1)$
9	$20 \pm 3(4.4)$
Total	1,000

TABLE 96D

Class	$f(x)$
1	25
2	75
3	175
4	200
5	210
6	170
7	80
8	50
9	15
Total	1,000

servations are actually taken from the population of Table 95A, we might obtain frequency distributions with the variation in the class frequencies as indicated in Table 96C. So that if we were sampling from Table 95A and should secure a sample with the frequencies given by Table 96D, we would be inclined to suspect

that randomness went awry since the frequencies in Classes 3 and 7 are outside the limits set by Table 96C.

In general, if the frequency of the kth class of the parent distribution of population S be $F_k(x)$, then the probability of an observation's falling in that class is $p_k\left(=\frac{F_k(x)}{S}\right)$ and the probability of the observation's not falling in that class is $q_k\left(=1-\frac{F_k(x)}{S}\right)$. So, when a sample of N is chosen the expected frequency of the kth class of the sample distribution is Np_k with the standard deviation $\sqrt{Np_kq_k}$.

In applications we do not know the parent population and hence the true value of p_k is unknown. Let $f_k(x)$ be the observed frequency of the kth class of the sample. If N is fairly large we accept $f_k(x)/N$ as an approximation to p_k. Then we have

$$\sigma^2_{f_k(x)} = N \cdot \frac{f_k(x)}{N}\left(1-\frac{f_k(x)}{N}\right) = f_k(x)\left(1-\frac{f_k(x)}{N}\right)$$

Hence the frequency [1] of the kth class may be written with its probable error as

$$\text{frequency of } k\text{th class} = f_k(x) \pm 0.6745\sqrt{f_k(x)\left(1-\frac{f_k(x)}{N}\right)}$$

This means that if a sample of N is taken from some unknown parent distribution, the chances are even that the observed frequency of the kth class, $f_k(x)$, will not differ from the expected frequency of the kth class by more than $\pm\, 0.6745\sqrt{f_k(x)\left(1-\frac{f_k(x)}{N}\right)}$.

If each class frequency of a distribution of N variates is divided by N, we obtain a distribution of *relative frequencies or percentages.* As a corollary to the theorem for finding $\sigma_{f_k(x)}$ we can immediately derive a formula for finding the variation in the *relative frequency of the* kth *class,* $\sigma_{p_k(x)}$, where $p_k(x) = f_k(x)/N$.

From Exercise 21 on page 148 we have $\sigma_{AX} = A\sigma_X$. Employing this theorem we have

[1] See Rietz, H. L., *Mathematical Statistics*, pp. 119–122, for a formula which gives a closer approximation.

$$\sigma_{p_k(x)} = \frac{1}{N}\sigma_{f_k(x)} = \frac{1}{N}\sqrt{f_k(x)\left[1 - \frac{f_k(x)}{N}\right]}$$

$$= \sqrt{\frac{1}{N}\frac{f_k(x)}{N}\left[1 - \frac{f_k(x)}{N}\right]} = \sqrt{\frac{p_k(x)q_k(x)}{N}}$$

where $q_k(x) = 1 - p_k(x)$.

This formula, when used in its broad meaning to measure the variation in a percentage, is usually written

$$\sigma_p = \sqrt{\frac{pq}{N}}$$

where $q = 1 - p$.

Example. Suppose that of a large number of men examined for military service about 70 per cent have been accepted. If the same standards are imposed in future examinations, what are the limits of percentage acceptances expected from a sample of 1,000?

Solution. We have $p = 0.70 \quad q = 0.30 \quad N = 1{,}000$

$$\sigma_p = \sqrt{\frac{(0.70)(0.30)}{1{,}000}} = 0.014 = 1.4 \text{ per cent}$$

Adopting $\pm 3\sigma_p$ as the limits of the percentage accepted, we should expect the percentage accepted to vary from 70 − 4.2 per cent to 70 + 4.2 per cent. That is, we should expect from 65.8 to 74.2 per cent of the men examined to be accepted.

110. AN EXPERIMENT IN SAMPLING

In order to clarify the problem of the sampling process, let us consider the parent universe of 64 variates distributed according to the point binomial $64(\frac{1}{2} + \frac{1}{2})^6$. Table 97 exhibits the parent distribution in tabular form. We indicate the mean and the standard deviation of the universe by M_u and σ_u respectively.

TABLE 97

X	$f(x)$
0	1
1	6
2	15
3	20
4	15
5	6
6	1
Total	64

For this *universe* we have:

$$M_u = np = 6(\tfrac{1}{2}) = 3$$

$$\sigma_u = \sqrt{npq} = \sqrt{6(\tfrac{1}{2})(\tfrac{1}{2})} = 1.225$$

$$\alpha_3 = \frac{q - p}{\sigma} = 0$$

In order that we may draw random samples from the given parent population we prepare 64 cards in the following manner. On 1 card

we write $X = 0$ and $X^2 = 0$; on 6 cards we write $X = 1$ and $X^2 = 1$; on 15 cards we write $X = 2$ and $X^2 = 4$; on 20 cards we write $X = 3$ and $X^2 = 9$; and so on for the entire parent distribution. We now have a parent population of 64 members, one card for each individual, from which we may draw random samples. Suppose we draw samples of 10 cards. The remaining 54 cards constitute a sample of 54 cards. With each drawing we therefore obtain samples of $N = 10$ and $N = 54$. The sum of X for all 64 cards is 192 and the sum of X^2 for all 64 cards is 672. We shuffle the cards well and take a sample of 10 cards. We total the values of X and of X^2 on the 10 cards and find for the first sample of 10 that $\Sigma X = 26$ and $\Sigma X^2 = 100$. For the first sample of 10 we now find $M = 2.6$ and $\sigma = 1.8$. We thus have one sample mean and one sample standard deviation for $N = 10$. For $N = 54$ we also have $\Sigma X = 192 - 26 = 166$ and $\Sigma X^2 = 672 - 100 = 572$, from which we compute $M = 3.1$ and $\sigma = 0.99$. We thus have for $N = 54$ one sample mean and one sample standard deviation. We place the cards again on the pack, shuffle them well again, and draw 10 cards, from which we again compute the sample means and the sample standard deviations. We can continue this process and select as many samples as we please. Obviously ${}_{64}C_{10}$ distinct samples can be secured. We show below the distributions of 100 actual sample means for the case in which $N = 10$ and the case in which $N = 54$. We denote by Z any sample mean and its frequency by $f(z)$. (See Table 98.)

Distribution (a), which has 100 sample means, was derived by drawing samples of 10 variates from the previously described parent population of 64 variates and computing the means of the samples drawn. Distribution (b), with samples of 54 variates, was similarly derived. Each distribution is therefore a distribution of sample means that has its mean (the mean of the means), its standard deviation (the standard deviation of the means), its skewness (the skewness of the means), and so on. It is the standard deviation of the means in which we are especially interested, for it gives a measure of the variability of the distribution of means.

We shall leave it as an exercise for the student to verify the following values:

$N = 10$	$N = 54$
$M_Z = 2.997$	$M_Z = 3.00$
$\sigma_Z = 0.298$	$\sigma_Z = 0.078$

TABLE 98

(a)	
$N = 10$	
Z	$f(z)$
2.2	1
2.3	1
2.4	2
2.5	3
2.6	5
2.7	6
2.8	9
2.9	14
3.0	20
3.1	13
3.2	8
3.3	7
3.4	4
3.5	3
3.6	1
3.7	2
3.8	1
Total	100

(b)	
$N = 54$	
Z^*	$f(z)$
3.15	1
3.13	1
3.11	2
3.09	3
3.07	5
3.06	6
3.04	9
3.02	14
3.00	20
2.98	13
2.96	8
2.94	7
2.93	4
2.91	3
2.89	1
2.87	2
2.85	1
Total	100

* These are rounded values.

Figure 59 represents the curve for the parent distribution and Figure 60 the ordinates of the distribution of sample means for $N = 10$.

FIGURE 59

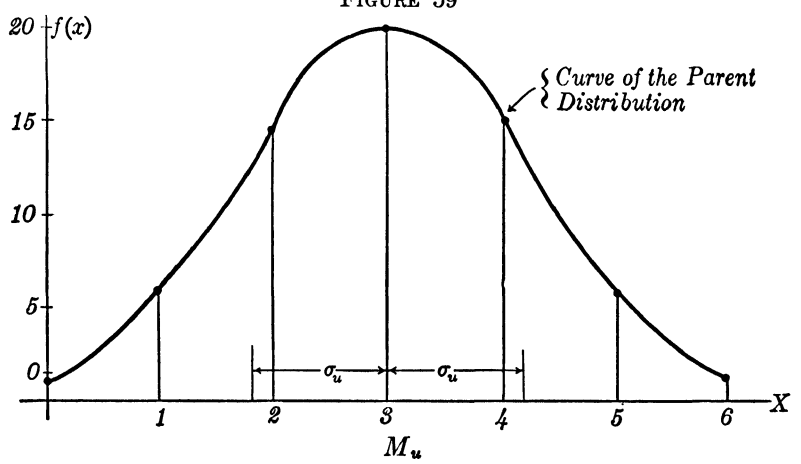
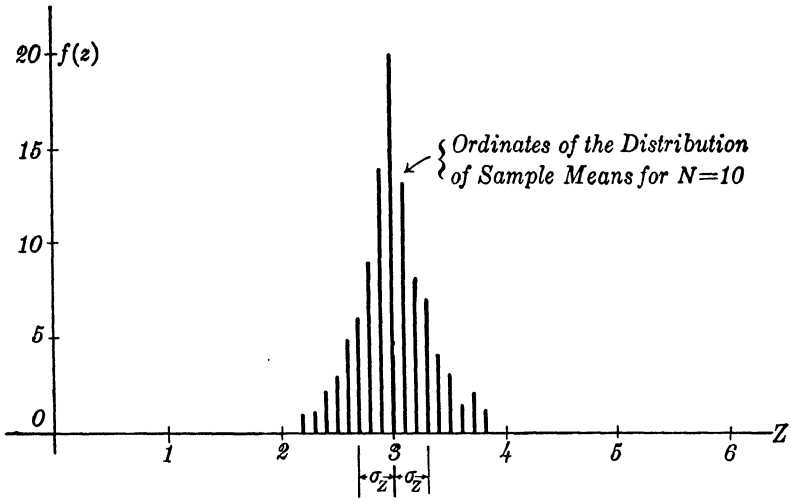


FIGURE 60



It will be observed that the sample means are approximately normally distributed above and below $M_Z = 2.997$, but with a dispersion much less than that of the parent population. In the next section we shall derive some theorems that should explain these phenomena.

The following exercises are given primarily to prepare the student for a facile reading of the succeeding section. The various numbers should therefore be solved in detail.

EXERCISES

1. Consider the parent population of 5 variates: X_1, X_2, X_3, X_4, X_5. Write down the 10 distinct samples of 3 variates that may be drawn. For example, X_1, X_2, X_3; X_1, X_2, X_4.

2. Let Z_i represent the ith sample mean and write down the 10 distinct sample means for the parent population in Exercise 1. For example,

$$Z_1 = \frac{(X_1 + X_2 + X_3)}{3}$$

3. Show that for the sample means in Exercise 2:

$$M_Z = \frac{\Sigma Z_i}{10} = \frac{\Sigma X_i}{5} = M_X$$

State in words the theorem of this formula.

4. Show that:

$$(\Sigma X_i)^2 = \Sigma X_i^2 + 2\Sigma X_i X_j$$

where

$$\Sigma X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

5. For the values of Z_i found in Exercise 2 show that:

$$\frac{\Sigma Z_i^2}{10} = \frac{1}{15}[\Sigma X_i^2 + \Sigma X_i X_j]$$

6. Using the relationship in Exercise 4 show that:

$$\frac{\Sigma Z_i^2}{10} = \frac{1}{30}[\Sigma X_i^2 + (\Sigma X_i)^2]$$

7. From equation (7) of Chapter IV we have $\sigma_Z = \sqrt{\frac{\Sigma Z_i^2}{10} - M_Z^2}$.

Use this relationship and those established in Exercises 3 and 6 above to show that, for the distribution of means here considered:

$$\sigma_Z = \frac{\sigma_X}{\sqrt{6}}$$

111. THE DISTRIBUTION OF MEANS

Let us now consider the general problem of characterizing the distribution of sample means derived by drawing samples of N variates from a parent population of S variates. Obviously ${}_SC_N$ distinct samples may be drawn. Each sample has its mean and the ${}_SC_N$ samples give us a distribution of sample means.

We shall undertake to characterize this distribution of means as we should characterize any distribution, that is, by finding its mean, its standard deviation, its skewness, and so on.

A. The Mean of the Means. Let the parent universe be denoted by $X_1, X_2, X_3, \ldots, X_S$. Denoting any sample mean by Z_i, we have:

$$\left.\begin{aligned} Z_1 &= \frac{1}{N}[X_1 + X_2 + \cdots + X_{N-1} + X_N] \\ Z_2 &= \frac{1}{N}[X_1 + X_2 + \cdots + X_{N-1} + X_{N+1}] \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_{{}_SC_N} &= \frac{1}{N}[X_{S-N+1} + X_{S-N+2} + \cdots + X_{S-1} + X_S] \end{aligned}\right\} \quad (1)$$

We desire to find the mean of this distribution of sample means. We must find $\Sigma Z \div {}_SC_N$. Note that each parenthesis contains N terms and that the ${}_SC_N$ lines contain $(N \cdot {}_SC_N)$ terms which are not all distinct. One X occurs as frequently as another. Hence each of the S X's occurs $\left(\frac{N}{S} \cdot {}_SC_N\right)$ times. That is:

$$M_Z = \frac{1}{{}_SC_N}\Sigma Z_i = \frac{1}{N}\left[\frac{N}{S} \cdot {}_SC_N \Sigma X_i\right] \div {}_SC_N = \frac{1}{S}\Sigma X_i = M_X \quad (2)$$

We may express this important result as follows:

Theorem: *The mean of the* ${}_SC_N$ *sample means formed by selecting samples of* N *variates from a parent population of* S *variates is equal to the mean of the* S *variates.*

Stated less accurately, we may say that *the mean of the distribution of means is equal to the mean of the parent universe:* $M_M = M_u$.

B. The Standard Deviation of the Means. We shall now proceed to find the standard deviation of the ${}_SC_N$ sample means with which we can measure the variability of the distribution of means. We should recall that the standard deviation of the parent universe is given by

$$\sigma_X = \sqrt{\frac{\Sigma X^2}{S} - M_X^2}$$

and that the standard deviation of the distribution of sample means is given by:

$$\sigma_Z = \sqrt{\frac{\Sigma Z^2}{{}_SC_N} - M_Z^2} \quad (3)$$

Since M_Z is known, σ_Z can be determined if we can find ΣZ^2.

From equations (1) we have:

$$Z_1^2=\frac{1}{N^2}[X_1^2+X_2^2+\cdots+X_N^2]+\frac{2}{N^2}[X_1X_2+X_1X_3+\cdots+X_{N-1}X_N]$$

$$Z_2^2=\frac{1}{N^2}[X_1^2+X_2^2+\cdots+X_{N-1}^2+X_{N+1}^2]+\frac{2}{N^2}[X_1X_2+X_1X_3+\cdots+X_{N-1}X_{N+1}]$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$Z_{{}_SC_N}^2=\frac{1}{N^2}[X_{S-N+1}^2+\cdots+X_S^2]+\frac{2}{N^2}[X_{S-N+1}X_{S-N+2}+\cdots+X_{S-1}X_S]$$

To find the sum of these ${}_SC_N$ squared means we note that the sum of the parentheses containing terms of the type X_i^2 may be found as follows: Each parenthesis contains N terms and the ${}_SC_N$ parentheses contain $(N\cdot {}_SC_N)$ terms which are not all distinct. One X^2 occurs as frequently as another. Hence each of the given S X^2's occurs $(N\cdot {}_SC_N\div S)$ times.

To sum the parentheses containing the cross-product terms of the type X_iX_j, note that each parenthesis contains ${}_NC_2$ terms and the ${}_SC_N$ parentheses contain $({}_NC_2\cdot {}_SC_N)$ terms which are not all distinct. One cross-product term occurs as frequently as another. Since we can get ${}_SC_2$ cross-product terms from the given S letters, each of the $({}_NC_2\cdot {}_SC_N)$ cross-product terms must occur $({}_NC_2\cdot {}_SC_N\div {}_SC_2)$ times. Therefore:

$$\Sigma Z_i^2=\frac{1}{N^2}\left[\frac{N\cdot {}_SC_N}{S}\Sigma X_i^2+\frac{{}_NC_2\cdot {}_SC_N}{{}_SC_2}2\Sigma X_iX_j\right]$$

and

$$\frac{\Sigma Z_i^2}{{}_SC_N}=\frac{1}{N^2}\left[\frac{N}{S}\Sigma X_i^2+\frac{N(N-1)}{S(S-1)}2\Sigma X_iX_j\right]$$

Since

$$(\Sigma X_i)^2=\Sigma X_i^2+2\Sigma X_iX_j$$

we have:

$$2\Sigma X_iX_j=(\Sigma X_i)^2-\Sigma X_i^2$$

Hence:

$$\frac{\Sigma Z_i^2}{{}_SC_N}=\frac{S-N}{N(S-1)}\frac{\Sigma X^2}{S}+\frac{S(N-1)}{N(S-1)}\left(\frac{\Sigma X}{S}\right)^2$$

Substituting this in (3), recalling from (2) that $M_z = M_x$, we have upon simplifying:

$$\sigma_z = \sigma_x \sqrt{\frac{S - N}{N(S - 1)}} \tag{4}$$

Since, in general, S is very large when compared with N, we can obtain a simpler relationship if we assume that S is infinite. For this case we obtain [1]

$$\sigma_z = \frac{\sigma_x}{\sqrt{N}} \tag{5}$$

in which, we repeat for emphasis, N is the *number of variates in the sample* and σ_x is the standard deviation of the *parent universe.* In fact, in (4) and (5) we may replace σ_z and σ_x by σ_M and σ_u. As the constants describing the parent universe are usually not known, formulas (4) and (5) are apparently of value only theoretically. Since we have stated that it is our problem to make certain inferences about the parent universe from a consideration of the sample we shall see in a later section how (5) will assist us in doing it. Experiment justifies our making the assumption that *the standard deviation of the parent universe is approximately equal to the standard deviation of the sample, the goodness of the approximation increasing as N is increased.* This assumption makes possible our expressing the formula for the standard deviation of the mean in a workable form. We have, finally

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \text{ (approximately)} \tag{6}$$

where M is the mean of the sample, σ is the standard deviation of the sample, and N is the number of variates in the sample. That is:

$$\text{the standard deviation of the arithmetic mean} = \frac{\text{the standard deviation of the sample}}{\sqrt{\text{the number in the sample}}}$$

C. The Probable Error of the Mean. In Chapter 4 we insisted that the standard deviation is an excellent measure of the variability of a distribution. We also insist that the standard deviation of the

[1] This is easily seen if we divide numerator and denominator of the quantity under the radical by S and note that N/S and $1/S$ approach zero as S becomes infinite.

mean as computed from (6) is an excellent measure of the variability of the distribution of means. Tradition, however, has been a potent influence in commending the use of the *probable error* to measure the variation in the several statistical constants. In the preceding chapter, we defined the probable error of any measure by the equation:

$$E_X = 0.6745\sigma_X$$

Therefore, the probable error of the mean is defined by the relation:

$$E_M = 0.6745\sigma_M = \frac{0.6745\sigma}{\sqrt{N}} \tag{7}$$

The quantities, σ_M and E_M, are frequently used as measures of the reliability of the arithmetic mean. Since the smaller the variation, the greater the reliability, a small standard deviation of the mean or a small probable error of the mean means "accurate shooting." It is therefore evident from (6) and (7) that the smaller the σ_M or E_M, the greater the reliability in M.

The language of variation used in the preceding paragraph is inverse. We can make the variation direct if we adopt the measure, h, as is done in the theory of errors, for the *index of precision* where h is defined by the equation (see Section 102, p. 399):

$$h_X = \frac{1}{\sigma_X\sqrt{2}}$$

For the distribution of means we have

$$h_M = \frac{1}{\sigma_M\sqrt{2}} = \frac{\sqrt{N}}{\sigma\sqrt{2}} = \frac{1}{\sigma}\sqrt{\frac{N}{2}} \tag{8}$$

as the *index of precision of the mean.* It will be observed from (6), (7), and (8) that the reliability of the mean or the precision of the mean varies as the square root of the number in the sample. That is, the greater the number in the sample, the greater the reliability in the mean. For example, to double the reliability, we must quadruple the frequency.

It is not customary, however, in elementary statistics, to use h_M as the measure of the reliability of the mean. Rather do the workers in applied statistics prefer σ_M or E_M. In fact, it is the custom

(see Section 37, p. 143) to write the probable error of the mean immediately after the computed mean of the sample with a ± sign between them, thus:

$$M_u = M \pm E_M \tag{9}$$

For example, suppose a sample distribution of the heights of 1,000 men shows an arithmetic mean of 67.5 inches and a standard deviation of 2.5 inches. Then:

$$E_M = 0.6745 \frac{2.5}{\sqrt{1000}} = 0.053 \text{ inch}$$

and

$$M_u = 67.5 \pm 0.053 \text{ inches}$$

Since the distribution of sample means collected from a normal parent population is itself normal, this means simply that if a large number of the means of samples of the heights of 1,000 men were collected, half of the sample means would be within 0.053 inch of the mean of the universe $M_u (= M_M)$. Since $M_M \pm 3\sigma_M$ or $M_M \pm 4.5E_M$ includes nearly all the sample means, it is practically certain that no sample mean will differ from the mean of the universe M_u by more than ± 4.5(0.053) inches.

It should be emphasized that the expression $M_u = M \pm E_M$ is not to be interpreted as stating that the true mean of the universe *is somewhere* between $M - E_M$ and $M + E_M$; nor is it to be interpreted as stating that the true mean *probably differs* from the computed sample mean by the amount E_M. It means that, so far as *variation due to pure chance is concerned*, the odds are even that a sample mean M will not differ from the mean of the universe M_u by more than E_M.

If we were to write the arithmetic mean of the universe M_u in the form

$$M_u = M \pm \sigma_M$$

this would signify that the odds are about 2 to 1 that a sample mean M will not differ from M_u by more than σ_M. It does not mean that the odds are 2 to 1 that M_u is within the interval whose end values are $M - \sigma_M$ and $M + \sigma_M$. The probability pertains to the limits of the range embracing M_u. We do not state the probability of M_u lying within these limits for M_u is fixed. Thus, for the heights of the sample of 1,000 men noted above we have

$$\sigma_M = \frac{2.5}{\sqrt{1{,}000}} = \frac{2.5}{31.623} = 0.08 \text{ inch}$$

And we say that the odds are 2 to 1 that the mean of the sample, 67.5 inches, is within 0.08 inches of the mean of the universe. The odds are 95 to 5 (or 19 to 1) that the mean of the sample does not differ from the mean of the universe by more than $\pm$ 1.96(.08) inches, and the odds are 99 to 1 that the sample mean does not differ from the mean of the universe by more than $\pm$ 2.56(.08) inches.

It has thus become customary to write M_u in two different forms: $M_u = M \pm E_M$ and $M_u = M \pm \sigma_M$. In the first case we have "M with a *probable error* of E_M" and in the second case we have "M with a *standard error* of σ_M." To avoid ambiguity the statistician should state definitely what his symbols mean.

ILLUSTRATIVE EXAMPLES

Example 1. A corporation which sells a large number of automobile tires gathered data on the mileage obtained from a given type of tire. A large group of 100,000 users were questioned, and the data analyzed. For this universe of S = 100,000 it was found that M_u = 21,000 miles and σ_u = 2,000 miles.

At a later time in order to compare the quality of the product, the corporation secured data from 1,000 users of the same type of tire. For this sample of N = 1,000 it was found that M = 20,960 and σ = 1,980 miles.

Was the corporation correct in concluding that the quality of the tire was not impaired, or that the variation of M from M_u was not significant?

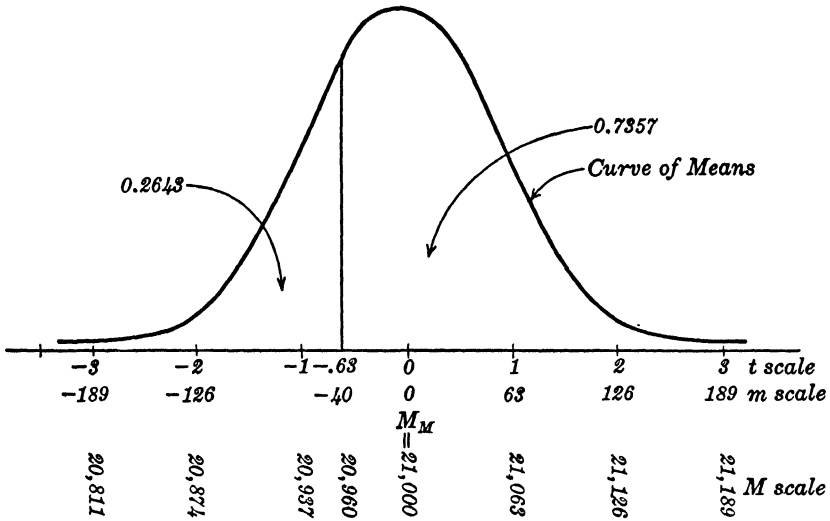
Solution. Translating formula (4) into better symbols, we have

$$\sigma_M = \sigma_u\sqrt{\frac{S-N}{N(S-1)}} = 2{,}000\sqrt{\frac{100{,}000-1{,}000}{1{,}000(100{,}000-1)}}$$

$$= 62.6 \text{ miles} = 63 \text{ miles (rounded)}.$$

Thus for the distribution of means, which is normal, we have $M_M = M_u$ = 21,000 miles and σ_M = 63 miles.

If many such samples were taken we could expect 68.27 per cent, or about two-thirds, of the means to fall within the interval $M_M \pm \sigma_M$. That is, we should expect about two-thirds of the sample means to fall in the interval 21,000 $\pm$ 63 miles, or between 20,937 and 21,063. Since 20,960 is within this interval, we conclude that the quality of the tire is not impaired and that the difference is not statistically significant.



We can look at the problem from another point of view. We express the divergence $M - M_M$ in standard units. We find

$$t = \frac{m}{\sigma_M} = \frac{M - M_M}{\sigma_M} = \frac{20{,}960 - 21{,}000}{63} = -\frac{40}{63} = -.63$$

Looking up the probability table we find

$$A_\phi\Big]_{-\infty}^{-.63} = .5000 - A_\phi\Big]_{0}^{.63} = .5000 - .2357 = .2643$$

We would therefore expect 26 per cent of the sample means to be less than 20,960 miles and 74 per cent to be greater than 20,960 miles. In other words the probability of a sample mean being less in value than 20,960 is 26/100 or 13/50 and greater than 20,960 is 74/100 or 37/50.

Of course we can base our argument on the probable error of the mean instead of the standard error of the mean. We find

$$E_M = .6745\sigma_M = .6745(62.6) = 42 \text{ miles}$$

Then we can state that the chances are even that a sample mean will lie in the interval 21,000 ± 42 or between 20,958 and 21,042. The given mean 20,960 is within this interval, and such a small divergence as 40/42 probable errors from M_M is certainly within the tolerance limits of the most scrupulous.

Example 2. Suppose in the previous example we use formula (5)

$$\sigma_M = \frac{\sigma_u}{\sqrt{N}}$$

Will our results be affected?

Solution. This means, writing formula (4) in the form

$$\sigma_M = \frac{\sigma_u}{\sqrt{N}}\sqrt{\frac{1-\frac{N}{S}}{1-\frac{1}{S}}}$$

S is so large compared to N that we may consider $\frac{N}{S}$ and $\frac{1}{S}$ as negligible. We find for the data at hand

$$\sigma_M = \frac{2{,}000}{\sqrt{1{,}000}} = \frac{2{,}000}{31.63} = 63.2 = 63 \text{ miles (rounded)}$$

This approximation certainly will in no way alter our previous conclusion.

Example 3. If in Example 1 we use formula (6) for computing σ_M, will our conclusion be altered?

Solution:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{1{,}980}{\sqrt{1{,}000}} = \frac{1{,}980}{31.63} = 62.6 = 63 \text{ miles (rounded)}$$

And this approximation will also in no way alter our conclusion.

Example 4. The blood pressure of 10,000 young men of given age was measured and recorded. The analysis of the sample gave $M = 122$ and $\sigma = 9$. Find the standard error and the probable error of the mean, and interpret them. What is the 5 per cent level of significance?

Solution. In this case we do not know the statistics of the universe. Our information about the statistics of the universe must be *inferred* from the statistics of the sample. The mean of the sample is an estimate of the mean of the universe. How reliable is the estimate?

We compute the dispersion of the sample means by (6). We have

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{9}{\sqrt{10{,}000}} = 0.09$$

which indicates the dispersion of the sample means about the universe mean M_u. The universe mean is unknown. However, we can state that the odds are 2 to 1 that the sample mean 122 does not differ numerically from M_u by more than 0.09. Since 99.74 per cent of the sample means vary from M_u by not more than $\pm 3\sigma_M (= 0.27)$, we may conclude that the odds are 99.74 to 0.26, or about 385 to 1, that the sample mean 122 does not differ numerically from M_u by more than 0.27.

$$E_M = 0.6745\sigma_M = .6745(0.09) = 0.06$$

which indicates that the chances are even that the sample mean 122 does not differ numerically from M_u by more than 0.06.

The 5 per cent level of significance is at $\pm 1.96\sigma_M$ or at $\pm 1.96(.09) = \pm 0.176$. That is, the odds are 95 to 5, or 19 to 1, that the sample mean 122 does not differ numerically from M_u by more than 0.176. In other words, if many samples of 10,000 were measured and their means computed, we should expect 95 per cent of the means to lie within the interval $M_u - 0.176$ and $M_u + 0.176$.

EXERCISES

1. Professor Sorenson [1] records an experiment in which fifty samples of fifty men each were taken from *American Men of Science* and the mean age of each sample computed. The means ranged in value from 41.40 years to 51.14 years. He found that this distribution of means was normal with $M_M = 46.34$ years and $\sigma_M = 2.30$ years.

a. What is the estimated mean age of the 2,500 men? Dr. Sorenson gives 46.34 years as the computed mean age of the 2,500 men.

b. How many of the means would you expect to find between $M_M - \sigma_M$ and $M_M + \sigma_M$? Dr. Sorenson found 34, or 68 per cent of them.

c. Compute E_M. What does it mean?

d. How many of the means would you expect to find between $M_M - E_M$ and $M_M + E_M$? Dr. Sorenson found 25, or 50 per cent of them.

e. Consider the 2,500 ages as a sample of all the men, 22,000, whose names appeared in the book. For this large sample Dr. Sorenson gives $M = 46.34$ and $\sigma = 12.46$. Compute E_M and interpret it with regard to the average age of all men in the book.

2. A sample of $N = 625$ gave $E_M = 0.27$. What size sample would be required to give $E_M = 0.09$? 0.045?

3. A distribution of the weights at birth of a sample of 402 infants gave $M = 7.29$ pounds and $\sigma = 1.006$ pounds. Compute E_M and interpret it.

4. A study of "red blood cell count" for 40 normal men gave $M = 4.973$ millions per cu. mm. and $\sigma = 0.332$ millions per cu. mm. Find σ_M and interpret it.

5. For a group of 1,000 college students the mean height was 68.2 inches and the standard deviation was 2.5 inches. (a) Find the probability that in a sample of 100, the mean height will be between 67.82 and 68.78 inches. (b) Find the probability that in a sample of 100, the mean will be greater than 68.9 inches.

6. Consider the table of the weights of men found on page 140. Does the difference between each sample mean and the universe mean lie within the 5 per cent level of significance?

7. Consider the table of the heights of men found on page 141. Does

[1] Herbert Sorenson: *Statistics for Students of Psychology and Education*, pp. 295–300.

the difference between each sample mean and the universe mean lie within the 5 per cent level of significance?

8. Using the probable error notation, E. S. Pearson gave the mean length of cubit for 1,063 British males as 18.31 ± 0.019 inches. Show that $\sigma = 0.92$ inch. Does this mean that, assuming a normal distribution, about 709 of these men had cubit lengths between 17.39 and 19.23 inches?

9. If the statement is made that the mean height of 1,000 men is 68.78 ± 0.046 inches, can you adduce evidence that 0.046 is E_M and not σ_M?

10. (*Freeman*) Two engineers made 1,306 readings during a 5-year period on the heat value in Btu. of a mixed gas. The distribution, approximately normal, gave $M_u = 534.99$ Btu. and $\sigma_u = 3.85$ Btu. On 64 days at irregular intervals, state inspection was conducted and the mean of the approximately normal sample was 536.72 Btu. Would you say that the 64 measures constituted a random sample?

11. The breaking strength of a certain type of cord has been established from considerable experience to be 18.3 ounces with a standard deviation of 1.2 ounces. A sample of 100 pieces of the same type of cord shows a mean breaking strength of 16.5 ounces. Would you say that the sample is inferior?

12. After observing a large number of cases it has been established that a certain disease is 10 per cent fatal. The hospital of the Good Shepherd found that during the period 1937–1942, of 100 patients admitted with the disease 12 died. May this difference be attributed to chance? Hint: $\sigma_q = \sqrt{pq/N}$.

13. At Bucknell University the freshmen who take College Algebra are previously screened by a placement test. Our records covering a period of years reveal that about 16 per cent fail the course. During the fall semester, 1942, of 400 freshmen enrolled in College Algebra 20 per cent failed. Adopting the 5 per cent level as a basis for judgment, would you say this difference is significant?

D. The Skewness and Excess of the Distribution of Means. We have derived formulas for the arithmetic mean and the standard deviation of the distribution of sample means given by (1). In order to characterize more completely the distribution, we should derive formulas for the skewness and the excess. In this section we shall give an abridged derivation for the skewness, leaving the details for the reader to work out, and shall give without proof the formulas for the excess. (See Exercise 9 at end of this chapter.)

The skewness for the distribution of means is given by:

$$\alpha_{3,z} = \frac{\nu_{3,z}}{\sigma_z^3}$$

where

$$\nu_{3,Z} = \frac{\Sigma Z^3}{{}_SC_N} - \frac{3\Sigma Z^2}{{}_SC_N} \cdot M_Z + 2M_Z^3 \tag{10}$$

Returning to equations (1) we have:

$$\frac{\Sigma Z^3}{{}_SC_N} = \frac{1}{N^2S}\left[\Sigma X_i^3 + \frac{3(N-1)}{(S-1)}\Sigma X_i^2X_j + \frac{6(N-1)(N-2)}{(S-1)(S-2)}\Sigma X_iX_jX_k\right] \tag{11}$$

Since

$$(\Sigma X)^2 = \Sigma X^2 + 2\Sigma X_iX_j$$

and

$$\Sigma X \Sigma X^2 = \Sigma X^3 + \Sigma X_i^2X_j$$

and

$$2\Sigma X \Sigma X_iX_j = 2\Sigma X_i^2X_j + 6\Sigma X_iX_jX_k$$

we have:

$$\Sigma X_i^2X_j = \Sigma X^2 \Sigma X - \Sigma X^3$$

$$6\Sigma X_iX_jX_k = (\Sigma X)^3 - 3\Sigma X^2\Sigma X + 2\Sigma X^3$$

Substituting these values in (11) we obtain:

$$\frac{\Sigma Z^3}{{}_SC_N} = \frac{1}{N^2}\left[\frac{(S-N)(S-2N)}{(S-1)(S-2)}\frac{\Sigma X^3}{S} + \frac{3S(S-N)(N-1)}{(S-1)(S-2)}\frac{\Sigma X^2}{S} \cdot M_X + \frac{S^2(N-1)(N-2)}{(S-1)(S-2)} \cdot M_X^3\right]$$

Substituting this and the other necessary values, previously found, into (10) we have:

$$\nu_{3,Z} = \frac{(S-N)(S-2N)}{N^2(S-1)(S-2)}\left[\frac{\Sigma X^3}{S} - \frac{3\Sigma X^2}{S} \cdot M_X + 2M_X^3\right]$$

$$\nu_{3,Z} = \frac{(S-N)(S-2N)}{N^2(S-1)(S-2)}\nu_{3,X}$$

and hence

$$\alpha_{3,Z} = \frac{S-2N}{S-2}\sqrt{\frac{S-1}{N(S-N)}} \cdot \alpha_{3,X} \tag{12}$$

If S is infinite:

$$\alpha_{3,Z} = \frac{1}{\sqrt{N}} \cdot \alpha_{3,X} \tag{13}$$

Further, if the parent population is normal,[1] $\alpha_{3,X} = 0$; hence $\alpha_{3,Z} = 0$. Therefore the *skewness of the distribution of sample means chosen from a parent normal distribution is zero.*

[1] See p. 405.

By a similar procedure, but with much more laborious algebra, it may be shown that for the distribution of sample means given by (1) the excess is given by the formula:

$$\alpha_{4,Z} - 3 = \frac{(S-1)(S^2+S-6NS+6N^2)}{N(S-N)(S-2)(S-3)}[\alpha_{4,X} - 3] - \frac{6S(N-1)(S-N-1)}{N(S-N)(S-2)(S-3)}$$

If S is infinite:

$$\alpha_{4,Z} - 3 = \frac{1}{N}[\alpha_{4,X} - 3]$$

If the parent population is normal, $\alpha_{4,X} = 3$, in which case $\alpha_{4,Z} = 3$. Therefore, we may say that the *excess of the distribution of sample means chosen from a normal parent population is zero.*

In the text we have stated that the distribution of means is a normal distribution. It has long been known, probably since the time of Gauss, that *if random samples are taken from a universe distributed normally, the means of the samples also form a normal distribution.* If the universe is non-normal, not a great deal is known at present, from analytic considerations, about the distributions of statistics of samples. However, even for small values of N, there is sufficient experimental evidence to support the conclusion that the *distribution of means of samples selected randomly from any finite universe* is practically normal.

We have shown that if S is unlimited and N is large, $\alpha_{2,M} = 1$, $\alpha_{3,M} = 0$, $\alpha_{4,M} = 3$. By a continuation of this same method,[1] under the stated hypotheses, it is easy to show that $\alpha_{5,M} = 0$, $\alpha_{6,M} = 1 \cdot 3 \cdot 5 = 15$, $\alpha_{7,M} = 0$, $\alpha_{8,M} = 1 \cdot 3 \cdot 5 \cdot 7 = 105$, and so on. That is to say, if fairly large samples are taken from an infinite universe, the moments of the distribution of means are those of a normal curve. Further, it is not difficult to show that if the parent universe is infinite and distributed according to the Pearson Type III curve, the moments of the distribution of sample means are also of the Pearson Type III curve. However, it is well known that as N increases the Type III curve approaches the normal, so again we have the property that as N grows large, the curve of means approaches normality.

[1] Richardson, C. H., *The Statistics of Sampling*, published by Edwards Brothers, Ann Arbor, Michigan.

We may say then that, so far as the practical needs are concerned, the distribution of means has been rather thoroughly explored. We regret that we cannot say so much for the distribution of sample standard deviations. If the universe is normal the curve of the standard deviations of samples is Type III. If the universe is non-normal, we do not know the distribution function of σ, *not even the values of the moments.* However, by working through the moments of the variance ($= \sigma^2$) we arrive at the facts contained in the next section.

112. THE RELIABILITY OF THE STANDARD DEVIATION

In Section 110 (p. 425) we outlined an experiment that was intended to explain to the reader what is meant by a sample mean and a sample standard deviation. Each sample drawn has its mean, its standard deviation, et cetera. In order to introduce the reader to the problem of sampling, we have shown in considerable detail in the preceding section how we may characterize the distribution of sample means. We were especially interested, however, in finding measures of the reliability of the mean, which measures we found in σ_M and $E_M (= .6745\sigma_M)$.

The sample standard deviations in like manner form a distribution that may be characterized by its mean, M_σ (the mean of the standard deviations), its standard deviation, σ_σ (the standard deviation of the standard deviations), and so on. We are especially interested in σ_σ or E_σ, by which we measure the variability and the reliability of any sample standard deviation.

The algebraic development showing the derivation of M_σ and σ_σ would take us too far afield. It can be shown that *if the parent population is normal and N is large, the mean of the distribution of standard deviations is approximately equal to the standard deviation of the parent population, and the standard deviation of the distribution of standard deviations is approximately equal to the standard deviation of the parent population divided by the square root of twice the number of variates in the sample.*[1] That is:

$$M_\sigma = \sigma_u$$

$$\sigma_\sigma = \frac{\sigma_u}{\sqrt{2N}}$$

[1] See formula (24) of Section 114.

Since the standard deviation of the parent population, σ_u, is approximately equal to the standard deviation of the sample, σ, we have:

$$M_\sigma = \sigma \text{ approximately}$$

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \text{ approximately} \tag{14}$$

and

$$E_\sigma = 0.6745\sigma_\sigma = 0.6745\frac{\sigma}{\sqrt{2N}} \tag{15}$$

The meaning of E_σ is similar to that given for E_M. Thus, it is customary to write the standard deviation in the form

$$\sigma_u = \sigma \pm E_\sigma$$

which means that, assuming that the curve of sample standard deviations is approximately normal, half the sample standard deviations lie within the range whose end values are $\sigma_u - E_\sigma$ and $\sigma_u + E_\sigma$. It also means that the chances are even that the sample σ does not differ from σ_u by more than $\pm E_\sigma$, and it is practically certain that the sample σ does not differ from σ_u by more than $\pm 4.5(E_\sigma)$.

113. THE RELIABILITY OF THE DIFFERENCE BETWEEN TWO MEASURES

An important problem in applied statistics is the determination of some criterion that will assist one in judging whether an observed difference between two samples is apparent or real. That is, is the difference between two samples such that it might arise from sampling (that is, from pure chance), or is the difference *significant* of a greater variation in the two samples than can be explained by random sampling alone?

Suppose we select from a normal parent population two samples, each fairly large. Each sample has its mean, its standard deviation, et cetera. The two means will not likely be equal and hence we shall have a difference of two means. Also, the standard deviations will not likely be equal and hence we shall have a difference of two standard deviations. Continue this process until we have, say, m pairs of samples, m usually a large number, and hence m differences in means

that will constitute a distribution of differences in sample means. From these m pairs of samples we may also have m differences in standard deviations that will constitute a distribution of differences of sample standard deviations.

Let X_i and Y_i be used to distinguish corresponding characteristics — means, standard deviations, et cetera — of two groups when the ith pair of samples has been taken, and $X_i - Y_i$ be the difference in any pair of corresponding characteristics.

TABLE 99

Sample Pair	*Group I* X	*Group II* Y	*Difference* $X - Y$
1	X_1	Y_1	$X_1 - Y_1$
2	X_2	Y_2	$X_2 - Y_2$
..	..	..	.....
i	X_i	Y_i	$X_i - Y_i$
..	..	..	.....
m	X_m	Y_m	$X_m - Y_m$

We shall find the arithmetic mean and the standard deviation for the distribution of differences:

$$M_{X-Y} = \frac{\Sigma(X_i - Y_i)}{m} = \frac{\Sigma X_i}{m} - \frac{\Sigma Y_i}{m} = M_X - M_Y \tag{16}$$

$$\sigma^2_{X-Y} = \frac{\Sigma(X - Y)^2}{m} - \left[\frac{\Sigma(X - Y)}{m}\right]^2 \qquad \text{(by (7), Chapter 4)}$$

$$= \frac{\Sigma X^2}{m} - \left(\frac{\Sigma X}{m}\right)^2 + \frac{\Sigma Y^2}{m} - \left(\frac{\Sigma Y}{m}\right)^2 - 2\left[\frac{\Sigma XY}{m} - \frac{\Sigma X}{m}\frac{\Sigma Y}{m}\right]$$

Using (7) on p. 128, and (7) on p. 245, we have:

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2r_{XY}\sigma_X\sigma_Y \tag{17}$$

If the two distributions, X and Y, are independent so that r_{XY} is zero, then:

$$\sigma_{X-Y} = \sqrt{\sigma^2_X + \sigma^2_Y} \tag{18}$$

We are especially interested in (18) when X and Y are the corresponding means of two samples. Then

$$\sigma_{M_X - M_Y} = \sqrt{\sigma_{M_X}^2 + \sigma_{M_Y}^2} \tag{19}$$

gives a measure of the variability (or the reliability) of the differences in two sample means. Also

$$\sigma_{\sigma_X - \sigma_Y} = \sqrt{\sigma_{\sigma_X}^2 + \sigma_{\sigma_Y}^2} \tag{20}$$

gives a measure of the variability and the reliability of the differences in two sample standard deviations. The formulas for the corresponding probable errors are found by multiplying (19) and (20) by 0.6745. Thus:

$$E_{M_X - M_Y} = 0.6745\sqrt{\sigma_{M_X}^2 + \sigma_{M_Y}^2} \tag{21}$$

$$E_{\sigma_X - \sigma_Y} = 0.6745\sqrt{\sigma_{\sigma_X}^2 + \sigma_{\sigma_Y}^2} \tag{22}$$

Let us consider, for illustration, the results on the placement examination in mathematics of two different freshman classes at Bucknell University.

Group I	*Group II*
$N_X = 329$	$N_Y = 302$
$M_X = 32.75$	$M_Y = 30.60$
$\sigma_X = 8.05$	$\sigma_Y = 6.95$

The difference between the two means is $32.75 - 30.60 = 2.15$. Is this difference so large that it could not be due to chance or does it indicate that Group I really demonstrated a significantly better training in elementary mathematics?

There is no question about the observed difference in the two means. It is certainly 2.15. Could such a difference be due to chance? Yes, such a difference *could* be due to chance but we shall show that the likelihood that it did arise from chance is so small that we feel justified in neglecting it and in assuming that the difference has been caused by other factors than pure chance. When such is the case, the statistician says "the difference is significant."

Using (6), (19), and (21):

$$\sigma_{M_X} = 0.444 \qquad \sigma_{M_X - M_Y} = 0.597$$
$$\sigma_{M_Y} = 0.399 \qquad E_{M_X - M_Y} = 0.403$$

We may now state that the chances are even that an observed $(M_X - M_Y)$ is within $\pm$ 0.403 of the true (unknown) value, and it is practically certain that an observed $(M_X - M_Y)$ is within $\pm$ 3(0.597) or $\pm$ 4.5(0.403) of the true value. Following custom, we describe this variation by writing 2.15 $\pm$ 0.403 which, translated into English, reads "2.15 with a probable error 0.403." It may be noticed incidentally that the difference 2.15 is 3.6 times its standard error and 5.3 times its probable error. Such a large numerical difference as this would rarely occur by pure chance, in fact, about 4 times in 10,000. When the happening of an event, such as this under discussion, is extremely unlikely, we conclude that some factors other than pure chance have influenced the result.

While proofs for all the statements are beyond the scope of this text, other pertinent facts are the following. If many independent sample pairs are taken from normal parent populations, the differences (indicated by D) of means, standard deviations, etc. also form approximately normal distributions. As may be expected, the mean of the distribution of differences, M_D, is zero and the standard deviation, σ_D, is given by (18). The probable error of the D distribution is of course $E_D = 0.6745\sigma_D$. It is customary to take

$$\frac{|D|}{\sigma_D} = t \quad \text{or} \quad \frac{|D|}{E_D} = k$$

as the criteria whereby one can quickly determine if the difference D is significant. As a "rule of thumb" we say:

if $t > 3$, (or if $k > 4.5$), the difference is certainly significant;
if $t > 2$, (or if $k > 3$), the difference is possibly significant;
if $t < 2$, (or if $k < 3$), the difference is probably not significant.

These limits, however, are arbitrary, and consequently vary among the authorities.

In the particular problem of this section:

$$t = \frac{D}{\sigma_D} = \frac{M_X - M_Y}{\sigma_{M_X - M_Y}} = \frac{2.15}{0.597} = 3.6$$

Hence, from a comparison of the means, we would conclude that Group I and Group II came from statistically different parent populations; or, if from the same parent population, then other factors than pure chance must have caused a numerical difference as large as 2.15.

The assumptions underlying this procedure deserve a brief consideration. The universe difference in the means, or other statistics, is *assumed to be zero.* Is this a reasonable assumption? We think it is. Let the reader return to Table 99, and remember that each sample, fairly large, is drawn from a normal parent universe. It would seem then that of the m differences of $(X_i - Y_i)$, negative differences would occur about as frequently as positive differences and of equal numerical amounts so that their sum $\Sigma(X_i - Y_i)$ would theoretically equal zero. Hence, theoretically $M_X = M_Y$.

R. A. Fisher terms such an hypothesis a "*null hypothesis,*" the hypothesis that there is *no* difference. So in our applications we try to give the facts a chance to nullify the hypothesis. We make no effort to prove it or to disprove it; rather do we attempt to cast doubt upon it.

In our illustrative example we sought evidence that the two samples came from different universes. Very well, on the basis of large sample theory, we began by assuming they came from the *same universe* with $M_D = 0$ and $\sigma_D = .597$. It is expected that practically all of the actual differences will fall within $0 \pm 3\sigma_D$. If, therefore, the actual difference D exceeds $3\sigma_D$ numerically, then it is reasonable to conclude that our assumption of the same universe is probably wrong. Thus we conclude that the two samples came from *different universes.*

Of course we may wish to see what light a comparison of the variabilities of the samples will throw upon our problem. We find, using (14) and (20),

$$\sigma_{\sigma_X} = \frac{8.05}{\sqrt{2(329)}} = 0.314 \qquad \sigma_{\sigma_Y} = \frac{6.95}{\sqrt{2(302)}} = 0.283$$

$$\sigma_{\sigma_X - \sigma_Y} = \sqrt{(0.314)^2 + (0.283)^2} = 0.423$$

$$t = \frac{D}{\sigma_D} = \frac{8.05 - 6.95}{0.423} = 2.6$$

So a comparison of the standard deviations supports the previous conclusion since $t > 2$.

As this problem well illustrates, in investigating the significance in differences it is a wise procedure to penetrate the problem as deeply as possible.

Example. Two samples of weights of male students gave the following information: $N_1 = 100$, $M_1 = 140.4$ lbs., $\sigma_1 = 17.7$ lbs.; $N_2 = 100$, $M_2 = 136.8$ lbs., $\sigma_2 = 16.2$ lbs. If other samples are taken, what is the probability that an observed difference in the means will be numerically equal to or greater than $D = 140.4 - 136.8 = 3.6$ lbs.?

Solution.

$$\sigma_{M_1} = \frac{17.7}{\sqrt{100}} = 1.77 \qquad \sigma_{M_2} = \frac{16.2}{\sqrt{100}} = 1.62$$

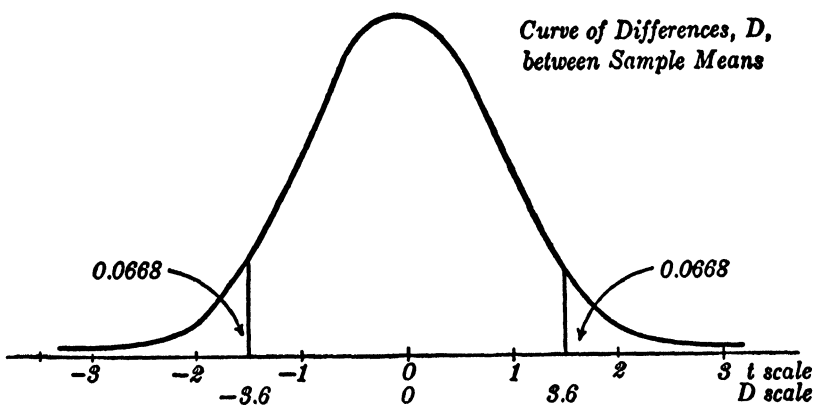
$$\sigma_D = \sqrt{(1.77)^2 + (1.62)^2} = 2.4$$

$$t = \frac{D}{\sigma_D} = \frac{3.6}{2.4} = 1.5$$

$$P = 2\left[A_\phi\right]_{1.5}^{\infty} = 2(0.5000 - 0.4332) = 0.1336$$

That is, we would obtain a difference numerically as large as 3.6 about 134 times in 1,000.

FIGURE 61



EXERCISES

1. The following table gives the distribution of the weights of 1,000 female students subdivided into ten random samples, each of 100 individuals. Measurements were recorded to nearest 1/10th pound.

Class	*Frequencies*										
	1st 100	*2nd 100*	*3rd 100*	*4th 100*	*5th 100*	*6th 100*	*7th 100*	*8th 100*	*9th 100*	*10th 100*	*Total*
74.95		1							1		2
84.95		1	4	1	1	4		3	2		16
94.95	9	7	9	4	2	10	14	9	9	9	82
104.95	22	18	23	19	29	23	26	30	20	21	231
114.95	30	24	23	31	28	17	21	31	22	21	248
124.95	25	21	19	22	18	21	15	15	16	24	196
134.95	9	16	15	9	12	12	10	7	19	13	122
144.95	2	7	5	6	6	8	10	5	6	8	63
154.95	2	3	1	2	2	4	2		4	3	23
164.95	1	0	0	0	1	1	1		1	0	5
174.95		2	0	3	1		0			1	7
184.95			0	0			1				1
194.95			0	2							2
204.95			1	0							1
214.95				1							1
Total	100	100	100	100	100	100	100	100	100	100	1000
M											
σ											

Let the universe be the "total" group.

a. Compute M and σ for each sample and for the total.

b. Compare the mean of the ten sample means with M_u.

c. Compare the mean of the ten sample σ's with σ_u.

d. Using $\sigma_M = \frac{\sigma_u}{\sqrt{100}}$, how many of the ten sample means are within the five per cent level of significance?

e. Using $\sigma_\sigma = \frac{\sigma_u}{\sqrt{200}}$, how many of the ten sample σ's are within the five per cent level of significance?

f. Do you believe that randomness went awry on any sample?

2. (*Tippett*, p. 70) The lengths of 4,000 hairs of an Indian cotton gave $M = 2.33$ cm. and $\sigma = 0.4806$ cm. "The first thousand hairs were selected

by a different method from the rest and gave a mean of 2.54 cm. Is this deviation compatible with the hypothesis that the 1,000 are a random sample from the 4,000 and that the difference in means is due to random errors, or is the difference large enough to indicate that the change in technique has had an effect?"

3. A contractor purchased a certain type of copper sheeting from a manufacturer. The contract specified that the sheets were to meet a theoretical standard — universe mean — of thickness 0.022 inch. The contractor measured a sample of 100 sheets and found $M = 0.020$ inch and $\sigma = 0.003$ inch. Did the contractor have reason to complain?

4. For ten years we at Bucknell University have given to the in-coming freshmen a standardized test in pre-college mathematics. Based upon this experience with $S = 4{,}000$ we have established the norms for the test: $M_u = 62$, $\sigma_u = 18$. The freshman class of 400 admitted in September 1939, Class of 1943, took the test with the results: $M = 58$ and $\sigma = 16$. Would you agree that the Class of 1943 was significantly ill-prepared in mathematics? The Class of 1945 with $N = 400$ took the test with the results: $M = 60.5$ and $\sigma = 16.5$. Is the Class of 1945 within the five per cent level?

5. During a given month one machine produced 900 units but spoiled 3.2 per cent of them. During the same month another machine with a more experienced operator produced 1,000 units but spoiled 2.8 per cent of them. Is the percentage difference in spoilage significant?

6. A. S. Parkes and J. C. Drummond (*Proc. Roy. Soc.*, B, XCVIII, p. 147) gave the following data showing the effect of vitamin B on the sex-ratio of offspring in rats. May the percentage change in males be attributed to chance, or is the evidence sufficient to warrant that the change was due to the increased vitamin B?

Diet	*Males*	*Females*	*Total Young*	*Per cent Males*
Vitamin B Deficient	123	153	276	44.57
Vitamin B Sufficient	145	150	295	49.15
Totals	268	303	571	

114. SMALL SAMPLES

The formulas for estimating the reliability of a statistic that we have given previously are suitable when N is reasonably large, say 30 or more, but require modification when N is small. When N is small, the σ of a sample which is used as an estimate of σ_u gives values too small and thus our standard errors have a downward bias. To overcome this bias we need to develop a theory that will give

us a better estimate of the standard deviation of the universe σ_u than is given by σ. We shall now attack the problem of finding the standard deviation which gives the better estimate for σ_u.

Consider the parent universe $X_1, X_2, \ldots, X_S$. Transform the S variates to the mean of the universe M_u as origin and denote them by $x_1, x_2, \ldots, x_S$ where $x_i = X_i - M_u$. We then have *for the universe* $\sum_{i=1}^{S} x_i = 0$.

From this universe we choose samples of N. In all we may choose ${}_SC_N$ samples. Each sample has its second moment and thus in all we have ${}_SC_N$ second moments. These ${}_SC_N$ second moments give us a distribution of sample second moments. It is our immediate problem to find the mean of these ${}_SC_N$ sample second moments.

Let

$m_{2,k}$ = the second moment of the kth sample about the mean of the sample

Then, for the kth sample, we have

$$m_{2,k} = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2 = \frac{\Sigma x^2}{N} - \frac{(\Sigma x)^2}{N^2}$$

Since $(\Sigma x)^2 = \Sigma x^2 + 2\Sigma x_i x_j$, we have

$$m_{2,k} = \frac{1}{N^2}[(N-1)\Sigma x^2 - 2\Sigma x_i x_j]$$

where $i \neq j$, and the Σ's cover only the sample.

The mean of the distribution of second moments is given by

$$M_{\mu_2} = \frac{1}{{}_SC_N}\sum_{k=1}^{{}_SC_N} m_{2,k} = \frac{1}{N^2}\left[(N-1)\frac{N}{S}\Sigma x^2 - \frac{2N(N-1)}{S(S-1)}\Sigma x_i x_j\right]$$

where the Σ's cover the entire universe.

Again returning to

$$(\Sigma x)^2 = \Sigma x^2 + 2\Sigma x_i x_j,$$

we note that for the universe $\Sigma x = 0$, and hence

$$-2\Sigma x_i x_j = \Sigma x^2 = S\sigma_u^2$$

Then, substituting and simplifying,

$$M_{\mu_2} = M_{\sigma^2} = \frac{S(N-1)}{N(S-1)}\sigma_u^2 \tag{23}$$

If S becomes infinite,

$$M_{\sigma^2} = \frac{N-1}{N}\sigma_u^2 \tag{24}$$

That is, if the parent population is very large, the expected μ_2 or σ^2 is $(N-1)/N$ times the parent $\mu_{2,u}$ or σ_u^2.

If we replace in (24) M_{σ^2} by $\sigma^2{}_{sample}$ or σ^2, and σ_u by $\sigma_{u,\,estimated}$ or $\sigma_{u,\,est.}$, where $\sigma_{u,\,est.}$ is *the best estimate of the standard deviation of the universe from the sample,* (what R. A. Fisher calls the *maximum likelihood estimation* of σ_u from a sample) we have

$$\sigma^2 = \frac{N-1}{N}\sigma_{u,\,est.}^2$$

$$\sigma_{u,\,est.} = \sqrt{\frac{N}{N-1}}\,\sigma \tag{25}$$

If, as is customary, we find σ for a sample of N items by the formula

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \tag{26}$$

we obtain

$$\sigma_{u,\,est.} = \sqrt{\frac{N}{N-1}} \cdot \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N-1}} \tag{27}$$

Consequently, if we must *estimate* σ_u from a sample, formula (27) gives a better estimate than the customary one (26). Of course if N is large, it is a matter of little consequence whether we divide by N or by $(N-1)$, but when N is small, say less than 30, the use of $(N-1)$ is particularly important.

We immediately find, for N small,[1]

[1] The introduction of the factor $\sqrt{\frac{N}{N-1}}$ in (25) is called "Bessel's correction," and the formula for the standard error of the mean $\sqrt{\frac{\Sigma x^2}{N(N-1)}}$ is called "Bessel's formula" [Friedrich Wilhelm Bessel (1784–1846)].

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}} \tag{28}$$

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2(N-1)}} \tag{29}$$

where σ is computed from (26). Corresponding formulas for E_M and E_σ are immediately found.

We may thus compute standard and probable errors for the various statistics, mean, standard deviation, differences, and so on when N is small by using formulas (27), (28), (29), and their substitutions in (21) and (22). A word of caution is in order with regard to applying them to establish probability levels.

In our previous discussion we have used the values of the normal curve to assist in interpreting the values of

$$\frac{M - M_u}{\sigma_M}, \quad \frac{\sigma - \sigma_u}{\sigma_\sigma}, \quad \frac{M_1 - M_2}{\sigma_{M_1 - M_2}}$$

because the distributions of these quantities are closely normal *when* N *is large.* When N is small, these distributions deviate from normality, the amount of the deviation increasing as N decreases. A special table has been devised by R. A. Fisher which gives values of t for various "*degrees of freedom*" n, ($n = N - 1$ in the above formulas) and various probabilities P that an observed value may differ from zero by more than $\pm t$. Or it gives values of t for given levels of significance and given values of n.

This table differs considerably from that of the normal curve. For example, in the normal curve with $N = 11$ or $n = 10$, the 1 per cent level of significance is at $t = \pm 2.58$ whereas in the Fisher table the value of t is $t = \pm 3.17$. When N is larger than 20, the differences are not so appreciable, and when N is greater than 30 the normal table may be used with slight error. This Fisher table is found in the texts by Fisher and by Croxton and Cowden listed in the Appendix. A general idea of the table may be obtained from the portion that we reproduce on page 454.

In the use of this table remember that a "level of significance" refers to both tails of the distribution. Note too that it is set up differently from the table of areas for the normal curve. A tail of the normal curve is found by subtracting the tabulated value from 0.5000, and doubling this value yields the level of significance.

TABLE 100. VALUES OF t FOR DEGREES OF FREEDOM n AND LEVELS OF SIGNIFICANCE P

	Level of Significance						
n \ P	.9	.7	.5	.3	.1	.05	.01
4	.134	.414	.741	1.190	2.132	2.776	4.604
5	.132	.408	.727	1.156	2.015	2.571	4.032
6	.131	.404	.718	1.134	1.943	2.447	3.707
8	.130	.399	.711	1.108	1.860	2.306	3.355
10	.129	.397	.706	1.093	1.812	2.228	3.169
15	.128	.393	.691	1.074	1.753	2.131	2.947
20	.127	.391	.687	1.064	1.725	2.086	2.845
30	.127	.389	.683	1.055	1.697	2.042	2.750
∞	.1257	.3853	.6745	1.0364	1.6449	1.9600	2.5758

Table 100, however, shows n (degrees of freedom) in the stub, P (the level of significance) in the caption, and t in the body of the table. The last line of the table for $n = \infty$ shows values of t obtained from the normal curve.

Exercise: Show that $\sigma_{u,\,est.}$ may be found from

$$\sigma_{u,\,est.} = \sqrt{\frac{N\Sigma X^2 - (\Sigma X)^2}{N(N-1)}} \tag{30}$$

Illustrative Example 1. A corporation has set as a standard the mean breaking strength of a certain type of wire at 582 pounds. A sample of 10 specimens was tested with the results shown in

TABLE 101. BREAKING STRENGTH OF WIRE

Specimen	*Breaking Strength (pounds)* X	x	x^2
1	581	2	4
2	576	− 3	9
3	584	5	25
4	586	7	49
5	575	− 4	16
6	573	− 6	36
7	574	− 5	25
8	572	− 7	49
9	588	9	81
10	581	2	4
Total	5790	0	298

Table 101. For the purposes for which the wire is used, values within the 5 per cent level are tolerated. Does this sample meet the requirements?

Solution:

We have
$$M = \frac{5790}{10} = 579 \text{ pounds.}$$
$$\sigma = \sqrt{\frac{298}{10}} = 5.46 \text{ pounds.}$$
$$\sigma_M = \frac{5.46}{\sqrt{9}} = 1.82 \text{ pounds.}$$

In the t table for $n = N - 1 = 9$, we have at the 5 per cent level, $t = \pm 2.3$. That is, a variation of $\pm 2.3(1.82)$ or ± 4.14 pounds on either side of 582 pounds is tolerated. Hence the toleration limits are (582 ± 4.14) pounds or from 577.76 pounds to 586.76 pounds. Certainly 579 pounds, the mean of the sample, is well within these limits.

Illustrative Example 2. Table 102 gives data on strength tests (lbs. per sq. in.) on two types of wool fabric. Is the difference in the means sufficient to warrant the conclusion that Type 2 is superior to Type 1?

TABLE 102

Type 1		*Type 2*	
Specimen	*Strength*	*Specimen*	*Strength*
1	139	1	137
2	127	2	132
3	134	3	135
4	125	4	144
5	141	5	131
6	144	6	133
7	128	7	136
8	138	8	134
9	131	9	139
10	133	10	129

For these data we find

$N_1 = 10$ $\qquad$ $N_2 = 10$

$M_1 = 134$ lbs. per sq. in. $\qquad$ $M_2 = 135$ lbs. per sq. in.

$\sigma_1 = 6.05$ lbs. per sq. in. $\qquad$ $\sigma_2 = 4.09$ lbs. per sq. in.

$$\sigma_{M_1} = \frac{\sigma_1}{\sqrt{N_1 - 1}} = 2.02 \text{ lbs. per sq. in.}$$

$$\sigma_{M_2} = \frac{\sigma_2}{\sqrt{N_2 - 1}} = 1.36 \text{ lbs. per sq. in.}$$

$$\sigma_{M_1 - M_2} = \sqrt{(2.02)^2 + (1.36)^2} = 2.4 \text{ lbs. per sq. in.}$$

$$t = \frac{D}{\sigma_D} = \frac{135 - 134}{2.4} = .417$$

From Table 100, for $n = N - 1 = 9$ and $t = .417$ we estimate P at about 0.7, indicating that a difference of $\pm$ 1 lb. per sq. in. might occur 7 times in 10. There is thus no evidence to support a contention that Type 2 is superior to Type 1.

115. CONCLUDING REMARKS ON SAMPLING

The statistical theory of sampling is a fundamental and basic problem in mathematical statistics. It has challenged and continues to challenge some of our best minds. The reader who may wish to pursue the problem further will find the following articles interesting and not too difficult.

H. C. Carver, *Fundamentals of the Theory of Sampling*, Annals of Math. Statistics, Vol. I, page 101.

C. C. Craig, *An Application of Thiele's Semi-invariants to the Sampling Problem*, Metron, Vol. VII, No. 4.

W. E. Deming and R. T. Birge, *Statistical Theory of Errors*, The Graduate School of U.S. Dept. of Agriculture, Wash., D.C.

Dunham Jackson, *The Theory of Small Samples*, Amer. Math. Monthly, June–July, 1935.

C. H. Richardson, *The Statistics of Sampling*, Edwards Brothers, Ann Arbor, Michigan.

H. L. Rietz, *Topics in Sampling Theory*, Bulletin of the American Mathematical Society, April, 1937.

W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Co., New York City.

116. SUMMARY OF RELIABILITY FORMULAS

In this chapter we have undertaken only to *introduce* the reader to what Karl Pearson has called the fundamental problem in statistics,

namely, the problem of sampling. To do more in an elementary text would not be good judgment on our part. A list of the probable errors that are needed most frequently follows and includes a few which we are not in a position to derive here.[1]

Statistical Constant	*Probable Error*
The arithmetic mean	$\frac{0.6745\sigma}{\sqrt{N}}$
The median (normal distribution)	$\frac{0.8454\sigma}{\sqrt{N}}$
The standard deviation (normal distribution)	$\frac{0.6745\sigma}{\sqrt{2N}} = \frac{0.4769\sigma}{\sqrt{N}}$
The coefficient of correlation (normal distribution)	$0.6745\frac{1-r^2}{\sqrt{N}}$
α_3 for a normal distribution	$0.6745\sqrt{\frac{6}{N}}$
α_4 for a normal distribution	$0.6745\sqrt{\frac{24}{N}}$

EXERCISES

1. For the distribution of scores in English, (a) of Exercise 4, page 102, we have found $N = 334$, $M = 149.8$, $\sigma = 42.47$. Find E_M and interpret it. Also find σ_σ and interpret it.

2. For the distribution of the lengths of eggs, (a) of Exercise 15, page 105, we have found $N = 450$, $M = 56.323$ mm., $\sigma = 2.386$ mm. What is the probability that the sample mean does not differ from the universe mean by more than ± 0.09 mm.? What is the probability that the sample dispersion does not differ from the true dispersion of the universe by more than ± 0.07 mm.?

3. Find σ_M and σ_σ for the distribution of pulse beats, Table 29, page 165. Find the probability that the sample mean does not differ from the universe mean by more than ± 1.0 pulse beats per minute.

4. Assuming normality, find σ_r and E_r for the data of Table 59, and interpret them.

5. Find E_M for the data of the chest measurements of men, Exercise 10, page 168, and interpret it.

[1] For the probable errors of other constants, see Rietz and others, *op. cit.* p. 77.

6. The following are summaries of the results on placement tests in English which were given to two freshman classes entering Bucknell University.

Group I	*Group II*
$N = 334$	$N = 302$
$M = 149.79$	$M = 158.37$
$\sigma = 42.47$	$\sigma = 36.28$

Is the difference between the means significant?

7. The heights of two groups of soldiers were measured and the following results were secured:

Group I	*Group II*
$N = 10{,}000$	$N = 10{,}000$
$M = 67.51$ inches	$M = 62.24$ inches
$\sigma = 2.20$ inches	$\sigma = 2.25$ inches

Is the difference in the means sufficient to warrant belief that the two groups were chosen from different races?

8. We present below two frequency distributions based upon the batting averages of players in the National and the American leagues during the early part of the 1925 season. (See the accompanying table.)

FREQUENCY DISTRIBUTION OF BATTING AVERAGES [1]

Batting Average	*Number of Players in the National League with the Given Average*	*Number of Players in the American League with the Given Average*
.050–.099	3	0
.100–.149	7	11
.150–.199	11	11
.200–.249	21	22
.250–.299	31	35
.300–.349	34	28
.350–.399	18	13
.400–.449	4	6
.450–.499	0	0
.500–.549	3	2
.550–.599	0	1

Is the difference in the means of these distributions significant?

[1] New York *Herald Tribune*, May 17, 1925. See also F. C. Mills and D. H Davenport, *Manual of Problems and Tables of Statistics*, 1925, p. 65.

9. In the theory of the chapter we have assumed that the parent population consisted of the S variates $X_1, X_2, \ldots, X_S$. We proved that $M_Z = M_X$, where $M_{\bar{z}}$ is the mean of the distribution of sample means and M_X is the mean of the parent population. Let us now transform the S variates to this mean as origin and denote them by $x_i = X_i - M_X$, $(i = 1, 2, \ldots, S)$.

Let z_i be the ith sample mean of N variates chosen from the population $x_1, x_2, \ldots, x_S$. We may have the ${}_SC_N$ distinct sample means which are given by the following equations:

$$z_1 = \frac{1}{N}\left[x_1 + x_2 + \cdots + x_{N-1} + x_N\right]$$

$$z_2 = \frac{1}{N}\left[x_1 + x_2 + \cdots + x_{N-1} + x_{N+1}\right]$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$z_{{}_SC_N} = \frac{1}{N}\left[x_{S-N+1} + x_{S-N+2} + \cdots + x_{S-1} + x_S\right]$$

Recalling that $\sum_{i=1}^{S} x_i = 0$:

a. Show that:

$$\Sigma z_i = 0$$

b. Show that:

$$\frac{\Sigma z^2}{{}_SC_N} = \frac{1}{N^2}\left[\frac{N}{S}\Sigma x_i^2 + \frac{2N(N-1)}{S(S-1)}\Sigma x_i x_j\right]$$

which, upon applying the proper symmetric products, reduces to:

$$\frac{\Sigma z^2}{{}_SC_N} = \frac{S-N}{N(S-1)}\frac{\Sigma x_i^2}{S} = \frac{S-N}{N(S-1)}\sigma_X^2$$

c. Use a. and b. and show that:

$$\sigma_z = \sqrt{\frac{\Sigma z^2}{{}_SC_N}} = \sqrt{\frac{S-N}{N(S-1)}}\sigma_X$$

d. Show that:

$$\frac{\Sigma z^3}{{}_SC_N} = \frac{1}{N^3}\left[\frac{N}{S}\Sigma x_i^3 + \frac{3N(N-1)}{S(S-1)}\Sigma x_i^2 x_j + \frac{6N(N-1)(N-2)}{S(S-1)(S-2)}\Sigma x_i x_j x_k\right],$$

which, upon applying the proper symmetric products, reduces to

$$\frac{\Sigma z^3}{{}_SC_N} = \frac{(S-N)(S-2N)}{N^2(S-1)(S-2)}\frac{\Sigma x_i^3}{S}$$

and finally to:

$$\nu_{s,z} = \frac{(S-N)(S-2N)}{N^2(S-1)(S-2)}\nu_{s,x}$$

10. Distributions of the heights of men born in England and in Scotland gave the following results:

England	*Scotland*
N = 6,194	N = 1,304
M = 67.4375 inches	M = 68.5456 inches
σ = 2.548 inches	σ = 2.480 inches

Is the difference in the means sufficient to conclude that Scots are really taller than Englishmen?

11. A distribution of 150 people in normal condition gave an average pulse rate of 79.68 $\pm$ 0.15 beats per minute but after being administered a certain drug they showed an average pulse rate of 81.12 $\pm$ 0.20 beats per minute. Is it probable that the increase in the pulse rate was due to the drug, or is the increase simply a result of variation due to sampling?

12. For the distributions of wages received by clothing workers in Cincinnati, Cleveland, and St. Louis we have found the values given in the table. Are the differences of the means significant? [See page 75.]

	Cincinnati	*Cleveland*	*St. Louis*
M	\$16.77	\$21.48	\$15.90
σ	6.86	6.28	6.04

13. The average grades of sorority and non-sorority women on a certain campus were as follows:

Sorority Group	*Non-sorority Group*
N = 175	N = 150
M = 81.23	M = 79.62
σ = 10.18	σ = 9.37

Is the difference of the arithmetic means sufficient to conclude that there was a *real* difference in the scholarship of the two groups?

14. Desiring to test the milk-producing qualities of two different kinds of food, a dairy association separated, by a random selection, 800 cows into two different herds. All other conditions were kept identical as far as possible. Observing the cows for a certain period, the following results were obtained:

Herd Number 1	*Herd Number 2*
$N_1 = 400$	$N_2 = 400$
$M_1 = 36$ quarts per cow	$M_2 = 40$ quarts per cow
$\sigma_1 = 5.4$ quarts per cow	$\sigma_2 = 4.5$ quarts per cow

Determine whether the difference between the average yields of the two herds is or is not significant.

15. The following table exhibits two frequency distributions relating to the earnings of coal miners in two different sections of Illinois. Is the difference between their means sufficient to conclude that these two sections do not belong to the same homogeneous group?

PICK MINERS IN ILLINOIS COAL MINES CLASSIFIED ACCORDING TO AVERAGE DAILY EARNINGS, 1918–1921 [1]

Range of Average Daily Earnings	*Number of Pay Checks*	
	In 21 Central Illinois Mines	*In 52 Southern Illinois Mines*
$ 2.00– 2.99	501	87
3.00– 3.99	1,288	131
4.00– 4.99	3,222	306
5.00– 5.99	6,293	563
6.00– 6.99	9,821	973
7.00– 7.99	13,089	1,530
8.00– 8.99	11,869	2,684
9.00– 9.99	9,484	5,584
10.00–10.99	6,748	2,426
11.00–11.99	4,418	1,433
12.00–12.99	2,551	853
13.00–13.99	1,304	577
14.00–14.99	696	364
15.00–15.99	362	197
16.00–16.99	196	105
17.00–17.99	115	71
18.00–18.99	57	35
19.00–19.99	39	33
20.00–20.99	25	13
21.00–21.99	16	6
22.00–22.99	13	7
23.00–23.99	10	4
24.00–24.99	10	4
Total	72,127	17,986

[1] See Mills and Davenport, *op. cit.*, p. 107.

16. Two types of electric bulbs were observed as to the length of life, and the following data were secured:

Type 1	*Type 2*
$N_1 = 100$	$N_2 = 100$
$M_1 = 1224$ hours	$M_2 = 1036$ hours
$\sigma_1 = 36$ hours	$\sigma_2 = 40$ hours

Is the difference in the two means sufficient to warrant the conclusion that Type 1 is a bulb superior to Type 2?

17. A large number of men were measured as to height giving $M_u = 68.1$ inches and $\sigma_u = 2.5$ inches. How large a sample should be taken in order to be fairly sure (probability 0.95) that the sample mean may not differ from the true mean by more than $\pm$ 0.5 inch?

18. The weights of 400 male babies of same nationality were analyzed. The analysis yielded $M = 7.29$ pounds and $\sigma = 1.01$ pounds. What statements can you make about the universe mean weight of babies of this nationality? If the universe mean were known to be 7.5 pounds, would you consider the above described sample a random one?

19. (*Treloar*, p. 143) "Data secured from the archives of the Sloane Hospital, New York City, for length of new-born infants of Irish parents yielded the following statistics":

Male (X)	*Female* (Y)
$N = 1{,}136$	$N = 1{,}071$
$M = 51.96$ cm.	$M = 51.22$ cm.
$\sigma = 2.181$ cm.	$\sigma = 2.189$ cm.

Do these results justify the inference that Irish male offspring are in general longer than females at birth? Do the results justify the inference that male babies are generally less variable in length than females at birth?

20. The cost of building an identical house in various parts of the United States in 1940 gave

$M =$ \$6,029 $\quad \sigma =$ \$459 $\quad N =$ number of cities $= 77$.

The cost of building the same house during the first quarter of 1941 gave

$M =$ \$6,232 $\quad \sigma =$ \$504 $\quad N =$ number of cities $= 68$.

Is this increase in average-cost significant?

21. The British Cotton Industry Research Association tested the breaking load on two types of yarn with the following results:

Type I	*Type II*
$N = 1{,}782$	$N = 1{,}914$
$M = 6.83$ oz.	$M = 7.48$ oz.
$\sigma = 1.23$ oz.	$\sigma = 1.33$ oz.

Is the difference in the mean breaking-load significant?

22. Karl Pearson and Alice Lee (*Biometrika*, Vol. II, p. 415) secured the measurements of the stature of 1078 fathers and sons. The analysis yielded the results:

Fathers	*Sons*
$N = 1{,}078$	$N = 1{,}078$
$M = 67.70$ inches	$M = 68.66$ inches
$\sigma = 2.72$ inches	$\sigma = 2.75$ inches

$r = 0.51$

Determine if the difference in the means is significant.

23. The following exercise is based upon data given in the "Proceedings of the American Society for Testing Materials," 1930, Vol. 30, Part II, pp. 448–455. A Committee of the Society, appointed to study corrosion, made numerous studies of the length of life of steel plates immersed in city water. The Committee found that the length of life was distributed normally. Numerous tests on No. 16 gauge sheets immersed in Washington tap water gave: $M_u = 1940$ days and $\sigma_u = 224$ days.

a. What is the probability that the mean of a sample of 100 sheets will not differ more than 25 days from M_u?

b. Find the 5 per cent level of significance for the mean of a sample of 100 sheets.

c. What should be the size of sample in future tests in order that the probability will not be greater than $\frac{1}{10}$ of the sample mean being in error by more than 74 days?

24. The following item appeared in the *New York Times* November 22, 1942. "TALL FRESHMEN — From Yale comes the news that the class of 1945 is the youngest and tallest that ever entered the university. Average freshman age is 18 years, 1 month and 11 days. Average height 5 feet 8.5 inches. Compared with his predecessor of World War I the Yale freshman of today is ten pounds heavier and 1.7 inches taller. Of all this year's Yale freshmen 21.6 per cent (227 in actual numbers) are over six feet tall."

Assuming $N = 1{,}000$, $\sigma_{weight} = 17$ pounds and $\sigma_{height} = 2.5$ inches, would you say the above item was noteworthy?

25. The ages of 5,317 husbands and wives were secured and the analysis of the data yielded the results:

Husbands	*Wives*
N = 5,317	N = 5,317
M = 42.8 years	M = 40.6 years
σ = 13.1 years	σ = 12.7 years

$r = 0.91$

Basing your judgment on these data would you state that the difference in the means is significant?

APPENDIX A

SELECTED BOOKS FOR SUPPLEMENTARY READING

A. General Texts

B. H. Camp, *Mathematical Part of Elementary Statistics*, D. C. Heath and Company, 1931.

F. E. Croxton and D. J. Cowden, *Applied General Statistics*, Prentice-Hall, 1941.

R. A. Fisher, *Statistical Methods for Research Workers, 7th edition*, Oliver and Boyd, London, 1938.

John F. Kenney, *Mathematics of Statistics*, D. Van Nostrand Company, 1939.

H. L. Rietz, *Mathematical Statistics*, Open Court Publishing Company, 1927. An excellent monograph for students who have had calculus and an elementary course in statistics.

L. H. C. Tippett, *The Methods of Statistics, 3rd edition*, Williams and Norgate, London, 1941. This book is mainly one of interpretations with the illustrations biological.

Alan E. Treloar, *Elements of Statistical Reasoning*, John Wiley and Sons, 1939. This book is concerned mainly with interpretations.

Albert E. Waugh, *Elements of Statistical Method*, McGraw-Hill Book Company, 1938.

G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics, 12th edition*, Charles Griffin and Company, London, 1940.

B. Texts in Special Fields

R. E. Chaddock, *Principles and Methods of Statistics*, Houghton Mifflin Company, 1925. Mainly descriptive and philosophical. It is intended for students of the social sciences.

Karl Holzinger, *Statistical Methods for Students in Education*, Ginn and Company, 1928.

F. C. Mills, *Statistical Methods Applied to Economics and Business, Revised*, Henry Holt and Company, 1938.

Raymond Pearl, *Medical Biometry and Statistics, second edition*, W. B. Saunders and Company, 1930.

George W. Snedecor, *Statistical Methods Applied to Experiments in Agriculture and Biology*, The Iowa State College Press, 1940.

Herbert Sorenson, *Statistics for Students of Psychology and Education*, McGraw-Hill Book Company, 1936.

C. Texts on Related Mathematical Topics

J. L. Coolidge, *An Introduction to Mathematical Probability*, Oxford University Press, 1925. A careful analysis of the fundamentals of the theory of probability.

Mordecai Ezekiel, *Methods of Correlation Analysis, second edition*, John Wiley and Sons, 1941.

Joseph Lipka, *Graphical and Mechanical Computation*, Part II. John Wiley and Sons, 1918. An excellent reference for curve-fitting which may be used in connection with that of Running.

H. L. Rietz, and others, *Handbook of Mathematical Statistics*, Houghton Mifflin Company, 1924. A useful reference book for one who has a good background in mathematics and statistics. A collection of chapters on important statistical topics.

T. R. Running, *Empirical Formulas*, John Wiley and Sons, 1917. A valuable reference for curve-fitting.

J. B. Scarborough, *Numerical Mathematical Analysis*, Johns Hopkins Press, 1930. Excellent chapters on interpolation, the normal curve, least squares, and empirical formulas. A valuable book.

Hugh H. Wolfenden, *The Fundamental Principles of Mathematical Statistics*, The Actuarial Society of America, New York, 1942.

D. Graphical Methods

W. C. Brinton, *Graphic Methods for Presenting Facts*, Engineering Magazine Company, New York, 1914.

S. C. Haskell, *How to Make and Use Graphic Charts*, Codex Book Company, New York, 1923.

K. G. Karsten, *Charts and Graphs*, Prentice-Hall, New York, 1923.

E. Aids in Calculation

J. W. Glover, *Tables of Applied Mathematics and Statistics*, George Wahr, Ann Arbor, Mich., 1923. This book contains many helpful tables.

Karl Pearson, *Tables for Statisticians and Biometricians*, Part I, *second edition*, Cambridge University Press. These tables are indispensable to the advanced student.

Mathematical Tables from Handbook of Chemistry and Physics, Chemical Rubber Company, Cleveland.

F. Sources for Current Statistical Data

Statistical Abstract of the United States, published annually by the Government Printing Office, Washington, D.C.

Yearbook of Agriculture, published annually by the Government Printing Office, Washington, D.C.

Survey of Current Business, United States Department of Commerce, Washington, D.C. Published monthly.

World Almanac and Book of Facts, New York World, New York.

APPENDIX B

AREAS AND ORDINATES OF THE NORMAL CURVE

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

The following table gives the values of the area under the curve from the ordinate at $t = 0$ to the ordinate for the values of t given in the column at the left. Values of the ordinate are also given.

t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$
.00	.0000	.3989	.40	.1554	.3683	.80	.2881	.2897	1.20	.3849	.1942
.01	.0040	.3989	.41	.1591	.3668	.81	.2910	.2874	1.21	.3869	.1919
.02	.0080	.3989	.42	.1628	.3653	.82	.2939	.2850	1.22	.3888	.1895
.03	.0120	.3988	.43	.1664	.3637	.83	.2967	.2827	1.23	.3907	.1872
.04	.0160	.3986	.44	.1700	.3621	.84	.2996	.2803	1.24	.3925	.1849
.05	.0199	.3984	.45	.1736	.3605	.85	.3023	.2780	1.25	.3944	.1827
.06	.0239	.3982	.46	.1772	.3589	.86	.3051	.2756	1.26	.3962	.1804
.07	.0279	.3980	.47	.1808	.3572	.87	.3079	.2732	1.27	.3980	.1781
.08	.0319	.3977	.48	.1844	.3555	.88	.3106	.2709	1.28	.3997	.1759
.09	.0359	.3973	.49	.1879	.3538	.89	.3133	.2685	1.29	.4015	.1736
.10	.0398	.3970	.50	.1915	.3521	.90	.3159	.2661	1.30	.4032	.1714
.11	.0438	.3965	.51	.1950	.3503	.91	.3186	.2637	1.31	.4049	.1692
.12	.0478	.3961	.52	.1985	.3485	.92	.3212	.2613	1.32	.4066	.1669
.13	.0517	.3956	.53	.2019	.3467	.93	.3238	.2589	1.33	.4082	.1647
.14	.0557	.3951	.54	.2054	.3448	.94	.3264	.2565	1.34	.4099	.1626
.15	.0596	.3945	.55	.2088	.3429	.95	.3289	.2541	1.35	.4115	.1604
.16	.0636	.3939	.56	.2123	.3411	.96	.3315	.2516	1.36	.4131	.1582
.17	.0675	.3932	.57	.2157	.3391	.97	.3340	.2492	1.37	.4147	.1561
.18	.0714	.3925	.58	.2190	.3372	.98	.3365	.2468	1.38	.4162	.1540
.19	.0754	.3918	.59	.2224	.3352	.99	.3389	.2444	1.39	.4177	.1518
.20	.0793	.3910	.60	.2258	.3332	1.00	.3413	.2420	1.40	.4192	.1497
.21	.0832	.3902	.61	.2291	.3312	1.01	.3438	.2396	1.41	.4207	.1476
.22	.0871	.3894	.62	.2324	.3292	1.02	.3461	.2371	1.42	.4222	.1456
.23	.0910	.3885	.63	.2357	.3271	1.03	.3485	.2347	1.43	.4236	.1435
.24	.0948	.3876	.64	.2389	.3251	1.04	.3508	.2323	1.44	.4251	.1415
.25	.0987	.3867	.65	.2422	.3230	1.05	.3531	.2299	1.45	.4265	.1394
.26	.1026	.3857	.66	.2454	.3209	1.06	.3554	.2275	1.46	.4279	.1374
.27	.1064	.3847	.67	.2486	.3187	1.07	.3577	.2251	1.47	.4292	.1354
.28	.1103	.3836	.68	.2518	.3166	1.08	.3599	.2227	1.48	.4306	.1334
.29	.1141	.3825	.69	.2549	.3144	1.09	.3621	.2203	1.49	.4319	.1315
.30	.1179	.3814	.70	.2580	.3123	1.10	.3643	.2179	1.50	.4332	.1295
.31	.1217	.3802	.71	.2612	.3101	1.11	.3665	.2155	1.51	.4345	.1276
.32	.1255	.3790	.72	.2642	.3079	1.12	.3686	.2131	1.52	.4357	.1257
.33	.1293	.3778	.73	.2673	.3056	1.13	.3708	.2107	1.53	.4370	.1238
.34	.1331	.3765	.74	.2704	.3034	1.14	.3729	.2083	1.54	.4382	.1219
.35	.1368	.3752	.75	.2734	.3011	1.15	.3749	.2059	1.55	.4394	.1200
.36	.1406	.3739	.76	.2764	.2989	1.16	.3770	.2036	1.56	.4406	.1182
.37	.1443	.3726	.77	.2794	.2966	1.17	.3790	.2012	1.57	.4418	.1163
.38	.1480	.3712	.78	.2823	.2943	1.18	.3810	.1989	1.58	.4430	.1145
.39	.1517	.3697	.79	.2852	.2920	1.19	.3830	.1965	1.59	.4441	.1127

t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$
1.60	.4452	.1109	2.00	.4773	.0540	2.40	.4918	.0224	2.80	.4974	.0079
1.61	.4463	.1092	2.01	.4778	.0529	2.41	.4920	.0219	2.81	.4975	.0077
1.62	.4474	.1074	2.02	.4783	.0519	2.42	.4922	.0213	2.82	.4976	.0075
1.63	.4485	.1057	2.03	.4788	.0508	2.43	.4925	.0208	2.83	.4977	.0073
1.64	.4495	.1040	2.04	.4793	.0498	2.44	.4927	.0203	2.84	.4977	.0071
1.65	.4505	.1023	2.05	.4798	.0488	2.45	.4929	.0198	2.85	.4978	.0069
1.66	.4515	.1006	2.06	.4803	.0478	2.46	.4931	.0194	2.86	.4979	.0067
1.67	.4525	.0989	2.07	.4808	.0468	2.47	.4932	.0189	2.87	.4980	.0065
1.68	.4535	.0973	2.08	.4812	.0459	2.48	.4934	.0184	2.88	.4980	.0063
1.69	.4545	.0957	2.09	.4817	.0449	2.49	.4936	.0180	2.89	.4981	.0061
1.70	.4554	.0941	2.10	.4821	.0440	2.50	.4938	.0175	2.90	.4981	.0060
1.71	.4564	.0925	2.11	.4826	.0431	2.51	.4940	.0171	2.91	.4982	.0058
1.72	.4573	.0909	2.12	.4830	.0422	2.52	.4941	.0167	2.92	.4983	.0056
1.73	.4582	.0893	2.13	.4834	.0413	2.53	.4943	.0163	2.93	.4983	.0055
1.74	.4591	.0878	2.14	.4838	.0404	2.54	.4945	.0159	2.94	.4984	.0053
1.75	.4599	.0863	2.15	.4842	.0396	2.55	.4946	.0155	2.95	.4984	.0051
1.76	.4608	.0848	2.16	.4846	.0387	2.56	.4948	.0151	2.96	.4985	.0050
1.77	.4616	.0833	2.17	.4850	.0379	2.57	.4949	.0147	2.97	.4985	.0049
1.78	.4625	.0818	2.18	.4854	.0371	2.58	.4951	.0143	2.98	.4986	.0047
1.79	.4633	.0804	2.19	.4857	.0363	2.59	.4952	.0139	2.99	.4986	.0046
1.80	.4641	.0790	2.20	.4861	.0355	2.60	.4953	.0136	3.00	.4987	.0044
1.81	.4649	.0775	2.21	.4865	.0347	2.61	.4955	.0132	3.01	.4987	.0043
1.82	.4656	.0761	2.22	.4868	.0339	2.62	.4956	.0129	3.02	.4987	.0042
1.83	.4664	.0748	2.23	.4871	.0332	2.63	.4957	.0126	3.03	.4988	.0041
1.84	.4671	.0734	2.24	.4875	.0325	2.64	.4959	.0122	3.04	.4988	.0039
1.85	.4678	.0721	2.25	.4878	.0317	2.65	.4960	.0119	3.05	.4989	.0038
1.86	.4686	.0707	2.26	.4881	.0310	2.66	.4961	.0116	3.06	.4989	.0037
1.87	.4693	.0694	2.27	.4884	.0303	2.67	.4962	.0113	3.07	.4989	.0036
1.88	.4700	.0681	2.28	.4887	.0297	2.68	.4963	.0110	3.08	.4990	.0035
1.89	.4706	.0669	2.29	.4890	.0290	2.69	.4964	.0107	3.09	.4990	.0034
1.90	.4713	.0656	2.30	.4893	.0283	2.70	.4965	.0104	3.10	.4990	.0033
1.91	.4719	.0644	2.31	.4896	.0277	2.71	.4966	.0101	3.11	.4991	.0032
1.92	.4726	.0632	2.32	.4898	.0271	2.72	.4967	.0099	3.12	.4991	.0031
1.93	.4732	.0620	2.33	.4901	.0264	2.73	.4968	.0096	3.13	.4991	.0030
1.94	.4738	.0608	2.34	.4904	.0258	2.74	.4969	.0094	3.14	.4992	.0029
1.95	.4744	.0596	2.35	.4906	.0252	2.75	.4970	.0091	3.15	.4992	.0028
1.96	.4750	.0584	2.36	.4909	.0246	2.76	.4971	.0089	3.16	.4992	.0027
1.97	.4756	.0573	2.37	.4911	.0241	2.77	.4972	.0086	3.17	.4992	.0026
1.98	.4762	.0562	2.38	.4913	.0235	2.78	.4973	.0084	3.18	.4993	.0025
1.99	.4767	.0551	2.39	.4916	.0229	2.79	.4974	.0081	3.19	.4993	.0025

t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$	t	$A_\phi\rceil_0^t$	$\phi(t)$
3.20	.4993	.0024	3.50	.4998	.0009	3.80	.4999	.0003	4.10	.5000	.0001
3.21	.4993	.0023	3.51	.4998	.0008	3.81	.4999	.0003	4.11	.5000	.0001
3.22	.4994	.0022	3.52	.4998	.0008	3.82	.4999	.0003	4.12	.5000	.0001
3.23	.4994	.0022	3.53	.4998	.0008	3.83	.4999	.0003	4.13	.5000	.0001
3.24	.4994	.0021	3.54	.4998	.0008	3.84	.4999	.0003	4.14	.5000	.0001
3.25	.4994	.0020	3.55	.4998	.0007	3.85	.4999	.0002	4.15	.5000	.0001
3.26	.4994	.0020	3.56	.4998	.0007	3.86	.4999	.0002	4.16	.5000	.0001
3.27	.4995	.0019	3.57	.4998	.0007	3.87	.5000	.0002	4.17	.5000	.0001
3.28	.4995	.0018	3.58	.4998	.0007	3.88	.5000	.0002	4.18	.5000	.0001
3.29	.4995	.0018	3.59	.4998	.0006	3.89	.5000	.0002	4.19	.5000	.0001
3.30	.4995	.0017	3.60	.4998	.0006	3.90	.5000	.0002	4.20	.5000	.0001
3.31	.4995	.0017	3.61	.4999	.0006	3.91	.5000	.0002	4.21	.5000	.0001
3.32	.4996	.0016	3.62	.4999	.0006	3.92	.5000	.0002	4.22	.5000	.0001
3.33	.4996	.0016	3.63	.4999	.0006	3.93	.5000	.0002	4.23	.5000	.0001
3.34	.4996	.0015	3.64	.4999	.0005	3.94	.5000	.0002	4.24	.5000	.0001
3.35	.4996	.0015	3.65	.4999	.0005	3.95	.5000	.0002	4.25	.5000	.0001
3.36	.4996	.0014	3.66	.4999	.0005	3.96	.5000	.0002	4.26	.5000	.0001
3.37	.4996	.0014	3.67	.4999	.0005	3.97	.5000	.0002	4.27	.5000	.0000
3.38	.4996	.0013	3.68	.4999	.0005	3.98	.5000	.0001	4.28	.5000	.0000
3.39	.4997	.0013	3.69	.4999	.0004	3.99	.5000	.0001	4.29	.5000	.0000
3.40	.4997	.0012	3.70	.4999	.0004	4.00	.5000	.0001			
3.41	.4997	.0012	3.71	.4999	.0004	4.01	.5000	.0001			
3.42	.4997	.0012	3.72	.4999	.0004	4.02	.5000	.0001			
3.43	.4997	.0011	3.73	.4999	.0004	4.03	.5000	.0001			
3.44	.4997	.0011	3.74	.4999	.0004	4.04	.5000	.0001			
3.45	.4997	.0010	3.75	.4999	.0004	4.05	.5000	.0001			
3.46	.4997	.0010	3.76	.4999	.0003	4.06	.5000	.0001			
3.47	.4997	.0010	3.77	.4999	.0003	4.07	.5000	.0001			
3.48	.4998	.0009	3.78	.4999	.0003	4.08	.5000	.0001			
3.49	.4998	.0009	3.79	.4999	.0003	4.09	.5000	.0001			

APPENDIX C

TABLES OF LOGARITHMS AND ANTILOGARITHMS

FOUR-PLACE LOGARITHMS

N	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
N	0	1	2	3	4	5	6	7	8	9

N	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
N	0	1	2	3	4	5	6	7	8	9

Logarithms	0	1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155
	0	1	2	3	4	5	6	7	8	9

Logarithms	0	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882
59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977
	0	1	2	3	4	5	6	7	8	9

ANSWERS TO EXERCISES

CHAPTER 1

Page 8

1. $\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots + \frac{1}{n^2}$.

2. $2^1 + 2^2 + 2^3 + 2^4 + \cdots + 2^n$.

3. $(1-3) + (2-3) + (3-3) + (4-3) + \cdots + (n-3)$.

4. ${}_{10}C_1 + {}_{10}C_2 + {}_{10}C_3 + \cdots + {}_{10}C_{10}$.

5. $a + 2a^2 + 3a^3 + 4a^4 + \cdots + na^n$.

6. ${}_{10}C_1 + 2 \cdot {}_{10}C_2 + 3 \cdot {}_{10}C_3 + 4 \cdot {}_{10}C_4 + \cdots + 10 \cdot {}_{10}C_{10}$.

7. $10f(10) + 20f(20) + 30f(30) + \cdots + 100f(100)$.

8. $25f(5) + 100f(10) + \cdots + 6400f(80)$.

9. $\sum_1^n x(x+1)$. **10.** $\sum_{i=1}^n (X_i - \bar{M})^2$.

Page 10

4. (1) 1,911. (2) 5,635.

Pages 13–14

1. a. $\Sigma x = 0$, $(\Sigma x)^2 = 0$, $\Sigma x^2 = 1{,}308$, $\sqrt{\Sigma x^2} = 36.16$.

b. $\Sigma U = 200$, $\Sigma U^2 = 5{,}486$, $\Sigma X = 700$, $\Sigma X^2 = 50{,}486$.

c. $\Sigma X^2 = 220$, $\Sigma Y^2 = 275$, $(\Sigma X)(\Sigma Y) = 1{,}050$, $\Sigma XY = 176$.

d. $\Sigma x = 0$, $\Sigma y = 0$, $\Sigma x^2 = 40$, $\Sigma xy = -34$, $\frac{\Sigma xy}{\Sigma x^2} = -0.85$.

Pages 17–18

1. (1) 4. (2) 3. (3) 2. (4) 3. (5) 5.

3. 0.00004. **4.** 0.00004. **5.** 2%. **6.** 0.147%. **7.** 0.04%.

8. 5.165×10^9; about 0.01%. **10.** (1) 2,142. (2) 2,774.

11. (1) 178.55. (2) 178.55. **12.** (1) 310.53. (2) 310.53.

Pages 21–22

1. 363 ± 0.5. **2.** $24{,}725 \pm 87.5$. **3.** $\frac{163}{25} \pm 0.112$. **4.** 4.05 sq. ft.

7. The former. **11.** $4n^3 + 4n^2 + 3n$. **12.** $\frac{1}{6}[4n^3 + 33n^2 + 89n]$.

13. 42,540. **15.** (1) 8,888. (2) 123,464. **17.** (1) 154,198. (2) 109,802.

18. $\frac{n(n+1)(2n+4)}{6}$. **19.** $\frac{n(n+1)(2n+7)}{6}$.

21. 24,001,875. **22.** $\frac{n(n+1)(n+2)(3n+5)}{12}$.

23. \$5.13 per ton; 0.000192.
24. \$150,000,000; 3.8%.
25. \$469,098,000; 0.178%.
26. 105.9 bu.; 0.29%.
27. \$330.75.
28. A: 39.37%; B: 37.25%.
29. 20%.
30. 9%.
31. \$187.50.
32. 93.5% of value in 1933.
33. 5.3 persons; 0.004.

CHAPTER 3

Pages 68–71

1. $M = 7.5$.
2. $M = 1.956$.
3. $M = 53.7$.
8. $M_X = 27$.
9. $M_X = 360$.
10. $M_X = 237.25$.
11. $M_X = 260$.
12. $M_X = 537.3$.
13. $M_X = 206.25$.
14. $M_X = 448$.
16. $M_X = 20.2$.
22. $M_X = 53.7$.

Pages 74–75

1. $M = 67.42$ inches.
2. $M = 139.39$ pounds.
3. $M = 6.06$ inches.
5. \$31.87.
6. \$35.08.
11. 22 cents.
12. \$16.77; \$21.48; \$15.90.
13. 1,000 lbs. per sq. in.

Page 79

1. $M_d = 1.98$ cm.
2. $M_d = \$449.50$.
3. a. $M_d = 67.52$ inches.
b. $M_d = 138.12$ pounds.
4. \$15.71; \$21.89; \$15.32.
5. $M_d = 991.3$ lbs. per sq. in.
7. 22.
8. $M_d = 53.7$ rays.

Page 85

	King	*Parabola*	*Pearson*	*Unit*
1.	1.998	1.997	2.02	cm.
2.	53.07	53.34	53.64	rays
3.	68.00	68.01	67.72	in.
4.	6.01	6.02	6.03	in.

Page 91

a. 3.27%. b. 3.45%. c. 2.5%.

Pages 91–92

1. $M_g = 17{,}043$.
2. 9.32%.
3. 20.5%; \$795; \$632.03; \$502.46.
4. 2.62%.

Pages 97–98

1. 26 mi. per day.
2. $66\frac{1}{4}$¢ per bu.
3. 15¢ per gal.
4. $24\frac{2}{3}$ days.
5. a. $9\frac{3}{5}$ units per hr.
b. $6\frac{1}{4}$ min.
c. 384.
6. 45 mi. per hr.
7. a. 8 problems per hr.
b. 7.5 min. per problem.

Pages 101–10

2. 86%. **3.** 5.8%. **4.** a. $M = 149.799$. b. $M = 150.77$.
5. a. $M_d = 151.09$. b. $M_d = 153.73$.
6. a. $M_o = 165.46$. b. $M_o = 164.74$.
7. a. 0.154. b. $M_g = 8{,}535.6$. **8.** $M = 38.1$ years.
9. $M = 6.9$ years. $M_d = 4.1$ years.
10. 4.76%. Estimated values: 42,222; 53,278; 67,224; 170,428.
11. a. 6.0759 pounds. b. 16.46¢ per lb.
14. 11.12¢ per lb.; 8.99¢ per lb.
15. a. 56.32. b. 41.92. **16.** a. 56.25. b. 41.95.
17. 166.6 years. **18.** 542.9 millions.
19. Group A: $M_d = \$7.54$; $M_X = \$7.395$.
Group B: $M_d = \$7.54$; $M_X = \$7.015$.
20. 1.3%; 11,085,900. **21.** $6(10)^6$.
22. 21.4%; \$485.58. **23.** $M = 648.7$; $M_g = 399.1$.
24. 8.56%. **25.** 7.875%; 2,120,350; 4.93%. **26.** About 7.5 years.
27. 4.99%. **28.** $\frac{11}{120}$ units per minute. **29.** 22 cents.
31. $M = \$1{,}371.72$; $M_d = \$1{,}365.37$; $M_o = \$1{,}432.83$.
32. $M = 3.39\%$. **33.** 108.9 millions of barrels.
35. Cincinnati: \$12.46; Cleveland: \$22.33; St. Louis: \$15.72.
36. $M = \$1{,}280.01$; $M_d = \$1{,}264.07$; $M_o = \$1{,}266.92$.
37. 10.6%. **38.** 1,267.9 millions of dollars.
39. At an infinite speed. **42.** np.

43. $M = \frac{1}{n+1}[2^{n+1} - 1]$; $M_g = \sqrt[n]{2^2}$; $M_h = \frac{(n+1)2^n}{2^{n+1} - 1}$.

CHAPTER 4

Page 114

1. 27%. **2.** 47%. **3.** No. **4.** 16%; 51%.
5. 7.5 to 8 inches. **6.** 50 to 55 pounds.
7. $M = M_d = M_o = 35$ pounds for A and B. No.

Page 119

1. 7.2%.
2. a.
$Q_1 = 65.85$ inches.
$Q_3 = 69.06$ inches.
$Q = 1.6$ inches.
$V_q = 2.4\%$.

b.
$Q_1 = 127.19$ pounds.
$Q_3 = 149.63$ pounds.
$Q = 11.22$ pounds.
$V_q = 8.1\%$.

4.

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
a.	90.8	114.4	128.0	139.4	151.1	162.3	171.4	186.8	203.8
b.	92.2	114.5	129.0	140.6	153.7	163.0	171.2	187.3	205.3

5. Absolute: $D_6 - D_4$, $D_7 - D_3$, etc. Relative: $\frac{D_6 - D_4}{D_6 + D_4}$, $\frac{D_7 - D_3}{D_7 + D_3}$, etc.

7. 5.731, 5.859, 6.192 inches. **8.** $Q_1 = 55.6$, $Q_3 = 75.4$.

9. Cincinnati: \$12.17 to \$20.17.
Cleveland: \$17.89 to \$25.48.
St. Louis: \$12.27 to \$18.57.

10. $Q_1 = 5.94$ inches, $Q_3 = 6.19$ inches, $Q = 0.125$ inch. No.

11. $Q_1 = 52.2$ rays, $Q_3 = 55.1$ rays.

12. \$1,197.00 to \$1,506.72. **13.** No.

Page 124

1. a. 0, 86, 14.3.
b. 0, 56, 9.3.
c. 0, 156, 26.0.

2. Wheat: 1.08 bu. per acre.
Rye: 1.3 bu. per acre.
Oats: 2.03 bu. per acre.

3. $M = 69.5$; $M.D.$ about $M = 11.32$; 58.1%. **4.** No.

Pages 127–28

1. Wheat: $M = 14.34$ bu. per acre; $\sigma = 1.23$ bu. per acre.
Rye: $M = 12.15$ bu. per acre; $\sigma = 1.6$ bu. per acre.
Oats: $M = 30.27$ bu. per acre; $\sigma = 2.4$ bu. per acre.

3. $M = 20$, $\sigma = 6.45$

Page 133

2. Cincinnati: $\sigma =$ \$6.86, $M =$ \$16.77.
Cleveland: $\sigma =$ \$6.28, $M =$ \$21.48.
St. Louis: $\sigma =$ \$6.04, $M =$ \$15.90.

3. $M =$ \$1,280.01, $\sigma =$ \$150.01.

4. $M =$ \$1,371.72, $\sigma =$ \$247.03, $V_\sigma = 18\%$.

5. A: $M = 100.95$, $M_d = 101.02$, $M_o = 101.1$, $\sigma = 13$.
B: $M = 47.71$, $M_d = 47.68$, $M_o = 48.08$, $\sigma = 5.88$.

Page 139

4. a.

	1st 100	*2nd 100*	*3rd 100*	*4th 100*	*5th 100*	*6th 100*	*7th 100*	*8th 100*	*9th 100*	*10th 100*	*Total*
M	142.35	138.75	138.65	139.05	138.35	139.35	137.05	138.75	140.65	138.55	139.15
σ	22.8	20.1	19.1	18.2	14.9	17.2	16.2	16.8	17.7	15.4	18.03

b. $\sigma_{means} = 1.36$ pounds. $M_{means} = 139.15$ pounds.
c. Seven. d. They are equal.

Pages 147–49

1. $\sigma = 0.20$ inch, $M.D. = 0.159$ inch.
2. $\sigma = 0.19$ cm., 49.3, Yes, $M = 1.956$ cm.
3. $M = 5$, $\sigma = 1.58$, 971.
4. a. $M = 67.42$ inches, $\sigma = 2.43$ inches, $V_\sigma = 0.036$.
b. $M = 139.39$ pounds, $\sigma = 17.2$ pounds, $V_\sigma = 0.12$.
5. a. $M = 149.8$, $\sigma = 42.5$, $V_\sigma = 0.28$.
b. $M = 150.8$, $\sigma = 42.2$, $V_\sigma = 0.28$.
6. a. $M = 56.323$ mm., $\sigma = 2.404$ mm., $V_\sigma = 0.043$.
b. $M = 41.917$ mm., $\sigma = 1.385$ mm., $V_\sigma = 0.033$.
7. b. $\frac{1}{3}\sqrt{3(N^2 - 1)}$. **9.** 17.
10. $Q = 5.3, Q_1 = 66.7, Q_3 = 77.3, M_o = 72, E_X = 5.3, E_M = 0.17, E_\sigma = 0.12$
11. a. $E_M = 0.04$ inch, $E_\sigma = 0.03$ inch.
b. $E_M = 0.3$ pound, $E_\sigma = 0.2$ pound.
12. $E_M = 0.004$ inch, $E_\sigma = 0.003$ inch.
13. $V_{\sigma_1} = 0.22$, $V_{\sigma_2} = 0.26$.
14. Scores: $E_M = 0.37$, Production: $E_M = 0.97$.
15. (1) 200. (2) 285, \$90 and \$30.
16. The first distribution. **17.** The distribution of weights.
27. $M = np$, $\sigma = \sqrt{npq}$. **28.** $X = \dfrac{\Sigma X_i}{N} = M_X$.

Page 154

	A	B	C	D
M	20	26.7	13.3	9.75
M_d	20	28.96	11.04	7.5
σ	7.2	7.25	7.25	6.42
Sk	0	−0.93	0.93	1.05

CHAPTER 5

Pages 167–68

1. 0.0083. **2.** a. − 0.125. b. 0.220. **3.** a. − 0.047. b. 0.026.
4. The unadjusted values are:

	a	b
M	149.8	150.77
σ	42.5	42.2
α_3	− 0.05	− 0.08
α_4	2.65	2.58

5.

	Unadjusted Values		*Adjusted Values*	
	a	*b*	*a*	*b*
M	56.322	41.917	56.322	41.917
σ	2.404	1.385	2.386	1.378
α_3	0.603	0.128	0.616	0.130
α_4	4.334	2.8904	4.373	2.8832

7. $M = 6.062$ inches, $\sigma = 0.20$ inch.

10.

	Unadjusted	*Adjusted*
M	39.835	39.835
M_d	39.831	.....
M_o	39.802	.....
σ	2.052	2.032
α_3	0.0287	0.0296

11.

	Unadjusted	*Adjusted*
M	171.8917	171.8917
M_d	171.8858	.....
M_o	173.407	.....
σ	6.8236	6.7992
α_3	0.1125	0.1137
α_4	3.194	3.197

CHAPTER 6

Pages 177–78

1.

Year	*Relatives to 1909*	*Year*	*Relatives to 1909*
1909	100	1916	90
1910	90	1917	80
1911	83	1918	72
1912	88	1919	78
1913	86	1920	76
1914	84	1921	61
1915	83	1922	71

2.

Year	*Relatives to 1909*	*Link Relatives*	*Year*	*Relatives to 1909*	*Link Relatives*
1909	100	...	1914	138	106
1910	106	106	1915	151	109
1911	111	105	1916	160	106
1912	120	108	1917	175	110
1913	130	108	1918	190	109

3.

Year	*Rel.*	*Year*	*Rel.*
1910	67	1920	87
1911	69	1921	87
1912	72	1922	101
1913	80	1923	120
1914	77	1924	130
1915	75	1925	141
1916	80	1926	144
1917	81	1927	151
1918	62	1928	154
1919	71	1929	150

4.

Year	*Rel. (Corn)*	*Rel. (Hogs)*	*Average Price Rel.*
1920	96	138	117
1921	60	84	72
1922	94	91	93
1923	103	75	89
1924	140	80	110
1925	96	117	107
1926	91	122	107
1927	103	99	101
1928	107	91	99
1929	111	100	106

Pages 183–84

1.

Year	*1921*	*1923*	*1925*	*1927*	*1929*
Aggregative Relative	100	104.5	103	99.6	96.9

2.

Year	*1921*	*1923*	*1925*	*1927*	*1929*
Harmonic Mean of Relatives	100	119	134	127	128

3.

Year	*1921*	*1923*	*1925*	*1927*	*1929*
Aggregative Relative	100	117	138	123	124
Arithmetic Mean of Rel.	100	123	137	132	132
Median of Relatives	100	116	140	122	125
Geometric Mean of Rel.	100	121	136	130	130
Harmonic Mean of Rel.	100	119	134	127	128

4.

Year	*1921*	*1923*	*1925*	*1927*	*1929*
Arith. Mean	100	105	102	115	108
Median	100	100	95	110	99
Geo. Mean.	100	105	100	113	102

Page 187

1.

Year	*1915*	*1918*	*1920*	*1922*
a. Simple Agg. Rel.	107	179	225	370
b. Simple A.M. of Rel.	103	168	231	144
c. Weighted Agg. Rel.	...	...	217	140

Page 196

1. 146. **2.** 151.

3.

	1921	*1929*
(1)	96.7	119.5
(2)	91.4	117.3
(3)	94.0	118.4
(4)	97.4	138.9
(5)	91.3	117.8

Page 200

6. 150.6.

7. 94.0; 118.3.

8.

	1921	*1923*
(1)	89.7	94.7
(2)	91.4	90.4
(3)	90.5	92.5
(4)	79.7	124.1
(5)	90.5	92.2

9. 147.

CHAPTER 7

Page 205

1. 2.5. **2.** 1. **3.** -4.5. **4.** $-\frac{10}{7}$. **5.** 0. **6.** 0. **9.** 3. **10.** -2.

Pages 209–10

1. $Y = 3X + 2$. **2.** $Y = 3X - 2$.
3. $Y = -3X + 2$. **4.** $Y = -3X - 2$.
5. a. Slope-intercept form. $m = 3$, $b = -4$.
6. a. $Y = 5X - 7$. b. $2Y = 3X + 10$. c. $Y = -X + 8$.
8. $7X - 5Y + 4 = 0$, $-\frac{4}{7}$, $\frac{4}{5}$. **13.** $Y = 3X - 1$.
14. $4X + 3Y = 17$. **15.** Yes. **16.** No

Page 216

$Y = 2.975X - 2.025$.

Pages 220–21

1. $R = 0.02799t + 10.122$. **2.** $l = 0.02w + 90.22$.

3.

l	w	l	w
50	91.22	350	97.22
100	92.22	400	98.22
150	93.22	450	99.22
200	94.22	500	100.22
250	95.22	550	101.22
300	96.22	600	102.22

Pages 225–26

1. a. $Y = 0.5X + 1$.
b. $Y = 2.6X - 2$.
2. $S = 0.52T + 54.2$, 80.2.
4. a. $W = 1.02L - 3.123$.
c. $Y = -0.85X + 12.1$.
d. $Y = -1.5X + 50$.
3. $Y = 0.765X + 22.9$, 68.8.
5. $L = -0.675T + 603.5$, 552.875.

T	*Observed* L	*Computed* L
70	556	556.25
80	550	549.5
90	542	542.75
100	536	536.00
110	530	529.25
120	523	522.5
130	515	515.75

Pages 229–30

1. $Y = 5.75X + 111.475$ with $X = 0$ at 1919.
2. (1) $Y = 2.65X + 18.304$. (2) \$28.90.
3. (1) $Y = 3.483X + 26.44$. (2) \$43.86 millions.
4. $Y = -0.12X + 5.2$.

CHAPTER 8

Pages 236–37

3. (1) $Y = -0.46X + 5.53$.
(3) $S_y = 0.53$ thousand strikes and lockouts.
(4) 1.85 thousand strikes and lockouts.
(5) Computed $Y = 0.65$.
4. (1) $Y = 9.82X - 29.7$.
(3) $S_y = \$17.6$.
(4) Computed $Y = \$68.5$.
5. (1) $Y = 11.1X - 217.83$.
(3) $S_y = \$45.2$.
(4) Computed $Y = \$226.2$.

Page 243

1. $r = 0.95$, $Y = 1.02X - 12.62$, $S_y = 3.93$ c.u.
2. $r = 0.89$, $Y = 0.075X + 4.72$, $S_y = 0.58$ ton.
3. $r = 0.61$, $Y = 3.32X + 21.93$, $S_y = 3.3$ bu. per acre.

Pages 250–52

1. $r = -0.92$. **2.** $r = 0.61$. **3.** a. $r = -0.62$.
5. $r = 0.95$. **6.** $r = -0.84$. No. b. $r = -0.55$.
10. $\sigma_X = \sigma_Y = 3.42$. $r = m = 1$. $Y = X + 4$.
11. a. $\sigma_X = 1.414$. $\sigma_Y = 2.828$. $r = -1$. $m = -2$. $Y = -2X + 12$.
11. b. $\sigma_X = 1.414$. $\sigma_Y = 4.242$. $r = -1$. $m = -3$. $Y = -3X + 13$.
12. a. $\sigma_X = 4.87$. $\sigma_Y = 3.78$. $r = 0$. $m = 0$. $Y = 4$.
12. b. $\sigma_X = 3.78$. $\sigma_Y = 2.27$. $r = 0$. $m = 0$. $Y = 3$.

Pages 260–262

1. $M_X = 53.77¢$. $M_Y = \$7.82$. $r = 0.72$.
$\sigma_X = 25.2¢$. $\sigma_Y = \$4.34$. $Y = 0.124X + 1.15$.
If $X = 75$, $Y_{est.} = \$10.45$. $S_y = \$3.04$.
2. $r = .40$. $Y = 1.36X + 108.5$. $X = 0.12Y + 13.2$.
3. $M_X = 16.25$ min. $M_Y = 82.125\%$. $r = -0.92$.
$\sigma_X = 5.04$ min. $\sigma_Y = 9.25\%$. $Y = -1.68X + 109.4$
If $X = 20$, $Y_{est.} = 76\%$. $S_y = 3.63\%$.

Page 266

1. 0.92.
2. 0.85.
3. $\rho_{I\,II} = 0.64$. $\rho_{I\,III} = 0.62$. $\rho_{II\,III} = 0.78$.

Pages 270–76

1. (1) $Y = 0.075X + 4.72$.
(2) $r = 0.89$. Yes.
(3) If $X = 40$, $Y_{est.} = 7.72$ tons.
(4) $S_y = 0.58$ ton.
2. $r = 0.92$.
3. $r = 0.63$. $Y = 1.21X + 4.73$. $X = 0.32Y + 14.3$.
4. $r = -0.84$. Yes.
7. $r = 0.60$. $Y = 0.85X + 85.03$. $X = 0.43Y - 19.3$.
10. (1) $r = 0.829$. $Y = 0.65X + 2.75$.
(2) $Y_{est.} = 6\%$. (3) $M_Y = 6.14\%$.
(4) $X = 1.056Y - 1.26$. (5) $X_{est.} = 6.13\%$.
(6) $M_X = 6.29\%$. (7) $S_y = 0.39\%$, $S_x = 0.50\%$.

11. $r = 0.80$.
12. (2) $r = -0.71$.
(3) $m = -0.32$.
(4) $Y = -0.32X + 187.69$.
(5) 155.7¢, 123.7¢, 91.7¢.
(6) $S_y = 33.6¢$.
13. $r = 0.73$.
14. $r = 0.90$. Yes, spurious.
15. (2) If $X = \$175$, $Y = \$138.88$.
16. The Bucknell test.
(3) \$0.746.

CHAPTER 9

Page 282

2. (1) $X_1 = 0.384X_2 + 1.646X_3 + 1.438$.

Pages 284–285

2. (2) $X_1 = 0.839X_2 + 0.462X_3 - 0.270$.
(4) $R_{1(23)} = 0.83$, $S_{1(23)} = 1.47$.
3. (2) $R_{1(23)} = 0.96$, $S_{1(23)} = 5.02$.
(3) $X_1 = 0.258X_2 + 0.606X_3 + 14.2$.
4. (1) $X_1 = 0.575X_2 + 1.092X_3 + 15.982$.
(2) 158.6. (3) 35.9.

Page 288

6. (2) $R_{1(23)} = 0.96$, $S_{1(23)} = 2.36$.

Page 290

1. $(-1, 3)$. **2.** $(6, 0)$. **3.** $(24, -16)$. **4.** $(\frac{5}{4}, -\frac{3}{2})$.

Page 292

2. $(-2, 1, 3)$. **3.** $(-1, 2, -3)$.

Page 293

1. (1) -15. (2) -56.

Page 297

5. 72.88%.

Pages 301–303

1. a. Weight = 0.994 Length + 2.660 Breadth − 112.217.
b. 55.25 grams.
c. Weight = 0.046 Length + 1.056 Bulk − 2.081.
d. Weight = 1.098 Bulk − 0.098 Breadth + 2.416.
e. $S_{W(L\,Br)} = 0.924$, $S_{W(L\,Blk)} = 0.907$, $S_{W(Blk\,Br)} = 0.909$.
2. a. $X_1 = 0.55X_2 + 1.07X_3 + 0.083X_4 - 69$.
c. $R_{1(234)} = 0.826$.
d. $r_{12\cdot34} = 0.764$, $r_{13\cdot24} = 0.676$, $r_{14\cdot23} = 0.09$.

CHAPTER 10

Pages 310–311

5. $a = 1, b = -2, c = 2$. If $X = 5$, $Y = 17$.
6. $Y = X^3 - 2X$. If $X = 5$, $Y = 115$.

Page 313

$R = 0.02792t + 10.1368$.

Page 315

3. With $X = 0$ at 1924, $Y = 3.483X + 26.44$.
At 1929, $X = 5$, $Y = 43.86$ millions of dollars.

Page 323

2. Using L.S., $p = 30(0.99996)^h$.

h	1,000	2,000	5,000
p	28.9	27.9	24.9

3. L.S. gives $T = 17.8921(0.9865)^t$.
4. With $t = 0$ at 1920, L.S. gives $X = 100{,}006(1.127)^t$.
5. L.S. gives $H = 0.86(1.39)^D$.

Pages 329–330

2. The points (8, 23) and (20,360) give $Y = 0.045X^3$.
4. $Y = 0.119X^{1.03}$.
5. $V = 2.26l^{0.5}$.
7. $T = 49.5h^{1.38}$.

Page 344

1. a. $\log Y = (\log 3)X + \log 1$, $Y = 3^X$.
b. $\log Y = 0.2X + \log 1$, $Y = 10^{\frac{X}{5}}$.
c. $\log Y = 0.1X + \log 1$, $Y = 10^{\frac{X}{10}}$.
d. $\log Y = (\log 5^{\frac{1}{3}})X + \log 2$, $Y = 2(5^{\frac{X}{3}})$.
e. $\log Y = -0.1X + 1$, $Y = 10(10^{-0.1})^X$.
f. $\log Y = (\log 2^{\frac{1}{3}})X + \log 2^{-\frac{4}{3}}$, $Y = 2^{\frac{X-4}{3}}$.

Pages 350–354

2. $Y = 0.045X^3$.
4. $Y = 0.04X^2$.
6. $Y = 4(1.2)^X$.
7. $N = 125(1.649)^t$.
10. $Y = 2.54(1.16)^X$.
16. Choosing $X = 0$ at 1909, $Y = 0.305X + 7.36$.
17. Choosing $X = 0$ at 1907, $Y = 14.375X + 159.31$.
At 1915, $X = 8$, $Y = 274.31$.
At 1920, $X = 13$, $Y = 346.185$.
18. Choosing $X = 0$ at 1909, a. $Y = 74X + 1{,}988.7$.
b. $Y = 109.8X + 2{,}053.8$.
19. Choosing $X = 0$ at 1915, $Y = 1.35X + 31.8$.

Pages 357–361

1. With $X = 0$ at 1910, L.S. gives $Y = 0.435X + 35.2$.
2. L.S. gives $V = 499.82p^{-1.066}$.
3. With $X = 0$ at 1910, L.S. gives
   $Y = 0.0574X^2 + 2.67X + 94.66$.
4. With $X = 0$ at 1900, L.S. gives $Y = 0.714(1.031)^X$.
   At 1915, $X = 15$, $Y = 1.13$.
   At 1928, $X = 28$, $Y = 1.67$.
5. Using L.S., $S = 44.603(1.049)^\theta$.
   $S = 0.0014725\theta^2 - 0.474\theta + 49.548$.
6. L.S. gives $V = 3.1944 + 0.4516D - 0.7792D^2$.
   If $D = 0.9$, $V = 2.9697$.
7. Using first, seventh, and ninth points, $\theta = 31.5 + 60(0.9038)^t$.
8. With $X = 0$ at 1910, $Y = 19(1.086)^X$.
9. $I = 4.480D^{0.6691}$.
14. L.S. gives with $t = 0$ at 1909.5, $X = 393.3(1.0743)^t$.
15. Using first, sixth, and eleventh points, $y = 10.1344 + 1.7521(1.2404)^X$.

CHAPTER 11

Pages 365–366

1. 4.
2. 8.
3. 36.
4. 288.
5. 504.
6. 2,730.
7. 3,024.

Page 367

1. 156.
2. 6,720.
3. a. 362,880. b. 725,760. c. 725,760. d. 2,903,040.
4. 720.
5. 10.
7. 30,240.
8. 34,650.
9. 2,520.

Pages 369–370

1. 45; 45; 4,950.
2. 50,063,860.
3. 5,880.
4. 45.
5. 63.
6. a. 126. b. 84.
8. 302,400.
9. 878,948,939.
12. 3,600.
13. a. 700. b. 1,408.
15. $n = 11, r = 2$.
16. $n = 6$.
17. $n = 10$.
18. $n = 7$.

Pages 371–372

1. $\frac{2,048}{4,040}, \frac{1,992}{4,040}$.
2. $\frac{2}{128}, \frac{8}{128}$, etc. $M = 3.43$, $\sigma = 1.3$.
3. 0.0085.
4. 0.514.

Pages 373–374

1. $\frac{1}{6}, \frac{1}{3}, \frac{1}{2}$.
2. $\frac{1}{36}$.
3. $\frac{1}{18}, \frac{1}{6}, 7$.
4. $\frac{1}{13}, \frac{2}{13}, \frac{1}{4}, \frac{1}{2}$.
5. $\frac{5}{18}$.
6. $\frac{11}{36}$.
7. $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$.
8. $\frac{1}{8}, \frac{1}{8}, \frac{3}{8}$.
9. The former.
10. $\frac{1}{72}$.
11. $\frac{1}{6}$.
12. $\frac{11}{850}$.

Pages 376–377

1. a. $\frac{2}{1,001}$. b. $\frac{240}{1,001}$. **2.** $\frac{1}{11,050}$. **3.** $\frac{1}{16}$. **4.** $\frac{23}{24}$.
5. a. 0.06. b. 0.56. c. 0.38. **6.** $\frac{7}{9}$. **7.** $\frac{2}{3}$.

Pages 380–382

1. a. $\frac{1}{128}$. b. $\frac{7}{128}$. c. $\frac{21}{128}$, etc. **2.** 1, 7, 21, etc.

3. a. $\frac{15(5^4)}{6^6}$. b. $\frac{2,906}{6^6}$. **4.** $\frac{4,651}{6^5}$. **6.** $\frac{56}{2^{10}}$.

7. a. 0.2646. b. 0.3483. **8.** 0.09.

9. a. $\frac{80}{3^5}$. b. $\frac{51}{3^5}$. **10.** $\frac{276}{6^5}$. **11.** 25.

12. a. $10(.94)^3(.06)^2$. b. $(.06)^2(.94)^3$.
c. $10(.94)^3(.06)^2 + 10(.94)^2(.06)^3 + 5(.94)(.06)^4 + (.06)^5$.

13. a. $(.95)^5$. b. $10(.95)^2(.05)^3$.
c. $10(.95)^2(.05)^3 + 5(.95)(.05)^4 + (.05)^5$.

14. b. $(.95)^{10}$. a. $10(.95)^9(.05)$. c. $(.95)^{10} + 10(.95)^9(.05)$.
d. $\sum_{r=5}^{10} {}_{10}C_r(.95)^r(.05)^{10-r}$. **15.** ${}_{25}C_{20}(.9)^{20}(.1)^5$.

16. a. ${}_{10}C_6(.97)^6(.03)^4$. b. $\sum_{r=6}^{10} {}_{10}C_r(.97)^r(.03)^{10-r}$.

17. $\sum_{r=90}^{100} {}_{100}C_r(.95)^r(.05)^{100-r}$. **19.** $\frac{5}{6}$.

20. a. $\frac{108}{7^6}$. b. $\frac{799}{7^7}$. c. $\frac{36}{7^7}$. **21.** a. $\frac{3}{5}$. b. $\frac{19}{20}$.

22. a. $\frac{1}{1,024}$. b. $\frac{63}{256}$. c. $\frac{7}{128}$. **23.** a. 5. b. $\frac{63}{256}$. c. $\frac{7}{128}$.
24. 4. **25.** $\frac{1}{6}$. **26.** $\frac{1}{8}$. **27.** $\frac{7}{72}$.

CHAPTER 12

Page 390

1.

	a	b	c	d
M_o	5	1	4	2 and 3
M	5	1	3.6	2.4
σ	0.91	0.91	0.6	0.98
α_3	− 0.73	0.73	1.33	0.20

Page 395

1.

A		B	
X	*Graduated* $f(x)$	X	*Graduated* $f(x)$
60.5	3.5	29.5	2.0
70.5	28.3	33.5	17.6
80.5	99.0	37.5	70.3
90.5	198.0	41.5	164.1
100.5	247.5	45.5	246.1
110.5	198.0	49.5	246.1
120.5	99.0	53.5	164.1
130.5	28.3	57.5	70.3
140.5	3.5	61.5	17.6
		65.5	2.0
Total	905.1	*Total*	1,000.2

Page 404

1. a. 0.9773. d. 0.0227. g. 0.0227.
b. 0.9836. e. 0.9918. h. 0.9892.
c. 0.9834. f. 0.0084. i. 0.0027.

2. a. 2.14. b. 0.65. c. 1.655. d. 0.6553.

Pages 411–413

1. $t = 5$. No.

2. a. $t = 6.1$. Yes. b. $t = 3.46$. Yes. From point of view of chance, this might happen, but very improbable.

3. 0.0108.

4.

Grade	F	E	D	C−	C	C+	B−	B	B+	A−	A	A+
Number receiving grade	1	7	28	79	159	226	226	159	79	28	7	1

7. The values of t for the five parts are: $-\infty$ to -0.8415, -0.8415 to -0.2533, -0.2533 to $+0.2533$, 0.2533 to 0.8415, 0.8415 to $+\infty$.

8. 929.0.

9. a. Yes. $t = 0.88$. b. Yes. $t = 2.5$. c. No. $t = 7$.

10.

X	*Binomial Ordinates*	*Normal Ordinates*	X	*Binomial Ordinates*	*Normal Ordinates*
0	.000	.000	9	.175	.176
1	.000	.0005	10	.122	.121
2	.002	.002	11	.067	.065
3	.0085	.009	12	.028	.027
4	.028	.027	13	.0085	.009
5	.067	.065	14	.002	.002
6	.122	.121	15	.000	.0005
7	.175	.176	16	.000	.000
8	.196	.199			

11. 0.9.

12. a. 0.95. b. 0.95.

13. a. 0.076. b. 0.076.

14. 2.5 inches.

15. a. 720 men. b. 0.000. c. 0.12.

16. a. 50 units. b. 36 and 64 units.

17. a. 900. b. Yes. $\sigma = 28.6$.

19. $Q_1 = 53.3$, $Q_3 = 66.7$, $Q = 6.7$, $M.D. = 8$, $\alpha_3 = 0$, $\alpha_4 = 3$, 87th percentile = 71.2.

20. 8.5 and 18.7.

21. a. 0.24. b. 0.02. c. 0.06. d. 0.007.

Page 417

1. Using $M = 39.835$, adjusted $\sigma = 2.0322$, $\frac{1}{\sigma} = 0.492078$, and rounding the values of t to two decimal places, we have

X	*Theoretical $f(x)$ Ordinates*	*Theoretical $f(x)$ Areas*	X	*Theoretical $f(x)$ Ordinates*	*Theoretical $f(x)$ Areas*
33	7	8	42	1108	1110
34	32	34	43	582	592
35	116	123	44	240	252
36	329	339	45	78	81
37	737	746	46	20	21
38	1309	1295	47	4	4
39	1805	1818	48	1	1
40	1957	1929			
41	1669	1646	*Total*	9994	9999

2. $M = 67.92$.
$\sigma = 2.42$.
$\frac{1}{\sigma} = 0.4132$.

3. $M = 6.06$.
$\sigma = 0.20$.
$\frac{1}{\sigma} = 5.00$.

Values of t are rounded to two decimal places.

l_x	*Theoretical* $f(x)$
57.5	0.2
58.5	0.6
59.5	2.4
60.5	7.6
61.5	21.4
62.5	47.6
63.5	91.7
64.5	154.1
65.5	207.9
66.5	242.4
67.5	244.8
68.5	199.2
69.5	140.7
70.5	85.6
71.5	41.8
72.5	18.0
73.5	6.5
74.5	2.0
75.5	0.5
76.5	0.2
Total	1515.2

l_x	*Theoretical* $f(x)$
5.45	4.3
5.55	14.8
5.65	40.4
5.75	86.3
5.85	144.3
5.95	188.9
6.05	193.5
6.15	155.3
6.25	97.6
6.35	47.9
6 45	18.5
6.55	5.5
6.65	1.3
6.75	0.3
6.85	0.0
Total	998.9

4. $Y = \frac{924(4)}{11.0668\sqrt{2\pi}} e^{-\frac{t^2}{2}} = (333.972)\phi(t)$ where $t = \frac{X - 74.2}{11.0668}$.

CHAPTER 13

Pages 438–439

2. 5,625; 22,500.
3. 0.034 pound.
4. 0.0525 million per cu. mm.
5. a. 0.9255. b. 0.0026.
6. Yes.
12. Yes. $t = 0.62$.
13. Yes. $t = 2$.

Pages 449–450

1. a.

	1st 100	*2nd 100*	*3rd 100*	*4th 100*	*5th 100*	*6th 100*	*7th 100*	*8th 100*	*9th 100*	*10th 100*	*Total*
M	117.15	120.85	117.35	122.95	119.25	118.45	118.05	113.65	119.45	120.25	118.74
σ	13.8	17.4	17.6	21.4	15.5	17.8	18.0	13.4	17.5	16.2	17.2

   b. Mean of means = 118.74. $M_u = 118.74$.
   c. Mean of σ's = 16.9. $\sigma_u = 17.2$.
   d. Eight. The five per cent limits are 115.37 and 122.11.
   e. Seven. The five per cent limits are 14.8 and 19.6.
   f. Sampling probably went awry on the 4th 100 and the 8th 100, for both the mean and the standard deviation are outside the five per cent levels.

2. Difference not due to chance. $t = 13+$.
3. Yes. $t = 6+$.
4. Class of 1943 is poorly prepared. $t = 4.4$.
   Class of 1945 is within 5 per cent level. $t = 1.7$.
5. $\sigma_{p_1-p_2} = 0.0078$. Difference not significant. $t = 0.5+$.
6. $\sigma_{p_2-p_1} = 0.0417$. Difference not significant. $t = 1.1+$.

Pages 457–464

1. $E_M = 1.57$. Even chance that sample mean, 149.8, does not differ from M_u by more than $\pm$ 1.57.
   $\sigma_\sigma = 1.64$. A two to one chance that the sample σ, 42.47, does not differ from σ_u by more than $\pm$ 1.64.
2. a. About 0.58. b. About 0.62.
3. a. $\sigma_M = 0.364$, $\sigma_\sigma = 0.257$ p.b. per min. b. About 0.994.
4. $r = 0.77$, $\sigma_r = 0.05$. A two to one chance that the sample r does not differ from the universe r by more than $\pm$ 0.05. $E_r = 0.03$.
5. $E_M = 0.0137$ inch. An even chance that the sample mean, 39.835 inches, does not differ from the universe mean by more than 0.0137 inch.
6. $t = 2.7$ and the difference is probably significant.
7. $t = 167$, and the difference is certainly significant. In fact, Group I was American soldiers and Group II was Japanese soldiers.
8. For National League, $M = 0.283$, $\sigma = 0.086$.
   For American League, $M = 0.278$, $\sigma = 0.085$.
   $t = 0.47$, and the difference is not significant.
10. $t = 14.6$, and the difference in the means is sufficient to warrant the conclusion that Scots are taller than Englishmen.

11. Yes. $k = 5.8$.
13. $t = 1.5$. **14.** $t = 11.7$.

15.

	1st Group	*2nd Group*
N	72,127	17,986
M	\$8.37	\$9.59
σ	\$2.49	\$2.43

$t = 60$. Hence $(M_2 - M_1)$ is significant.
16. $t = 35.3$. **17.** 96.
18. $t = 4.15$. **19.** $t = 7.96$; $t = 0.12$.
20. $t = 2.53$. Probably significant.
21. Yes. $t = 16+$. **22.** Yes. $t = 11.6$.
23. a. About 0.73. b. About 44 days. c. $N = 25$.
24. Yes. For $t = 15+$ for heights and $13+$ for weights.
25. Yes.

INDEX

The numbers refer to pages.

Absolute error, 16
Absolute value of a number, 121
Accuracy, in measurements, 14; measurements of, 15
Aggregative relatives, simple, 179; weighted, 185
Analysis, statistical, 2
Arithmetic mean, as a moment, 62; calculation of, 60, 62, 73; criticism of, 99; defined, 60; of relatives, 181–182, 189–190; probable error of, 143, 432; standard deviation of, 144, 430; standard error of, 144
Array, nature of, 24
Asymmetry, defined, 43, 52; Pearson's measures of, 151, 152; positive and negative, 152–153; quartile measure of, 156; third moment as measure of, 157
Average, characteristics of a good, 59; uses of an, 59; of relatives, 182–184, 188–200
Average deviation, 120

Base, in index number construction, 175
Benson, Paul, Preface
Bernoulli, James, 395
Bernoulli Theorem, 395
Bessel, Friedrich Wilhelm, 138, 452
Bias, downward, 197; in averages of relatives, 197; in use of weights, 198; type, 197; upward, 197; weight, 197
Binomial expansion, 367, 383–395
Binomial, point, arithmetic mean of, 388; general form of, 384; graduation of data by, 391–395; mode of, 386; skewness and excess of, 389–390; standard deviation of, 388
Birge, R. T., 456
Bowley, A. L., 5, 156
Bradstreet's index number, 180
Brahe, Tycho, 339
Bravais, A., 237
Brinton, W. C., 466
Brown, W. and Thomson, G. H., 240
Bruce, C. W., 110
Burgess, Robert W., 98, 197

Camp, B. H., 36, 465
Carver, H. C., Preface, 456
Central tendency, 52, 98–101; measures of, 59–110
Chaddock, R. E., 269, 465
Charts, construction of, 37–53
Class, boundary, 26; frequency, 25; interval, 25, 26, 30; limits, 26, 30; mark, 27; unit, 70, 71, 72; width, 25
Classification of data, 23
Coefficient of, correlation, 238, 245, 246, 254; multiple correlation, 281, 284, 295, 300, 305; regression, 250, 295; skewness, 151, 152, 156, 157; variation, 131
Column diagram (*see* Histogram), construction of, 37; defined, 37
Combination, 366
Compound interest law, 316
Confidence limits, 410
Continuous variate, 6
Coolidge, J. L., 379
Correlation, by ranks, 263–265; coefficient of, 237, 245, 246, 254, 265; definition of, 240; index, 357; multiple, 277–305; non-linear, 355; partial, 295; perfect, 239, 287; summary of, 247; table, 254; versus causation, 267
Craig, A. T., 391
Craig, C. C., 456
Crathorne, A. R., Preface
Crowder, W. F., 35, 201, 288
Croxton and Cowden, 453, 465
Curve fitting, by averages, 313; by least squares, 307, 314, 331; by moments, 307, 331; by selected points, 311; of exponential function, 316; of hyperbola, 333; of modified exponential, 334; of modified power function, 337; of normal curve, 413–417; of parabola, 330–333; of power function, 323; of straight line, 216, 222, 311, 314
Curves, cumulative, 50; exponential, 316; hyperbolic, 333; J-shaped, 52; mound-shaped, 44, 52; normal, 53,

134, 363–409; parabolic, 82, 330; skewed, 52, 152, 153, 383, 384
Czuber, Emanuel, 85

Data, statistical, 1; grouped, 23; ungrouped, 23
Davenport, D. H., 458
Davies, George R., 35, 201, 288
Decile, 119
Degrees of freedom, 453
Deming, W. E., 456
De Moivre, Abraham, 363, 396, 397
Dependent events, 376
Dependent variable, 6
Determinants, defined, 289, 290, 292; finding mode by, 110; in multiple correlation, 293–305
Deviation, mean, 121; probable, 119; quartile, 115; standard, of a difference, 444; standard, of a distribution, 125–130; standard, of the mean, 144, 430; standard, of the standard deviation, 144, 443
Differences, standard error of, 444–445
Differencing, process of, 307; use of, in curve-fitting, 309–310
Discrete variate, 6, 15
Dispersion, 43, 111; meaning of, 111–115; measures of, 113
Distribution of means, defined, 143–144; excess of, 441; illustrated, 139, 140, 426; mean of, 144, 429; probable error of, 144, 432; standard deviation of, 144, 430; standard error of, 144; skewness of, 439
Distribution of standard deviations, defined, 144, 442; mean of, 145; standard deviation of, 145, 442; standard error of, 144
Distributions, asymmetrical, 52; cumulative frequency, 48; J-shaped, 52; mound-shaped, 44, 52; normal, 53, 414–416; simple frequency, 25; symmetrical, 52; temporal, 44; U-shaped, 52

Empirical curves, defined, 210, 306; limitations of, 338; methods of fitting, Chapter 10
Empirical equation, 210, 306, 338
Empirical probability, 369
Equation of, exponential functions, 316; hyperbola, 333; hyperplane, 298; modified exponential, 334; modified power function, 337; normal curve, 119, 396, 400; plane, 278; power function, 323; quadratic parabola, 330; straight line, 210, 311
Error, absolute, 16; possible, 15; probable, 137; probable, of mean, 143, 432; probable, of standard deviation, 144, 443; relative, 16; standard, of estimate, 233, 239, 280, 286, 295, 300, 304
Excess (kurtosis), 43, 158, 389, 439
Expectation, 374
Exponential function, fitting data to, 316, 343; when to use with empirical data, 317–318
Ezekiel, Mordecai, 466

Factor reversal test, 199
Fisher's Ideal Index, 198
Fisher, Irving, 195, 198, 199, 200
Fisher, R. A., 364, 421, 447, 452, 453, 465
Forsyth, C. H., Preface
Freeman, H. A., 439
Frequency, class, 25, 422; cumulative, 48; relative, 372
Frequency curves, 40, 397; types of, 52, 418
Frequency distribution, binomial, 382–395; cumulative, 48; normal, 395–410, 413–417; simple, 25
Frequency polygon, 38
Frequency table, 4, 25
Function, 7

Gale, A. S., 271
Gauss, Carl Friedrich, 138, 363, 396
Gavett, G. I., 271
Geometric mean, computation of, 90; criticism of, 101; defined, 87; of relatives, 180–182, 191–193; use of, 88
Glover, J. W., Preface, 379, 467
Goodness of fit, tests for, 211, 213, 230, 232, 233, 286, 300, 304, 416
Graduation of a frequency distribution, by point binomial, 391–395; by normal curve, 413–417
Graphical representation, 37, 340–350; of cumulative distributions, 50; of simple frequency distributions, 37, 38, 39; of temporal distributions, 43–48; with logarithmic paper, 346; with semi-logarithmic paper, 342
Growth, law of organic, 316

Harmonic mean, computation of, 93; defined, 92; of relatives, 181, 188; uses of, 93, 95–97, 181, 198
Hall, Winfield S., 171
Haskell, S. C., 46, 466
Histogram, 37

Holzinger, Karl, 465
Hotelling, Harold, 153
Huntington, E. V., 5
Hyperbola, 333

Independent, events, 375; variable, 6
Index numbers, 174–202; as average of relatives, 180–182, 188–193; bias in, 197–198; defined, 174, 177; Fisher's Ideal, 198; purpose of, 174; unweighted, 178–184; weighted, 185–200
Index of precision, defined, 399; use of, 433
Interpretation of statistical results, 3, 364, 410, Chapter 13
Interval, class, 25, 26, 30

Jackson, Dunham, 220, 456

Karsten, K. G., 466
Kendall, M. G., 1, 59, 145, 465
Kenney, J. F., 465
Kepler, Johann, 339
King, W. I., 49, 81, 200
Kurtosis, 43
Kurtz, Edwin B., 172

Laplace, Pierre Simon, 363, 396
Least squares, principle of, 211; fitting a parabola by, 330; fitting a straight line by, 214–222, 314
Lee, Alice, 463
Levels of significance, 410
Linear trends, 203
Lines of regression, 218, 248–249
Lipka, Joseph, 466
Logarithmic paper, 346

May, Mark, 301
Mean, arithmetic, 62; geometric, 87; harmonic, 92
Mean deviation, computation of, 122; defined, 121
Median, 49, 76; computation of, 51, 78, 79; defined, 49, 76
Mill, J. S., 269
Mills, F. C., 34, 188, 357, 458, 466
Mitchell, Wesley C., 200
Modal class, 81
Mode, 80; approximate, 80, 81, 84, 85; criticism of, 100; crude, 80; true, 80
Modified exponential function, fitting data to, 333; when to use with empirical data, 333
Modified power function, fitting data to, 337; when to use with empirical data, 337
Moment, arithmetic mean as, 62–66
Moments, adjusted, of a distribution, 163; computation of, 164; method of, in curve-fitting, 160; unadjusted, of a distribution, 159; of point binomial, 387–389; of normal curve, 405
Multiple correlation, coefficient of, 281, 287; defined, 277, 288
Mutually exclusive events, 374

Normal curve, defined by equation, 53, 119, 396; derivation of equation to, 397; graduation of distribution by, 413–417; history of, 363, 396; moments of, 405; properties of, 401; uses of, 134, 397, 405–409
Normal equations, 215, 217, 278, 283, 298, 303
Null hypothesis, 447
Numerical value of a number, 121

Ogive, 48
Organic growth, law of, 316
Organization of data, 1, 5

Parent population (universe), 3, 23, 363, 419, 451
Parkes, A. S., and Drummond, J. C., 450
Pearl, Raymond, 105, 301, 466
Pearson, E. S., 439
Pearson, Karl, 85, 151, 237, 391, 416, 419, 456, 463, 467
Percentiles, 119
Permutation, 364
Point binomial, 383–395
Power function, fitting data to, 323, 347; when to use with empirical data, 324
Precision, index of, 399, 433
Preliminary sheet, 25, 253
Probability, 369; a priori, 372; empirical, 369; theorems on, 374–378
Probable deviation, 119
Probable error, 118, 137, 403; defined, 137, 403; of any measure, 137, 403; of the arithmetic mean, 143, 432; of the standard deviation, 146, 443

Quadratic parabola, fitting data to, 330; in finding approximate mode, 83; when to use with empirical data, 330
Quartile deviation, 115
Quartiles, computation of, 117, 120; defined, 115; in measuring dispersion, 118; in measuring skewness, 156
Quetelet, Adolphe, 363

Range, 24, 113, 114
Regression, coefficients of, 250, 295; of Y on X, 250; of X on Y, 250; multiple, 295; plane, 278; hyperplane, 299
Relatives, defined, 174; chain, 176; fixed base, 176; link, 176; simple aggregative, 179; simple arithmetic mean of, 181; simple geometric mean of, 181; simple harmonic mean of, 181; weighted aggregative, 185; weighted arithmetic mean of, 188; weighted geometric mean of, 191; weighted harmonic mean of, 188
Relative error, defined, 16; in a product, 19; in a quotient, 20
Relative frequency, 369, 424
Relative variability, 113, 118, 122, 131
Reliability, defined, 143, 421; of a difference, 444–445; of the mean, 143, 453; of the standard deviation, 145, 453
Repeated trials, theorem of, 378
Residual, 210
Riebesell, Paul, 42
Rietz, H. L., 164, 416, 424, 456, 465, 466
Robinson, George, 168
Rounding off numbers, 16
Running, T. R., 310, 466

Sample, 3, 23, 363, 419; small, 450
Scarborough, J. B., 18, 466
Scatter diagram, 234, 253
Secrist, Horace, 53, 250
Secular trend, 226
Selection (*see* Combination), 366
Semi-logarithmic paper, 342
Sheppard's Corrections, 163–167
Shewhart, W. A., 456
Significant difference, 409, 443–448
Significant figures, 15
Simple frequency distribution, 25
Skewness, 150–157; defined, 43, 150; measurement of, 151–157
Slope of a straight line, 205
Snedecor, G. W., 466
Solomons, Leonard M., 153
Sorenson, Herbert, 35, 296, 438, 466
Standard deviation, computation of, 126, 127, 129, 130; defined, 125; in class frequencies, 125, 423; of the mean, 144, 430, 453; of a percentage, 425; of the standard deviation, 144–146, 443, 453
Standard error, of estimate, 233, 280; of the mean, 144; of the standard deviation, 144
Standard unit, 162, 250, 400
Statistical, analysis, 2; constant, 3; data, 1; induction, 364, 420; inference, 364; methods, 1
Stirling's Formula, 379
Straight line, fitting observed data to, 210, 311–315; intercepts of, 207; properties of, 206; slope of, 205
Summation, defined, 7; limits of, 8; theorems on, 9
Surface, F. M., 105, 301
Symmetrical distributions, 52

Tabular presentation, 23, 25
Tabulation of data, 23, 53
Tallying, 25, 253
Temporal distribution, 44, 226
Tests of significance, 409, 446
Thurstone, L. L., 34, 411
Time reversal test, 197
Time series, 43; fitting a straight line to, 226, 342
Tippett, L. H. C., 449, 465
Treloar, Alan E., 364, 462, 465
Trend, linear, 203; non-linear, 306
True class limits, 26
Tycho Brahe, 339
Tyler, R. W., Preface

Unit, class, 70, 71, 72, 128, 161, 245; standard, 162
Universe, 3, 23, 363, 419
Unweighted index numbers, 179–184

Variability, absolute, 113; relative, 118, 122, 131–133
Variable, dependent, 6; independent, 6
Variance, 126, 442
Variates, 6; continuous, 6; discrete, 6
Variation, coefficient of, 131

Walker, H. M., 1
Walsh, C. M., 200
Watkeys, C. W., 271
Waugh, Albert E., 412, 465
Weighted aggregative relative, 185
Weighted averages, 188–193
Weighted index numbers, 188, 191. 195–199
Weighted mean, 67, 90
Wembridge, H. A., 148
White, R. C., 34
Whittaker, E. T., 168
Winfrey, Robley, 172
Wolfenden, H. H., 466

Yoder, Dale, 35
Yule, G. U., 1, 59, 145, 465