

Sampling Statistics and Applications

FUNDAMENTALS OF THE THEORY OF STATISTICS

BY

JAMES G. SMITH

AND

ACHESON J. DUNCAN

*Department of Economics and Social Institutions,
Princeton University*

FIRST EDITION

McGRAW-HILL BOOK COMPANY, INC.

NEW YORK AND LONDON

1945

B. L. No. 64.

Imperial Veterinary Research Institute
Library,
" MUKTESWAR.

Class. *slatko*

Register No. *4980*

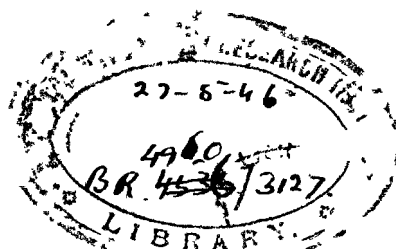
Room No.

Inward No. *3127*

Shelf No.

Received. *27-5-46.*

Book No.



SAMPLING STATISTICS AND APPLICATIONS

*This book is produced in full compliance
with the government's regulations for con-
serving paper and other essential materials.*

FUNDAMENTALS OF THE
THEORY OF STATISTICS

by JAMES G. SMITH
and ACHESON J. DUNCAN

*ELEMENTARY STATISTICS AND APPLICATIONS

**SAMPLING STATISTICS AND APPLICATIONS

SAMPLING STATISTICS AND APPLICATIONS

COPYRIGHT, 1945, BY THE
MCGRAW-HILL BOOK COMPANY, INC.

PRINTED IN THE UNITED STATES OF AMERICA

*All rights reserved. This book, or
parts thereof, may not be reproduced
in any form without permission of
the publishers.*

THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

Sampling Statistics and Applications is the second volume of *Fundamentals of the Theory of Statistics*; this second volume is intended for advanced students or research workers. The first volume is entitled *Elementary Statistics and Applications* and is designed for beginning courses in statistics.

After reviewing the basic concepts and definitions in *Sampling Statistics and Applications*, the authors discuss the general theory of frequency curves and the theory of random sampling. Important sampling distributions are derived, and their applications to a variety of problems are illustrated. Exact methods applicable primarily to normal populations and approximate methods used in sampling from discrete populations and from continuous non-normal populations are considered. Theoretical discussion is illustrated throughout to show real-life applications. The character of assumptions involved in theory is explicitly treated; the problems that such assumptions present when the statistician is confronted with practical applications are illustrated.

The arrangement of the material in the book is based partly on grounds of logic and partly on practical considerations. The theory of frequency curves is general in scope and is therefore presented first. The theory of sampling, being an elaboration of a special part of the theory of frequency curves, is discussed after the more general theory. Within the discussion of the theory of sampling, the more elementary aspects are examined before approaching complex problems. This leads to a separation of some material that might logically be run together. It gives the instructor greater freedom, however, in the selection of material appropriate for the level of his class, while the arrangement is such that he can assign material in logical sequence if he so desires.

For the most part theory and application have been discussed together in this volume. In certain cases, however, in which theoretical discussion is especially elaborate, its applications to

practical problems are treated separately. Numerical calculations for frequency curves, for example, are discussed in a separate chapter following the chapters devoted to the theory of frequency curves. Again, the uses of the sampling distributions of the mean, standard deviation, etc., are discussed in a chapter separate from that in which the theoretical derivation of these distributions is presented. On the whole, the arrangement of material is designed to facilitate the use of the book in the school or research laboratory as well as in the classroom.

Progress in sampling theory and its application has been rapid during the past 25 years. Much has been accomplished in deriving exact sampling distributions for important statistics and in clarifying the assumptions on which statistical analysis rests. Discussion has also been active regarding the concept of probability likely to be most fruitful for statistical research, and increasing attention has been paid to the logic of statistical inference. This development of theory has been accompanied by progress in method and application. The discovery of more exact sampling distributions, for example, has led to greater emphasis upon careful design of experiments and less emphasis upon size of sample. Quality of method has, in many instances, displaced reliance only on quantity of numbers.

Sampling Statistics and Applications represents an attempt to coordinate the new theory and applications with the old, to place the whole upon a logically consistent basis, and to put the subdivisions in proper relation to each other. To this end a concept of probability that at present appears most fruitful for statistical research was adopted, and the theory of sampling is explained in those terms. Elaboration of special techniques regarding the design of experiment and analysis of variance is not included since the primary intent is to emphasize fundamentals of the theory of statistics. Explanations both of old and of new methods start with the fundamentals; and, unless otherwise indicated, a particular exposition includes all the argument. If a highly mathematical step of an argument is omitted or abbreviated, this is so stated. Some of the mathematical portions are included in the text; other more advanced mathematical material is placed in appendixes to some of the chapters. These mathematical parts are presented in such a way as to be readily teachable.

The authors have drawn freely upon the many monographs and the periodical literature that have appeared during recent years. Care has been exercised to make acknowledgment in footnotes to the sources of new ideas that have been incorporated into the authors' own development of the subject. To all these vigorous workers in the field, too numerous to be listed by name, the authors are greatly indebted. The authors especially acknowledge a debt of gratitude to John H. Smith of the Bureau of Labor Statistics, who contributed many stimulating criticisms and suggestions that inspired important improvements in this book. The authors are grateful to the International Finance Section of Princeton University for the financial assistance that was given Acheson J. Duncan some years ago to study statistics and mathematical economics with the late Henry Schultz of the University of Chicago and with Harold Hotelling of Columbia University. The authors are indebted also to those men for help and guidance in the study of statistics.

Naturally it is not to be supposed that the whole or any part of the manuscript carries the endorsement of the authors' former teachers or those who have helped with criticisms of the manuscript. The authors assume full responsibility for any errors of theory or calculations that may be present in the volume.

They are indebted to Prof. R. A. Fisher, also to Oliver & Boyd, Ltd., of Edinburgh, for permission to reprint Tables III and IV from their book *Statistical Methods for Research Workers*. As indicated in notes making specific acknowledgment in the text and in the Appendix, the authors are also indebted to others for reproductions or abridgments of tables of several other sampling distributions.

JAMES G. SMITH,
ACHESON J. DUNCAN.

PRINCETON, N.J.,
March, 1945.

CONTENTS

	Page
PREFACE.	v
TABLE OF SYMBOLS.	xi

INTRODUCTION

CHAPTER

I. DEFINITIONS AND BASIC CONCEPTS.	1

PART I

GENERAL THEORY OF FREQUENCY CURVES

II. PROBABILITY AND THE PROBABILITY CALCULUS.	23
III. THE SYMMETRICAL BINOMIAL DISTRIBUTION AND THE NORMAL CURVE.	32
IV. THE PEARSONIAN SYSTEM OF FREQUENCY CURVES.	40
V. THE GRAM-CHARLIER SYSTEM OF FREQUENCY CURVES	82
VI. SUMMARY OF THE THEORY OF FREQUENCY CURVES, AND SOME EXAMPLES	100
VII. NUMERICAL CALCULATIONS FOR FREQUENCY CURVES. . . .	115

PART II

ELEMENTARY THEORY OF RANDOM SAMPLING

VIII. A PREVIEW OF SAMPLING THEORY.	153
IX. SAMPLING FROM A DISCRETE TWOFOLD POPULATION	186
X. SAMPLING FROM CONTINUOUS NORMAL POPULATIONS: I. VARIOUS SAMPLING DISTRIBUTIONS.	219
XI. SAMPLING FROM CONTINUOUS NORMAL POPULATIONS: II. USES OF THE SAMPLING DISTRIBUTIONS	267
XII. SAMPLING FLUCTUATIONS IN CORRELATION STATISTICS . . .	298

PART III

ADVANCED SAMPLING PROBLEMS

XIII. SAMPLING FROM A DISCRETE MANIFOLD POPULATION	308
XIV. JOINT SAMPLING FLUCTUATIONS IN MEAN AND STANDARD DEVIATION	340
XV. SAMPLING FLUCTUATIONS IN REGRESSION STATISTICS	372
XVI. PROBLEMS INVOLVING TWO SAMPLES.	391
XVII. ANALYSIS OF VARIANCE.	422
XVIII. THE PROBLEM OF NONNORMALITY	445

APPENDIX

	Page
TABLE I. FOUR-PLACE COMMON LOGARITHMS OF NUMBERS	457
TABLE II. SQUARES OF NUMBERS	461
TABLE III. SQUARE ROOTS OF NUMBERS FROM 10 TO 100	463
TABLE IV. SQUARE ROOTS OF NUMBERS FROM 100 TO 1000	465
TABLE V. RECIPROCAL OF NUMBERS.	467
TABLE VI. AREAS, ORDINATES, AND DERIVATIVES OF THE NORMAL CURVE.	469
TABLE VII. TABLE OF t	474
TABLE VIII. TABLE OF χ^2	475
TABLE IX. TABLE OF F	476
TABLE X. 5 PER CENT AND 10 PER CENT POINTS OF THE SAMPLING DISTRIBUTION OF $\sqrt{\beta_1}$ FOR SAMPLES OF VARIOUS SIZES	480
TABLE XI. LOWER AND UPPER 1 PER CENT AND 5 PER CENT POINTS OF THE SAMPLING DISTRIBUTION OF β_2 FOR SAMPLES OF VARIOUS SIZES	480
TABLE XII. LOWER AND UPPER 1 PER CENT, 5 PER CENT, AND 10 PER CENT POINTS OF SAMPLING DISTRIBUTION OF $a = A.D./\sigma$ FOR SAM- PLES OF VARIOUS SIZES	481
TABLE XIII. SAMPLING DISTRIBUTION OF THE RANGE $w = \frac{X_n - X_1}{\sigma}$	482
TABLE XIV. HYPERBOLIC TANGENTS.	483
INDEX.	485

TABLE OF SYMBOLS

Characteristics	Symbols for	
	Sample statistics	Population parameters
Arithmetic mean.....	$\bar{X}$	$\bar{X}$
Geometric mean.....	G.M.	G.M.
Harmonic mean.....	H.M.	H.M.
Median.....	Mi	Mi
Mode.....	Mo	Mo
Standard deviation:		
Zero order.....	σ_i	σ_i
Higher order.....	$\sigma_{i,jk} \dots$	$\sigma_{i,jk} \dots$
Average deviation.....	A.D.	A.D.
Quartiles:		
First.....	Q_1	Q_1
Third.....	Q_3	Q_3
Range:		
Absolute.....	R	R
Relative.....	w	w
Moments:		
About arbitrary origin.....	ν_1	ν_1
	ν_2	ν_2
	ν_3	ν_3
	.	.
	.	.
	.	.
	ν_n	ν_n
About arithmetic mean.....	μ_1	μ_1
	μ_2	μ_2
	μ_3	μ_3
	.	.
	.	.
	.	.
	μ_n	μ_n
Beta coefficients.....	β_1	β_1
	β_2	β_2
k statistics.....	k_1	k_1
(population parameters are called	k_2	k_2

Characteristics	Symbols for	
	Sample statistics	Population parameters
“cumulants,” which are related to the moments as follows: $k_1 = \mu_1, k_2 = \mu_2, k_3 = \mu_3, k_4 = \mu_4 - 3\mu_2^2, \dots$	k_3 . . . k_n	
g statistics.....	g_1	
($g_1 = \beta_1, g_2 = \beta_2 - 3$)	g_2	
Correlation coefficients:		
Simple.....	r_{ij}	
Partial.....	$r_{ij.k} \dots$	
Multiple.....	$R_{i.jk} \dots$	
Regression values:		
Constants.....	$a_{i.jk} \dots$	
Coefficients.....	$b_{ij.k} \dots$ $\beta_{ij.k} \dots$	
A point on a line or on a plane of regression.....	X'_i	
A percentage.....	p_i	
Standard errors:		
Of any specified statistic, theta (θ); thus, for example.....		
Of the arithmetic mean.....		
Of the variance.....		
Of a higher-order variance.....		
Of a regression constant.....		
Of a regression coefficient.....		
Of a percentage.....		
Etc.		

The breve above a symbol means that it represents the maximum likelihood estimate of the corresponding population parameter. For example, $\bar{b}_{ij.k}$ means the maximum likelihood estimate of $b_{ij.k}$; $\bar{\sigma}_x$ means the maximum likelihood estimate of σ_x ; etc.

INTRODUCTION

CHAPTER I

DEFINITIONS AND BASIC CONCEPTS

Frequency Distributions. Time-honored observation has made commonplace the eighteenth-century discovery that static variability in almost any conceivable attribute of an object, event, or condition follows a surprisingly uniform pattern.¹ The uniform pattern of static variability is revealed by arranging the variable in a frequency distribution or frequency series, which is simply a summary of an array of the variable arranged from smallest to largest. It is also called a "monovariate." When graphed the figure is called a "histogram," "frequency polygon," or "frequency curve," depending upon the manner of graphing. Two types of frequency series are to be carefully distinguished, discrete frequency series and continuous frequency series.

Discrete Frequency Series. A discrete frequency series is one in which the variation, by the nature of the variable, is in distinct steps. The variation in size of men's shoes occurs by distinct steps, not by infinitesimal differences. Accordingly, a frequency series describing the number of men's shoes of various sizes is a discrete frequency series. The same statement could be made with respect to almost any item of clothing manufactured in sizes for mass consumption.

The graph of a discrete frequency series should be in the form of a histogram, not a frequency curve; for the latter would suggest continuous variability, which is contrary to fact in the case of a discrete series.

Continuous Frequency Series. A continuous series is one representing a phenomenon that varies by infinitesimal amounts.

¹ Static variability refers to variability which is not correlated with time or in which time is "held constant." Variability correlated with time, *i.e.*, dynamic variability, may assume a great variety of patterns. See SMITH, J. G., and A. J. DUNCAN, *Elementary Statistics and Applications*, Chaps. V and XIX.

For example, the frequency distribution of weights or heights of people of some specified age is continuous in character; for the differences among a large number of people in weight or in height are in fact by infinitesimal amounts. A continuous series may have the appearance in a frequency table of the same discreteness as a discrete series; but this is because the arbitrarily discrete

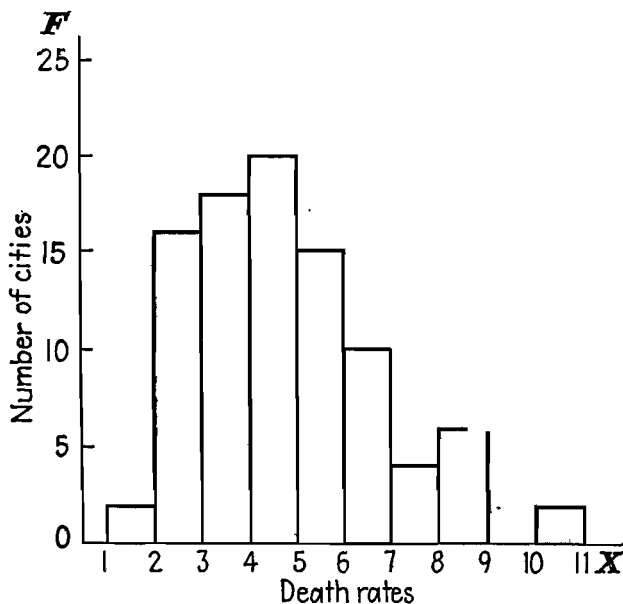


FIG. 1.—Maternal mortality in cities of 100,000 or more population in the United States in 1938. (Deaths per 1,000 live births.)

character of the unit of measurement eclipses the actual continuous character of the data.

The graph of a continuous-frequency series may appropriately be in the form of a frequency curve, but often when the number of cases observed is comparatively few a continuous frequency series is pictured by a histogram.

Histograms and Frequency Curves. Figure 1 is an illustration of a histogram; it is a graph of the frequency distribution presented in the first two columns of Table 1. In this figure the frequency of any class interval is represented by a rectangle erected on that interval as a base and with a height equal to the observed frequency.

For many types of analysis it is preferable to present the frequency distribution and the corresponding histogram in a form in which the area of a rectangle represents the proportional frequency of an interval. Figure 2 is an illustration of such a histogram; it is a graph of the third column of Table 1, with the same scale variation as that used in Fig. 1, namely, that shown in column (1) of Table 1.

TABLE 1.—MATERNAL MORTALITY IN CITIES OF 100,000 OR MORE POPULATION IN THE UNITED STATES IN 1938*

(1)	(2)	(3)
Deaths per 1,000 live births	Number of cities	Proportional number of cities
X	F	$\frac{F}{N}$
1-	2	.022
2-	16	.172
3-	18	.193
4-	20	.215
5-	15	.161
6-	10	.108
7-	4	.043
8-	6	.064
9-	0	.000
10-	2	.022
	$\Sigma = 93$	$\Sigma = 1.000$

In histograms of the type shown in Fig. 2, the total area of the histogram always equals 1, and each of the bars is a portion of 1.

Now suppose that the data from which the histogram has been constructed were a sample from a very large set of cases, theoretically an infinite set. For example, the data might be the heights of 100 adult males of the white race, instead of the mortality statistics above illustrated. The 100 heights, then, would be a sample of the heights of all adult men of that race, presumably millions of men. If the size of the sample were increased, the class interval could be reduced without causing irregularities in the form of the plotted histogram. In fact, if the number in the sample is made larger and larger and at the same time the size of the class interval is continuously reduced, the histogram will tend to become more and more regular and

the tops of the rectangles, which are getting narrower and narrower, will come closer and closer to forming a smooth continuous curve (a frequency curve).

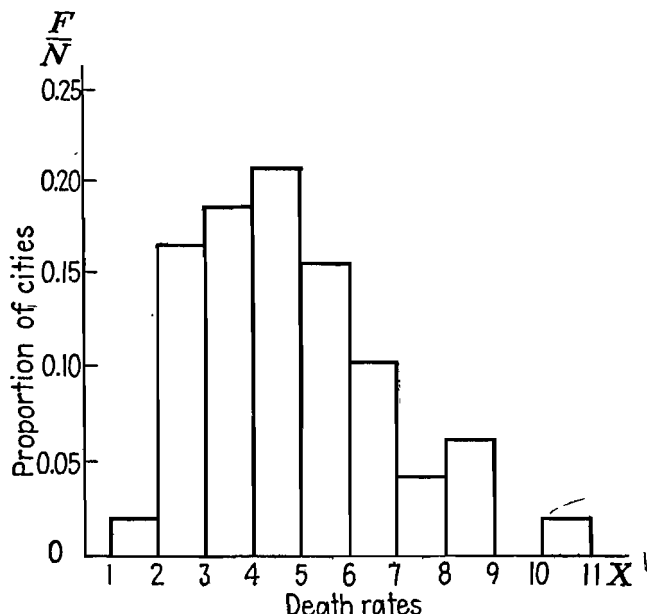


FIG. 2.—Maternal mortality in cities of 100,000 or more population in the United States in 1938. (Deaths per 1,000 live births.)

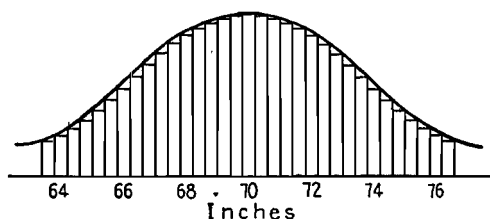


FIG. 3.—A frequency curve.

In such a manner, the frequency curve may be viewed as the limit that an area histogram of proportional frequencies approaches as the number of cases is increased and the size of the class interval is reduced indefinitely. The frequency curve depicts the distribution of a theoretically infinite set of data, with a theoretically infinitesimal class interval. Figure 3 illustrates a frequency curve.

Being the limit approached by an area histogram of proportional frequencies, the frequency curve has a total area between the curve and the X -axis that is always equal to 1. Furthermore, any section of area under the curve will give the relative frequency of the cases falling within the class interval marking off that section of area.

Populations, Parameters, and Statistics. To say that the population of the United States is some hundred and thirty⁴ million people is a familiar use of the word "population." In statistics the word is used in the same familiar sense, but it is also used in a more general sense. In statistics the term "population" (or sometimes "universe") refers to the enumeration of persons or animals of any kind or even to the enumeration of inanimate things. In statistics, population refers to all the objects of a defined kind in existence in some specified universe; for example, all the adult males in the United States constitute a population, as do all the automobiles in the United States or all the three-year-old steers in the United States.

The measurements of characteristics of a population are called "parameters." The average height of all adult males in the United States is a parameter of that population. The average weight of all the three-year-old steers in the United States is a parameter of that population. Rarely are the parameters of any of these populations actually measured. In practice, it is much easier and more practical to estimate the parameter by taking the average or measuring the corresponding characteristic of a sample from the population in question. This latter measure is called a "statistic." Thus parameters are the characteristics of the population, and the corresponding sample statistics are the measures of the corresponding characteristics of samples.

Averages. In addition to the familiar arithmetic mean, other averages are used as devices for summarizing or comparing. The median and mode are often used in statistical analysis; less frequently, but necessarily for certain purposes,¹ the geometric mean and the harmonic mean are used.

By definition the arithmetic mean is the sum of the variables divided by the number of variables, *i.e.*,

$$\bar{X} = \frac{\sum X}{N} \quad (1)$$

¹ Cf. SMITH and DUNCAN, *op. cit.*, pp. 173-179.

When the variable is arranged in a frequency distribution, this equation for the mean generally appears as

$$\bar{X} = \frac{\Sigma FX}{N} \quad (1')$$

to represent the fact that the several class-interval values of X have their respective frequencies, so that

$$\Sigma FX = F_1X_1 + F_2X_2 + \dots + F_nX_n.$$

It is important to note that, when the frequency distribution is expressed proportionally, as in the third column of Table 1, it becomes a distribution of probability.¹ The arithmetic mean of a distribution of probability, *i.e.*, of a proportional frequency distribution, is equal to $\sum \frac{F}{N} X$. Expressed in the symbols of probability, this equation is

$$\bar{X} = \Sigma P(X)X \quad (2)$$

Table 2 illustrates the calculation of the arithmetic mean of the frequency distribution shown in Table 1.

TABLE 2.—CALCULATION OF THE ARITHMETIC-MEAN MATERNAL MORTALITY IN CITIES OF 100,000 OR MORE IN 1938

Deaths per 1,000 live births		F	FX	$P(X)$	$\frac{F}{N} X$ or $P(X)X$
X	Mid-point of class interval				
1-	1.5	2	3.0	.022	.033
2-	2.5	16	40.0	.172	.430
3-	3.5	18	63.0	.193	.676
4-	4.5	20	90.0	.215	.968
5-	5.5	15	82.5	.161	.886
6-	6.5	10	65.0	.108	.702
7-	7.5	4	30.0	.043	.322
8-	8.5	6	51.0	.064	.544
9-	9.5	0		.000	
10-	10.5	2	21.0	.022	.231
		93	445.5	1.000	4.792

¹See *ibid.*, p. 254.

Accordingly,

$$\bar{X} = \frac{\Sigma FX}{N} = \frac{445.5}{93} = 4.790$$

or

$$\bar{X} = \Sigma P(X)X = 4.792$$

The sum of the column of FX 's must be divided by N ($= 93$) in order to find the mean, whereas the sum of the column of $P(X)X$'s is the mean. It will be noted that taking the mid-point of each class interval as the X , respectively, of the various class intervals is equivalent to assuming that the frequencies within each class interval are so distributed that their average is equal to the mid-point. This assumption causes negligible error in the calculation of the mean.¹

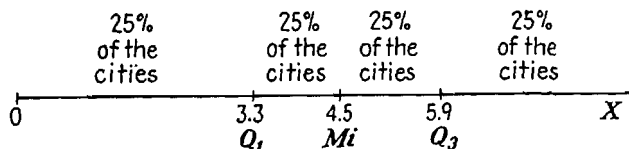


FIG. 4.—Median and quartiles of mortality rates in selected cities of the United States.

The median is a position average; by definition, the median is that value than which there is an equal number of cases larger and smaller. When the cases are arranged in an array, the median is either the value of the middle one (when there is an odd number of cases) or some value between the two middle ones (when there is an even number of cases). Ordinarily in the latter instance the arithmetic mean of the two middle cases is taken as the median value. The symbol for the median is Mi .

The first quartile, Q_1 , is the value below which one fourth of the cases fall and above which three fourths of the cases fall. The third quartile, Q_3 , is that value below which three fourths of the cases fall and above which one fourth of the cases fall. The median, obviously, is equivalent to the second quartile. If the cases are arranged in an array, it should be apparent from the definitions of the median and the quartiles that these three values divide the cases into four parts, with one fourth of the cases in each part. This is illustrated graphically in Fig. 4.

¹ When the standard deviation is so calculated, as it usually is, it is necessary to correct for error due to the assumption made; the correction is called "Sheppard's correction." Cf. pp. 10, 12.

For the frequency distribution shown in Tables 1 and 2, the median is found by interpolating the value of the middle ($46\frac{1}{2}$) case, to be 4.53; Q_1 is found by interpolating the $N/4$ case to be 3.3; and Q_3 is found by interpolating the $3N/4$ case to be 5.9. As shown in Fig. 4, the significance of these three values is that, in 1938, 25 per cent of the cities of 100,000 or more population had a maternal death rate of less than 3.3 per 1,000 live births; 25 per cent of the cities had maternal death rates between 3.3 and 4.5 per 1,000 live births; 25 per cent had maternal death rates between 4.5 and 5.9 per 1,000 live births; and the remaining 25 per cent of the cities had maternal death rates greater than 5.9 per 1,000 live births.

If it were desired to obtain a more detailed description of the distribution of these cities with respect to maternal death rates, the distribution could be divided in 10 parts each containing 10 per cent of the cities. The dividing points of these 10 parts would be called "deciles," instead of quartiles; there would be 9 deciles, and the fifth would be the same value as the median. Similarly, if with a larger number of cases it were desired to obtain a description of the distribution sufficiently refined to say within what limits lie each 1 per cent of the group of cities, the distribution could be divided in 100 parts each containing 1 per cent of the cities. The dividing points of these 100 parts would be called "percentiles"; and there would be 99 percentiles, of which the fiftieth would be the same value as the median.

Another commonly used average, the mode, is described in terms of relative frequency of occurrence. It is the magnitude that occurs more frequently than any other. The mode is the most probable value. When the data are presented in a frequency distribution, it is necessary to interpolate for the mode. The procedure may be illustrated by finding the mode in the frequency distribution shown in Tables 1 and 2. It may be assumed that the mode lies somewhere between 4 and 5, for more cases (20) lie in that class interval than in any other of the class intervals. The mode is equal to the lower limit of the modal class interval plus the interpolated part of the class interval established by the relationship of the frequencies above and below that class interval. Thirty-six frequencies lie below the modal class interval, and 37 frequencies lie above the modal class interval. Hence the mode is equal to $4 + \frac{37}{37 + 36} (1) = 4.5$.

The geometric mean is the n th root of the product of n variables X . Thus, the geometric mean of 5, 8, and 25 is the cube root of $5 \times 8 \times 25 = 100$. The geometric mean may also be defined as the antilogarithm of the arithmetic mean of the logarithms of the variables X , i.e.,

$$\log \text{G.M.} = \frac{\sum \log X}{N} \quad (3)$$

The harmonic mean is the reciprocal of the average of the reciprocals of a variable magnitude $X_1, X_2, \dots, X_n$, thus:

$$\text{H.M.} = \frac{N}{\sum \frac{1}{X}} \quad (4)$$

Moments. In statistics the term "moment" has been taken over from physics. In physics, moment is a measure of a force with respect to its tendency to produce rotation. The strength of the tendency depends on the amount of force and the distance from the origin of the point at which the force is applied. If the arithmetic mean is taken as the origin in a frequency distribution and the frequencies in each class interval ($F_1, F_2, \dots, F_n$) are taken as the forces, at distances $x_1, x_2, \dots, x_n$, the moments (more exactly moment coefficients since the physical moments are divided by N) are defined as follows:

$$\left. \begin{aligned} \mu_1 &= \frac{\sum Fx}{N} \\ \mu_2 &= \frac{\sum Fx^2}{N} \\ \mu_3 &= \frac{\sum Fx^3}{N} \\ &\vdots \\ \mu_n &= \frac{\sum Fx^n}{N} \end{aligned} \right\} \quad (5)$$

in which $x = X - \bar{X}$.

The Greek letter mu (μ) is used to describe the moments about the arithmetic mean. When the deviations are measured about an arbitrary origin, instead of about the arithmetic mean, the symbol used to represent the moments is the Greek letter nu (ν). Generally speaking, it is more convenient first to calculate the moments about an arbitrary origin and by the use

of the following equations to obtain the moments about the mean:

$$\left. \begin{aligned} \mu_1 &= \nu_1 - \nu_1 = 0 \\ \mu_2 &= \nu_2 - \nu_1^2 \\ \mu_3 &= \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 \\ \mu_4 &= \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 \end{aligned} \right\} \quad (6)$$

By the use of Sheppard's correction, μ_2 and μ_4 must be adjusted for errors involved in the use of mid-points of class intervals as the X 's in the process of calculation.¹ When it is desired to use a symbol for a population parameter in the present text, μ is printed in boldfaced type (**μ**). For the most part, statisticians deal with statistics and speculate about parameters. Seldom is it possible to specify the value of a parameter. But it is often possible to make a maximum likelihood estimate of a population parameter; the symbol for this is the symbol for the corresponding statistic, with a breve ($\smile$) above it. Thus the maximum likelihood estimates of the population moments are represented by $\check{\mu}_1, \check{\mu}_2, \dots, \check{\mu}_n$.

Types of Frequency Distribution. The moments are important because it is possible to calculate from them precise measures that will distinguish types of frequency distributions. The measures used by Karl Pearson to distinguish types of frequency distributions are the moments and functions of the moments.² Certain functions of the moments derived from them are called "betas." The first two are defined as follows:

$$\left. \begin{aligned} \beta_1 &= \frac{\mu_3^2}{\mu_2^3} \\ \beta_2 &= \frac{\mu_4}{\mu_2^2} \end{aligned} \right\} \quad (7)$$

If a frequency distribution is normal, $\beta_1 = 0$ and $\beta_2 = 3$. The degree to which the betas depart from these values is therefore a precise measure of the degree to which a frequency curve is not of the normal type.

¹ Cf. pp. 12, 132-134. It should be noted that Sheppard's correction is for the purpose of removing a bias; it does not correct for inaccuracies brought about by using class intervals that are too large. GOULDEN, C. H., *Methods of Statistical Analysis* (1939), p. 15.

² Cf. pp. 134-137.

The value of β_1 is related to the skewness of a curve, *i.e.*, whether or not the mode of the distribution is greater or less than the mean. If the mode is less than the mean, μ_3 is a plus quantity and the distribution is positively skewed; if the mode is greater than the mean, μ_3 is a minus quantity and the distribution is negatively skewed. The square root of β_1 , with the sign of the third moment, is often taken as a measure of skewness.¹

The value of β_2 is related to whether the frequency curve is flat topped or peaked. When it is flat topped, the shoulders of the curve are filled out and the tails depleted. When peaked, the frequency curve is higher at the center and the tails are also higher. The broad-shouldered frequency curve is said to be "platykurtic," and the narrow, or peaked, one is called "leptokurtic." The statistic β_2 is said, therefore, "to measure kurtosis." If β_2 is more than 3, the frequency curve is a peaked one; if β_2 is less than 3, the frequency curve is a broad-shouldered one. For the normal curve, $\beta_2 = 3$.

Because of their theoretical advantages R. A. Fisher² has suggested the use of certain k and g statistics instead of the moment and β statistics. These are defined as follows:³

$$\left. \begin{aligned} k_1 &= \frac{\Sigma x}{N} \\ k_2 &= \frac{\Sigma x^2}{N-1} \\ k_3 &= \frac{N}{(N-1)(N-2)} \Sigma x^3 \\ k_4 &= \frac{N}{(N-1)(N-2)(N-3)} \left[(N+1) \Sigma x^4 - 3 \frac{N-1}{N} (\Sigma x^2)^2 \right] \end{aligned} \right\} \quad (8)$$

and

$$\left. \begin{aligned} g_1 &= \frac{k_3}{\sqrt{k_2^3}} \\ g_2 &= \frac{k_4}{k_2^2} \end{aligned} \right\} \quad (9)$$

¹ For more complete discussion of skewness and of the normal frequency curve, see Smith and Duncan, *op. cit.*, pp. 226-230, 263-267, 285-306.

² *Statistical Methods for Research Workers*, Appendix B to Chap. III.

³ Cf. FISHER, R. A. *Statistical Methods for Research Workers*, Chap. III, Appendix on Technical Notation and Formulae (1932), p. 74; and GOULDEN, C. H., *op. cit.*, p. 29.

For large values of N , k_1 , k_2 , and k_3 are practically the same as μ_1 , μ_2 , and μ_3 ; and k_4 equals approximately $\mu_4 - 3\mu_2^2$.

It is necessary to apply Sheppard's correction to the second and fourth k statistics, as follows:

$$\begin{aligned}k_2 &= k_2 - \frac{1}{12} \\k_4 &= k_4 + \frac{1}{120}\end{aligned}$$

Also, for large values of N , g_1 is practically equal to $\sqrt{\beta_1}$, and g_2 equals approximately $\beta_2 - 3$. Thus, it is readily seen that g_1 relates to the measurement of skewness and g_2 relates to the measurement of kurtosis.

Standard Deviation and the Variance. From the second moment about the arithmetic mean a measure of the dispersion of the frequency distribution is obtained. The second moment itself is called the "variance." The square root of the second moment is called the "standard deviation."

$$\sigma = \sqrt{\mu_2} \quad (10)$$

The standard deviation gives a measure of the dispersion of the frequency distribution that for most purposes is preferable to the average deviation described below. The standard deviation of a normal frequency distribution, measured above and below the arithmetic mean, includes about 68 per cent of the cases. Twice the standard deviation measured above and below the arithmetic mean of a normal frequency distribution includes all but 5 per cent of the cases. When a normal distribution of proportional frequencies is regarded as a distribution of probabilities it follows that the probability of a case falling beyond the limits of $\bar{X} \pm 2\sigma$ of a normal frequency distribution is approximately .05, the probability of a case falling beyond $\bar{X} - 2\sigma$ is .025, and the probability of a case falling beyond $\bar{X} + 2\sigma$ is .025. These facts are of basic importance in sampling theory when the normal curve is used.

Average Deviation. The average deviation is a measure of dispersion that has its minimum value when deviations are measured from the median. To compute the average deviation from the median, subtract each of the N values of X from the median, add the absolute values of the deviations, and divide the sum by N . Thus,

$$\text{A.D.} = \frac{\sum |F(X - \text{Mi})|}{N} \quad (11)$$

The average deviation is less affected by extreme deviations than the more popular standard deviation, and for this reason it probably has greater sampling reliability from extremely leptokurtic (peaked) populations.

Range. The absolute range of a frequency distribution is the difference between the highest and lowest values of the distribution. The relative range is this difference divided by the standard deviation.

Bivariate and Multivariate Distributions. In the univariate frequency distributions discussed in the preceding sections, the data were classified according to a single characteristic. In

TABLE 3.—A BIVARIATE FREQUENCY DISTRIBUTION OF 81 MOUNT HOLYOKE FRESHMEN ACCORDING TO THEIR GRADES IN FIRST- (X_2) AND SECOND- (X_1) SEMESTER ENGLISH

$X_2 \backslash X_1$	60-	80-	100-	120-	140-	160-	180-	200-	220-	240-	260-	280-	F
60-				1									1
80-													0
100-	2		1										3
120-													0
140-				1			1						2
160-				5	3	1							9
180-					2	4	2						8
200-						3	4	7	2				16
220-							2	4	7	4			17
240-								2	7	3	1		13
260-									1	4	4		9
280-									1				1
300-												2	2
F	2	0	1	7	5	8	9	13	18	11	5	2	81

bivariate or multivariate distributions, data are classified according to two or more characteristics. Table 3 is an illustration of a bivariate frequency distribution. The frequency distribution showing grades of the 81 Mount Holyoke freshmen in second-semester English, shown in the total column at the

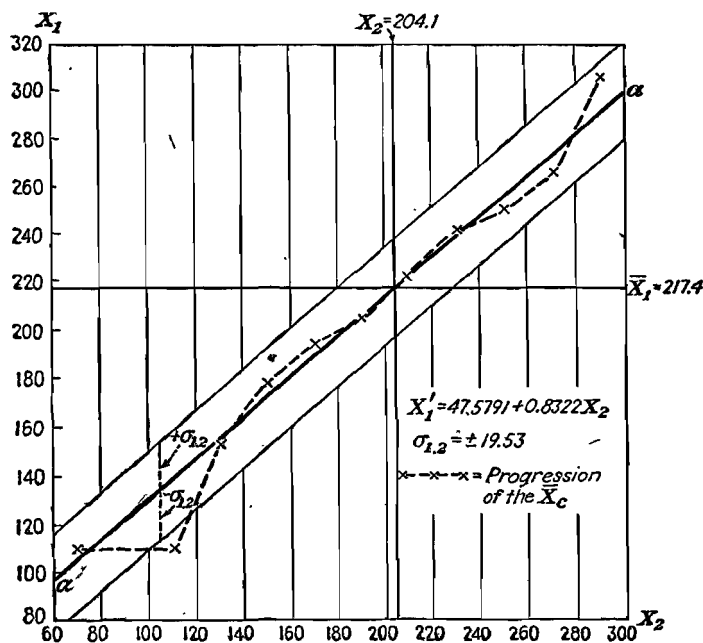


FIG. 5.—Progression of the means of X_1 with changes in X_2 .

right, under the caption F , is cross classified to show how each group of freshmen did in its first-semester English. Thus of the 8 students having second-semester grades between 180 and 200, row 7 of Table 3 shows that 2 had first-semester grades between 140 and 160, 4 had first-semester grades between 160 and 180, and 2 had first-semester grades between 180 and 200. This is a small univariate frequency distribution of the group of students who had grades between 180 and 200 in their second-semester course. In Table 3 there are 13 rows and 12 columns, of which 11 (in both instances) contain univariate frequency distributions. Since there are 11 subgroups of 11 groups, there are altogether 121 classes, represented by 121 squares or cells in the table, of which 28 contain frequencies.

The characteristics of a bivariate frequency distribution can be described by various statistics. Many of these are the same as the statistics employed in the description of a univariate frequency distribution, and some are new. Thus, the central tendency of one of the two variables may be measured by its

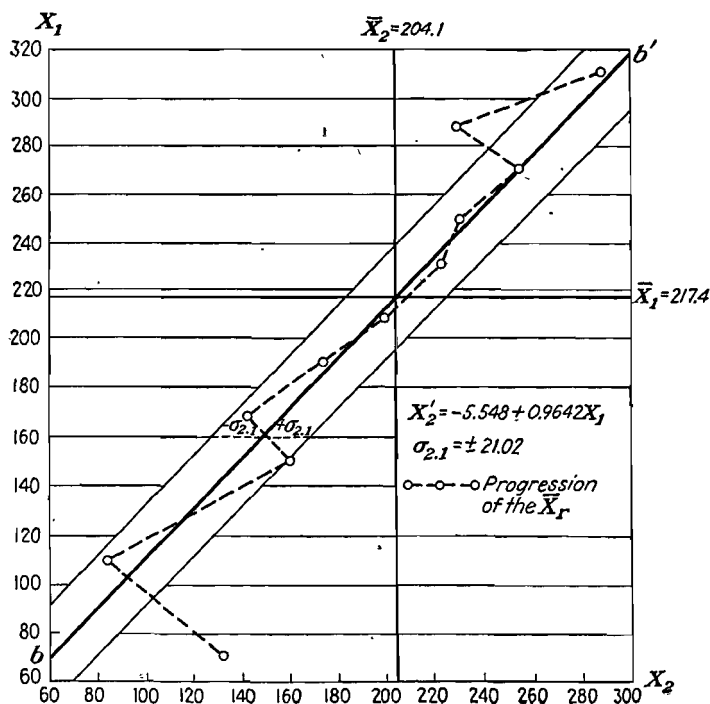


FIG. 6.—Progression of the means of X_2 with changes in X_1 .

mean, or its mode, or its median. Similarly, the dispersion of this variable may be measured by its range, its standard deviation, its average deviation, or its quartile deviation; and its skewness and kurtosis may be measured by β_1 and β_2 , respectively. The same is true of the other variable and of the numerous univariate frequency distributions that make up the details of a single bivariate distribution, *i.e.*, each frequency distribution of the rows and columns.

Progressions of Means. If the data are grouped in the form of a bivariate scatter diagram such as Table 3, one way to measure the association between the two variables is to compute the

mean values of one variable for various values of the other, *i.e.*, the means of the rows and the means of the columns of Table 3. The means of the columns show how the X_1 variable tends to change, on the average, with changes in X_2 ; and the means of the rows show how the X_2 variable tends to change, on the average,

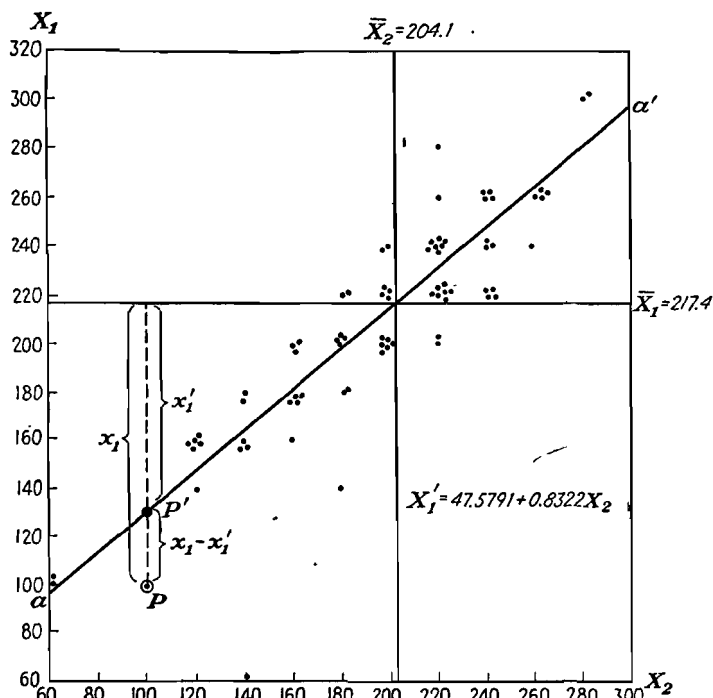


FIG. 7.—The fitting of the line of regression of X_1 on X_2 by the method of least squares (vertical deviations minimized).

with changes in X_1 . These means have been computed and shown, respectively, in Figs. 5 and 6, in which the means are connected by a series of straight lines; these are called “polygons of regression.”

Lines of Regression. The tendency of the progressions of the means to follow straight lines suggests the following hypothesis. Suppose that X_1 is so related to X_2 that an increase in X_2 of one unit always produces an increase in X_1 of, say, b units, b being a constant. If X_2 were the only factor affecting X_1 , all the values of X_1 , when plotted, would fall exactly on a straight line and the progression of all means would be perfectly linear;

all of them would be on the line representing the equation $X'_1 = 47.58 + .8322X_2$ shown in Fig. 7.

If, however, other forces also affect X_1 , causing it to be higher or lower than the value expected from its association with X_2 , the actual values of the means would not fall on the straight line

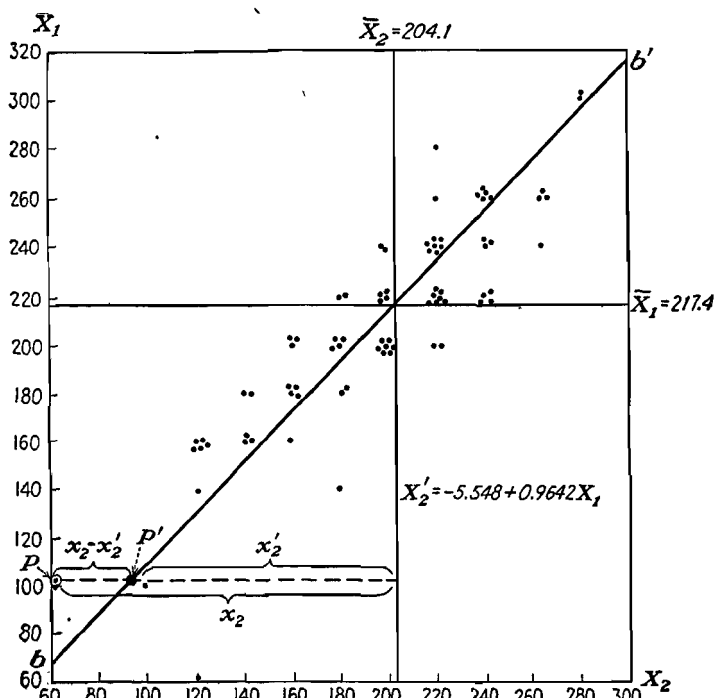


FIG. 8.—The fitting of the line of regression of X_2 on X_1 by the method of least squares (horizontal deviations minimized).

but would be scattered about the line in the manner shown in Fig. 7. According to such a hypothesis, a straight line fitted to the data should give the law of relationship between X_1 and X_2 , and the scatter about the line should give the deviation from this line caused by the other factors affecting X_1 .

A similar view could be taken of the variation in the mean value of X_2 with changes in X_1 and would justify drawing a straight line to show the law of relationship between X_2 and X_1 . This is illustrated in Fig. 8.

The lines that are derived to show the relationship between the mean value of one variable and the value of another are called "lines of regression," following Francis Galton's terminology, who used this term in his original study of the relationship between the heights of children and the heights of their parents.¹ A line of regression of one variable on another is to be interpreted as indicating the values of the first, the dependent variable, that would be obtained for various values of the second, the independent variable, if no other forces were affecting the dependent variable.

First-order Standard Deviations. In dealing with bivariate and multivariate frequency distributions, it becomes necessary to differentiate among zero-order standard deviations, first-order standard deviations, second-order standard deviations, etc. The zero-order standard deviations are standard deviations of the original variables. The first-order, second-order, and higher-order standard deviations are standard deviations of variation from lines and planes of regression.

In the case of the monovariate distribution, the representativeness of the mean depended upon how closely the cases were scattered around this mean value. This scatter was measured by the zero-order standard deviation. Similarly, in the case of a line of regression in a bivariate distribution, the representativeness of the line as a measure of the law of relationship between the two variables depends on the scatter of cases above and below the line or to the right and to the left of it. The representativeness of the line of regression depicting the equation $\bar{X}'_1 = 47.58 + .8322X_2$, used in Fig. 7, may be indicated by the scatter of cases above and below it. A measure of this scatter would be the standard deviation of the vertical deviations (of individual cases, not of means) from the line. The standard deviation about the line $\bar{X}'_2 = -5.548 + .9642X_1$, shown in Fig. 8, would be the standard deviation of the horizontal deviations from that line, made by the scatter of cases (not by the scatter of means) about the line. These two standard deviations are called "first-order standard deviations."

The first-order standard deviations will always be less than the zero-order standard deviations, for the part of the variation

¹ Cf. p. 19 for method used to derive the equations for these lines.

represented by the line of regression has been eliminated by taking the deviations from the line.

If the equation for the line of regression is $X'_1 = a_{1.2} + b_{12}X_2$, the first-order standard deviations are found as follows:

$$N\sigma_{1.2}^2 = \Sigma X_1^2 - a_{1.2}\Sigma X_1 - b_{12}\Sigma X_1X_2 \quad (12)$$

and

$$N\sigma_{2.1}^2 = \Sigma X_2^2 - a_{2.1}\Sigma X_2 - b_{21}\Sigma X_1X_2 \quad (12')$$

Sometimes these first-order standard deviations are called "standard errors of estimate" since they indicate the error involved in using the line of regression as an estimate of the dependent variable.

The Pearsonian Coefficient of Correlation. The progression of the means and the lines of regression described above are concerned with depicting the "law of relationship" between the two variables. They give the average value of one variable associated with given values of the other variable and show how these average values tend to change in unison with the other variable. Another statistic, called the "Pearsonian coefficient of correlation," aims to measure the degree of association between the two variables. This coefficient of correlation is found by the equation

$$r_{12} = \frac{\Sigma x_1x_2}{N\sigma_1\sigma_2} \quad (13)$$

in which x_1 and x_2 refer to deviations from the means and N to the number of pairs of cases.

Relationship between r and the First-order Standard Deviation. If the variables are measured from their mean values, Eq. (12) becomes

$$N\sigma_{1.2}^2 = \Sigma x_1^2 - b_{12}\Sigma x_1x_2 \quad \text{since } \Sigma x_1 = 0$$

If the value of b_{12} is determined by the method of least squares, as Eq. (12) and (12') presuppose,¹

$$b_{12} = \frac{\Sigma x_1x_2}{\Sigma x_2^2} = r_{12} \frac{\sigma_1}{\sigma_2} \quad \text{since } \Sigma x_1x_2 = N\sigma_1\sigma_2r_{12}$$

and the value of $N\sigma_{1.2}^2$ may be written

$$N\sigma_{1.2}^2 = N\sigma_1^2 - N\sigma_1^2r_{12}^2$$

¹ Cf. SMITH and DUNCAN, *op. cit.*, pp. 331-333.

so that

$$\sigma_{1.2}^2 = \sigma_1^2(1 - r_{12}^2)$$

and, finally,

$$\sigma_{1.2} = \sigma_1 \sqrt{1 - r_{12}^2} \quad (14)$$

In the same manner,

$$\sigma_{2.1} = \sigma_2 \sqrt{1 - r_{12}^2} \quad (14')$$

These equations, it is to be noted, may be put in the following form:

$$r_{12}^2 = 1 - \frac{\sigma_{1.2}^2}{\sigma_1^2} \quad (15)$$

$$r_{12}^2 = 1 - \frac{\sigma_{2.1}^2}{\sigma_2^2} \quad (15')$$

From this form [Eq. (15)] it is readily seen that r is closely related to the first-order variance, *i.e.*, to the scatter about the line of regression. If the scatter about the line of regression is a small percentage of the total scatter, this signifies that the line of regression itself accounts for a large part of the variation in X_1 , and r_{12} is high. If the scatter (the first-order standard deviation) about the line of regression is a large percentage of the total variation in the dependent variation, this signifies that only a small part of the variation is shown in the line of regression, and r_{12} is low. That is, the better the line of regression fits the data, the higher the value of r , and vice versa. The Pearsonian coefficient of correlation is thus a measure of the goodness of fit of the lines of regression.

The Pearsonian Coefficient of Correlation and the Breakup of Variance. For every point on a bivariate scatter diagram such as Fig. 7, there is a corresponding point on the line of regression of X_1 and X_2 . Geometrically, the former is obtained by projecting the point vertically onto the line of regression. This is illustrated by points P and P' in Fig. 7. Algebraically, the X_1 coordinate of a point on the line of regression is found by substituting the given values of X_2 in the regression equation $X'_1 = a_{1.2} + b_{12}X_2$ or, if the X 's are both in terms of deviations from their respective means, $x'_1 = r_{12} \frac{\sigma_1}{\sigma_2} x_2$.

When the variables are measured from their respective mean values, the mean of the various values of x_2 is zero. Hence the

mean of the corresponding values of x'_1 is zero also; and the standard deviation of these x'_1 values is accordingly as follows:

$$\sigma^2_{x'_1} = \frac{\Sigma(x'_1)^2}{N} = r_{12}^2 \frac{\sigma_1^2}{\sigma_2^2} \frac{\Sigma x_2^2}{N} = r_{12}^2 \sigma_1^2$$

This describes the part of the variance in X_1 that is depicted by the line of regression; and it is to be noted that Eq. (15) may consequently be written

$$\sigma_{1.2}^2 = \sigma_1^2 - \sigma_{x'_1}^2$$

or, in other words, .

$$\sigma_1^2 = \sigma_{x'_1}^2 + \sigma_{1.2}^2 \quad (16)$$

$$\sigma_2^2 = \sigma_{x'_2}^2 + \sigma_{2.1}^2 \quad (16')$$

Equation (16) says that the total variance in X_1 values is equal to the variance of the corresponding points on the line of regression plus the variance of the deviations from these points. Another way of looking at this is that the total variance in X_1 is made up of two parts, one consisting of the variance due to its association with X_2 ($\sigma_{x'_1}^2$) as represented by the line of regression, the other representing the variance of X_1 due to its association with factors independent of X_2 ($\sigma_{1.2}^2$).

It was also found just above that $\sigma_{x'_1}^2 = r_{12}^2 \sigma_1^2$; hence,

$$r_{12}^2 = \frac{\sigma_{x'_1}^2}{\sigma_1^2} \quad (17)$$

$$r_{12}^2 = \frac{\sigma_{x'_2}^2}{\sigma_2^2} \quad (17')$$

These equations shed further light on the meaning of r , which is consistent with the explanation of the significance of Eq. (16). Equation (17) shows that r_{12}^2 measures the proportion of the total variance in X_1 that is due to its association with X_2 . It also measures the proportion of the total variance in X_2 that is due to its association with X_1 , as indicated in Eq. (17').

The Correlation Ratio. The correlation ratio is a measure of correlation designed to be used in cases where correlation between the two variables is nonlinear. In such instances it is not appropriate to compute a straight line of regression; but an analysis similar to that outlined above may be applied. This latter analysis makes use of the polygon of regression rather than

the line of regression; *i.e.*, it makes use of the means of the columns and the means of the rows. The first-order variance is calculated by measuring the deviations in each column from its column mean, summing them, and dividing by N ; this is the variance in X_1 due to association with factors independent of X_2 . The amount of variance in X_1 explained by association or correlation with X_2 is then the variance in the column means from their mean, which is the mean of the whole distribution, namely, $\bar{X}_1$. The ratio of either of these variances to the total variance can be used to measure nonlinear correlation, and the ratio is called the "correlation ratio." The symbol used is the Greek lower-case letter eta (η). While it has been seen that $r_{12} = r_{21}$, it is not true that η_{12} equals η_{21} .

The correlation ratios may be defined by the following equations:

$$\eta_{12}^2 = 1 - \frac{\sigma_{\bar{X}_1 - \bar{X}_c}^2}{\sigma_1^2} = \frac{\sigma_{\bar{X}_c}^2}{\sigma_1^2} \quad (18)$$

$$\eta_{21}^2 = 1 - \frac{\sigma_{\bar{X}_r - \bar{X}_r}^2}{\sigma_2^2} = \frac{\sigma_{\bar{X}_r}^2}{\sigma_2^2} \quad (18')$$

PART I

General Theory of Frequency Curves

CHAPTER II

PROBABILITY AND THE PROBABILITY CALCULUS

The theory of frequency curves is in large measure a special application of the theory of probability. An understanding of probability and the probability calculus is therefore essential for a study of the theory of frequency curves. Probability theory in turn is principally based on combinatorial analysis. This chapter on probability and the probability calculus will accordingly begin with a review of permutations and combinations.

Permutations and Combinations. *Permutations.* A permutation is an arrangement. If there are N things, they may be arranged in $N!$ different ways; for the first selection may be made in N ways, the second selection in $N - 1$ ways, the third in $N - 2$ ways, etc. Hence the total number of arrangements, or permutations, that may be made of N things is

$$N(N - 1)(N - 2) \cdots 1 = N!$$

For example, the number of different permutations of five things is $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$.

Sometimes in forming permutations only a fraction of the total number of objects can be selected at one time. Thus an arrangement is to consist of three objects, but there are five objects from which to choose. In this instance the total number of different arrangements that can be made by the selection of three objects from the group of five will equal $5 \times 4 \times 3 = 60$, for there are five ways of selecting the first object, four ways of selecting the second, and three ways of selecting the third.

In general, the number of permutations of N things taken r at a time is

$$P_r^N = N(N - 1)(N - 2) \cdots (N - r + 1) = \frac{N!}{(N - r)!} \quad (1)$$

Combinations. Some of the permutations of Eq. (1) will be constituted alike; they will merely be different arrangements of the same combination of things. For example, suppose the five objects are the letters a, b, c, d, e . If permutations are made of these five letters three at a time, two of these permutations would be abc and acb , which are the same combination of the letters a, b , and c arranged in different order. To find the number of different combinations of three letters each that may be made from the group of five letters, it is necessary to allow for the number of permutations that can be made from any one combination. In the first section it was found that a given set of N objects can be arranged in $N!$ ways. Hence every combination of three letters can be made to yield $3! = 3 \times 2 \times 1 = 6$ permutations without changing the combination. It follows that the total number of combinations of three letters each that may be made from a group of five letters is equal to the total number of permutations of five things taken three at a time [$N!/(N-r)! = 5!/2!$] divided by the number of permutations of three things taken all at a time ($r! = 3!$); that is, $C_3^5 = 5!/2!3!$.

In general, the number of combinations of N things taken r at a time equals

$$C_r^N = \frac{N!}{r!(N-r)!} \quad (2)$$

The last equation may also be looked upon as the number of different combinations that may be made of N things by putting r of them in one category and $N-r$ in another. For obviously the number of different combinations that may be made of 10 men by putting 3 of them on a committee and leaving 7 of them off is the same as the number of different committees of 3 that may be picked from 10 men.

This new way of looking at the problem is especially helpful when more than two categories are involved. Suppose, for example, that three committees are to be chosen from 10 men; say a finance committee consisting of 3 men, a production committee consisting of 5 men, and a personnel committee consisting of 2 men. If each man is to serve on a committee, but no man is to serve on more than one committee, how many different committees of this kind could be formed from the 10 men? Here it is a question of how many different combinations can be made

of 10 things, 3 of them to be placed in one category, 5 in another, and 2 in another.

The answer to this broader question is obtained in the same way as the solution of the two-category case. The total number of different ways of picking 10 men from 10 men is $10!$. But not all these different ways of picking 10 men will lead to differently constituted committees. For any one set of committees could be picked in $(3!)(5!)(2!)$ different ways without changing the make-up of the committees. Thus, if the finance committee consisted of Mr. A, Mr. G, and Mr. J, the committee could be picked in that order, or in the order A, J, G, or the order G, A, J, or G, J, A, or J, A, G, or finally J, G, A. But each of these $3!$ ways of picking the finance committee could be combined with the $5!$ ways of picking the production committee, which would make $(3!)(5!)$ different ways of picking these two committees. Finally, each of these $(3!)(5!)$ ways of picking the finance and production committees could be combined with the $2!$ ways of picking the personnel committee, making a total of $(3!)(5!)(2!)$ different ways of selecting these three committees without changing their ultimate constituency. It follows, therefore, that the total number of different committees that can be picked is equal to $10!/3!5!2!$. In general, the number of different combinations that may be made of N things by putting N_1 of them in one category, N_2 of them in another category, N_3 of them in a third category, and N_k of them in a k th category, where

$$N_1 + N_2 + N_3 + \dots + N_k = N$$

is

$$C_{N_1, N_2, N_3, \dots, N_k}^N = \frac{N!}{N_1!N_2!N_3! \dots N_k!} \quad (3)$$

The Binomial Expansion. An important use of the combinatorial analysis is in the derivation of a formula for the expansion of the binomial $(a + b)^N$. The expression $(a + b)^N$ means the multiplication of $(a + b)$ by itself N times. Terms of the product are thus formed by multiplying the a 's of a various number of factors by the b 's of the remaining number of factors. The product $a^r b^{N-r}$ will therefore appear C_r^N times, since this is the number of different ways in which r of the a 's can be selected from N factors, no consideration being given to the order of selection.

The equation for the expansion of the binomial $(a + b)^N$ is accordingly

$$(a + b)^N = C_N^N a^N + C_{N-1}^N a^{N-1} b + \dots + C_{N-r}^N a^{N-r} b^r + \dots + C_0^N b^N \quad (4)$$

This is known as the "binomial expansion." As will be seen in the next chapter, there is a frequency distribution whose relative frequencies are computed in the same way as the terms of the binomial expansion. It is consequently known as the "binomial distribution."

The Multinomial Expansion. An argument similar to that just outlined shows that the general term of the expansion of the multinomial $(X_1 + X_2 + X_3 + \dots + X_k)^N$ is given by Eq. (3). There is also a frequency distribution whose relative frequencies are computed in the same way as the terms of this multinomial expansion, and it is hence called a "multinomial distribution."¹

Mathematical Probability. *Definition.* If, in a given set of t objects, m possess a given property and n do not possess this property, the probability of an object of this set having the given property is m/t , or the relative frequency of these objects in the set. The word "object" may include events that have the property of occurring or even propositions that have the property of being true, as well as material or immaterial things. To illustrate probability by a simple case, consider an ordinary deck of playing cards containing 13 cards in each suit so that the probability of a heart is $\frac{1}{2} \times \frac{3}{4} = \frac{1}{4}$. This is also the probability of a diamond, spade, or club in an ordinary deck of cards. In this illustration the probability set is finite. The definition of probability as a relative frequency is equally valid, however, for infinite probability sets.

In defining a probability, care must always be taken to see that the set of objects is precisely designated and that the property of the object to which the probability refers is carefully distinguished. It is obvious that the probability of an ace in an ordinary deck of cards is not the same as the probability of an ace in a pinochle deck, since the latter contains 48 cards of which 8 are aces, while an ordinary deck contains 52 cards of which 4 are aces. If a probability of an ace is defined with reference to an

¹ See Chap. XIII.

ordinary deck, it must not in subsequent stages of the analysis be taken as referring to a pinochle deck. This should be clear to anyone; yet the pitfalls of such a shift lie athwart the path to more intangible analysis, a path that is strewn with the intellectual bones of the unwary.

In this connection it is to be noted that the probability of a heart in an ordinary deck, say, is not necessarily the same as the probability of drawing a heart from an ordinary deck. The first probability refers to a set of 52 cards, 13 of which are hearts, so that the probability of a heart in an ordinary deck is clearly $\frac{13}{52} = \frac{1}{4}$. The second probability refers to a set of drawings from an ordinary deck. The probability of a heart in this set of drawings is the relative frequency of the number of hearts in the total set of cards drawn. The precise value of this second probability cannot be given until the number of cards drawn and the number of hearts among them have been counted. It may be $\frac{1}{4}$, or it may be some other value.

Law of Large Numbers. That the probability of drawing a heart from an ordinary deck is commonly said to be $\frac{1}{4}$ is the outcome of experience with certain kinds of mass phenomena. It has been found that if cards are drawn at random from an ordinary deck, the card being replaced and the deck reshuffled after each draw, the ratio of the number of hearts to the total number of cards drawn tends to approximate $\frac{1}{4}$ whenever the number of drawings is large. This experience with mass random¹ phenomena is generalized in the law of large numbers.

The law of large numbers says that, when a large number of random events is involved, it is possible to predict with reasonable accuracy the relative frequency of recurrence of a particular event by calculating a certain mathematical probability ascertainable from a carefully defined probability set. This law is the link between abstract calculations of probability and the prediction of relative frequencies of events in real life. With the assumption of its validity, the principal problem in most cases becomes that of finding the mathematical model that is the correct one for the particular phenomenon in question.

The probabilities in some sets are unknown but have been empirically approximated. Thus, if the empirically determined

¹ For a discussion of "randomness" see pp. 154-162.

probability of a man dying at age fifty is used in practice by life-insurance companies as a method of predicting, experience shows that these empirically determined probabilities yield good predictions if a large enough number of men is involved. With large masses of data, therefore, empirically determined probabilities may be used in the same way as known probabilities of a given set, as above illustrated, are used to make predictions.

Probability Distributions. *Definitions.* Any ordinary frequency distribution may be expressed in the form of a probability distribution by describing the frequencies as percentages of the total number of cases. A probability distribution can therefore be discrete or continuous. A continuous frequency curve that represents the distribution of relative frequency of an infinite population of cases is also a probability curve. Accordingly, all the measures of the various characteristics of frequency distributions, which have been summarized in the preceding chapter, apply to probability distributions; thus a probability distribution has a mean, a standard deviation, a coefficient of skewness, and a coefficient of kurtosis, like any frequency distribution. There are also bivariate and multivariate distributions of probability, corresponding to bivariate and multivariate distributions.

Probability Equations. Probability equations may be written in two ways. If the distribution is discrete, the probability of the attribute X may be written simply $Y = \varphi(X)$. The probability in this case is represented by the height of the ordinate Y at the abscissa point X . If the distribution is continuous, the equation $Y = \varphi(X)$ serves as an algebraic description of the curve but it is not a true measure of the probability. It merely represents the height of the curve at an abscissa point X .

For continuous distributions the proper form for representing probability is $d(F/N) = \varphi(X) dX$, or $dP = \varphi(X) dX$. In this form the probability, or relative frequency, of a case lying between X and $X + dX$ is expressed as a function of the attribute X . A probability, or frequency, curve is the limit approached by an area histogram as the class interval is made infinitesimally small. Thus the expression $d(F/N) = \varphi(X) dX$ or $dP = \varphi(X) dX$ merely says that, when the class interval (of size dX) is made infinitesimally small, the area under the curve for any class interval [that is, $d(F/N)$ or dP] is approximately equal to the area of a small rectangle whose base is dX and whose height

is the ordinate of the curve $[\varphi(X)]$ at some arbitrary value of X within the interval.

As an illustration, consider the normal probability curve. The algebraic equation for the normal curve is

$$Y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-\bar{X})^2}{2\sigma^2}} \quad (5)$$

This expresses the ordinate Y simply as a function of the abscissa X . The probability, however, of a normal variate lying between X and $X + dX$ is given by the equation

$$dP = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-\bar{X})^2}{2\sigma^2}} dX \quad (5')$$

In both cases $e (= 2.7183+)$ is the base of the Napierian system of logarithms, $\bar{X}$ is the mean of the distribution, and σ its standard deviation.

The Probability Calculus. *The Addition Theorem.* If the attributes of a given probability set are $X_1, X_2, \dots, X_s$ (representing either qualitative or quantitative characteristics) and their probabilities are $p_1, p_2, \dots, p_s$, then the probability of X_1, X_2 , or X_3 , say, *i.e.*, the attribute of being any one of these X 's, is $p_1 + p_2 + p_3$. If the variation in attributes within the probability set is continuous and if the distribution of probability is described by an equation such as $dP = \varphi(X) dX$, the probability of an attribute within any one of a number of small ranges dX whose sum constitutes the range X_1 to X_2 is given by

$$\sum_{X_1}^{X_2} dP = \sum_{X_1}^{X_2} \varphi(X) dX, \text{ or, in the symbolism of the integral calculus,}$$

$$\int_{X_1}^{X_2} dP = \int_{X_1}^{X_2} \varphi(X) dX.$$

The addition theorem is stated briefly as follows: The probability of either one of two mutually exclusive events (attributes) is the sum of their individual probabilities.

In using this theorem care must be taken to interpret correctly the term "mutually exclusive." What the term signifies is that the addition theorem applies only to probabilities of one and the same probability set. For the attributes of a probability set are by definition mutually exclusive. Hence the effect of this

term is to warn the student to define carefully his probability set at the start. It is to be noted that, while the attributes of a given probability set are mutually exclusive, not all mutually exclusive attributes are members of the same probability set.¹

The Multiplication Theorem. The multiplication theorem pertains to the calculation of a probability of a derived, or second-order, probability set from the probabilities of two or more first-order sets. In deriving the multiplication theorem two cases are distinguished, one pertaining to independent probabilities, the other to dependent probabilities. Consider first the case of independent probabilities.

The number of dots on the faces of a die form a set of mutually exclusive attributes the probability of each of which is $\frac{1}{6}$. Two dice form two such probability sets. Each of these may be viewed as a first-order probability set. Consider now the attributes given by the sum of two faces of a pair of dice. If there is no restriction on the way in which the face of one die can be paired with the face of the other die (the condition of independence), then each face of one die can be paired with every face of the other die and the sum of these faces may take on the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. If all possible pairs are formed, there will be $6 \times 6 = 36$ pairs. Only one of these will yield the sum 2, viz., the pair 1 and 1. Hence, in this derived, or second-order, probability set of 36 combinations, the probability of a sum having the value 2 is $\frac{1}{36}$. This, however, is equal to the product of the probability of a 1 on one die times the probability of a 1 on the other die, that is, $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. This illustrates the multiplication theorem for independent probabilities. It says that, if the probability of attribute B of set II is independent of the probability of attribute A of set I, then the joint probability of A and B in the probability set formed by combining each attribute of set I with every attribute of set II is equal to the product of the probability of A in set I times the probability of B in set II. Succinctly, if $P(B)$ is independent of A , $P(A, B) = P(A) \times P(B)$.

If the probability of B is dependent in some way on the occurrence of the attribute A , then the multiplication theorem for

¹ For further discussion, see SMITH and DUNCAN, *Elementary Statistics, and Applications*, p. 269.

independent probabilities cannot be applied. In its place must be put the multiplication theorem for dependent probabilities. This says that the joint probability of A and B is equal to the probability of A times the probability of B given A .¹ Succinctly, $P(A,B) = P(A) \times P(B/A)$.¹

¹ For further discussion, see *ibid.*, pp. 269–274.

CHAPTER III

THE SYMMETRICAL BINOMIAL DISTRIBUTION AND THE NORMAL CURVE

The preceding chapter has provided the tools that are now to be employed in shaping a general theory of frequency curves. This chapter is the first step in the development of that theory.

The argument will begin with a study of a simple problem in combinatorial analysis. The basic data will be 10 coins. These will be marked with a head on one side and a tail on the other. The problem will be to determine the relative frequencies or probabilities of various types of combinations in the whole set of combinations that might be made from various arrangements of the 10 coins. Immediate attention will center on the form of this derived distribution of probability, and exact and approximate equations will be determined. Ultimately, consideration will be given to how this combinatorial analysis may explain some of the frequency distributions that appear in real life.

The Symmetrical Binomial Distribution. *Derivation.* Suppose there are 10 coins all exactly alike and each having a head and a tail. Each of these coins represents an elementary probability set having two attributes, a head and a tail. Since there is only one head and one tail on a coin, the probability of each is $\frac{1}{2}$.

The 10 coins may be arranged in various ways so as to form different combinations of heads and tails. The various possible combinations are as follows: no heads, 10 tails; 1 head, 9 tails; 2 heads, 8 tails; 3 heads, 7 tails; 4 heads, 6 tails; 5 heads, 5 tails; 6 heads, 4 tails; 7 heads, 3 tails; 8 heads, 2 tails; 9 heads, 1 tail; and 10 heads, no tails. The probability of the combination no heads, 10 tails, is the product $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = (\frac{1}{2})^{10}$. The probability of the particular combination HTTTTTTTTT is also $(\frac{1}{2})^{10}$; but since there are nine other ways in which combinations containing 1 head and 9 tails can be formed, the total

probability of a combination containing 1 head and 9 tails is $10(\frac{1}{2})^{10}$. In general, the total probability of a combination containing r heads and $10 - r$ tails is $C_r^{10}(\frac{1}{2})^{10}$, which is the formula for the $(r + 1)$ th term in the expansion of the binomial $(\frac{1}{2} + \frac{1}{2})^{10}$. The probabilities of the various combinations of heads and tails are shown in Table 4.

TABLE 4.—PROBABILITIES OF VARIOUS COMBINATIONS OF HEADS AND TAILS AMONG 10 COINS

Combinations of heads and tails		Probability
H	T	
0	10	$\frac{1}{1,024}$
1	9	$\frac{10}{1,024}$
2	8	$\frac{45}{1,024}$
3	7	$\frac{120}{1,024}$
4	6	$\frac{210}{1,024}$
5	5	$\frac{252}{1,024}$
6	4	$\frac{210}{1,024}$
7	3	$\frac{120}{1,024}$
8	2	$\frac{45}{1,024}$
9	1	$\frac{10}{1,024}$
10	0	$\frac{1}{1,024}$

For N coins, the probability of a combination having N_1 heads and $N - N_1$ tails is as follows:¹

$$P(N_1) = \frac{N!}{N_1!(N - N_1)!} \left(\frac{1}{2}\right)^N \quad (1)$$

or, if $N_2 = N - N_1$,

$$P(N_1) = \frac{N!}{N_1!N_2!} \left(\frac{1}{2}\right)^N \quad (2)$$

¹ See p. 24.

Equation (2) is thus the general equation for what is called the "symmetrical binomial distribution."¹

Characteristics. Mathematical analysis shows that in general the symmetrical binomial distribution has the following characteristics:²

$$\left. \begin{aligned} \bar{X} &= \frac{N}{2} \\ \sigma &= \sqrt{\frac{N}{4}} \\ \beta_1 &= 0 \\ \beta_2 &= 3 - \frac{2}{N} \end{aligned} \right\} \quad (3)$$

The variable, it will be noted, is N_1 , the number of heads:

The Normal Curve. *Derivation from Binomial Distribution.* If 40 coins were used, the distribution of probability of N_1 heads, $N - N_1$ tails would be considerably more spread out than when 10 coins were tossed. In general, the equation $\sigma = \sqrt{N/4}$ indicates that the dispersion of the distribution increases in proportion to the $\sqrt{N}$. If the horizontal scale is reduced, however, and the vertical scale enlarged, in the same proportion in which the dispersion of the distribution is increased, then the effect of increasing N is to bring the ordinates of the distribution closer together and to raise them to the height of the original distribution. Under these conditions the tops of the ordinates tend to sketch out a smooth curve as N is increased.

Equations (3) suggest that the curve approached as a limit by the symmetrical binomial as N increases is the normal probability curve. For $\beta_1 = 0$, and $\beta_2 = 3 - \frac{2}{N}$ approaches 3 as N approaches infinity. It can also be shown that if a line is drawn between two ordinates of the symmetrical binomial the ratio of the slope of this line to the average of the two ordinates, *i.e.*, the relative slope of the binomial distribution at any mid-point, is

¹ The name follows from the fact that it is the equation for the general term of the expansion of $(\frac{1}{2} + \frac{1}{2})^N$ (see pp. 25-26).

² These equations are derived in the Appendix to Chap. IV (pp. 65-67). It will be noticed that the parameters $\bar{X}$, σ , etc., are in boldface type since they are population values and not sample statistics. This differentiation must be carefully watched in this volume. Cf. SMITH and DUNCAN, *Elementary Statistics and Applications*, pp. 123-124, 316.

the same as the relative slope of the normal curve that is traced out by the binomial ordinates.¹ Both these suggestions that the limit of the symmetrical binomial is a normal curve are borne out by strict mathematical analysis.² Thus the limit of the symmetrical binomial is³

$$y = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{x^2}{2\sigma^2} \right] \quad (4)$$

Here x represents a deviation from the mean value and equals $N_1 - \frac{N}{2}$.

If z is set equal to x/σ and if probabilities are measured by areas instead of ordinates, the equation becomes

$$dP = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{z^2}{2} \right] dz \quad (5)$$

which is the equation for a normal curve whose standard deviation is 1. It is subsequently referred to as the standard normal curve. If N is large, therefore, the probability that N_1 lies between N'_1 and N''_1 can be found approximately by computing $\frac{N'_1 - \frac{N}{2}}{\sqrt{N/4}}$ and $\frac{N''_1 - \frac{N}{2}}{\sqrt{N/4}}$ and finding the area under the standard normal curve between these limits. The latter is easily accomplished by reference to the normal probability table given in Table VI of the Appendix.⁴

Significance of the Symmetrical Binomial Distribution and the Normal Curve. The practical significance of the symmetrical binomial distribution and the normal curve is that certain conditions in real life appear to produce frequency distributions that have this form. Consider first a real situation that very closely parallels the combinatorial problem of the preceding sections. Let 10 unbiased coins (unbiased in the sense that they are physically uniform) be tossed at random⁵ a large number of times.

¹ See Appendix to Chap. IV, pp. 74-76.

² *Ibid.*, pp. 68-76.

³ The expression $\exp [x]$ is merely another way of writing e^x . It is frequently used in this text to facilitate the printing of complicated equations.

⁴ For further discussion on how to use the normal probability table, see pp. 120-123, 164-168.

⁵ Whether a method is random or not has to be determined largely by intuition and experience with similar experiments. See pp. 154-162.

Experience with such experiments indicates that the relative frequencies with which 0, 1, 2, . . . , 10 heads appear will be approximately the same as the probabilities of the binomial distribution. This is merely one instance of the law of large numbers referred to in Chap. II.

Again, suppose a large number of flour bags, each weighing 5 pounds at the start, are opened in succession and a certain quantity of flour is added or taken away in accordance with the following rule: When a bag is opened, 10 coins are tossed and an ounce of flour is added for each head that appears, and subtracted for each tail that appears. For this experiment the law of large numbers suggests that the outcome will be a set of bags varying in weight approximately as follows:

TABLE 5.—RELATIVE FREQUENCIES OF BAGS OF FLOUR OF SPECIFIED WEIGHTS

Weight of Bag	Relative Frequency
4 lb. 6 oz.	$\frac{1}{1,024}$
4 lb. 8 oz.	$\frac{10}{1,024}$
4 lb. 10 oz.	$\frac{45}{1,024}$
4 lb. 12 oz.	$\frac{120}{1,024}$
4 lb. 14 oz.	$\frac{210}{1,024}$
5 lb. 0 oz.	$\frac{252}{1,024}$
5 lb. 2 oz.	$\frac{210}{1,024}$
5 lb. 4 oz.	$\frac{120}{1,024}$
5 lb. 6 oz.	$\frac{45}{1,024}$
5 lb. 8 oz.	$\frac{10}{1,024}$
5 lb. 10 oz.	$\frac{1}{1,024}$

In other words, the distribution of weights will approximately conform to a symmetrical binomial distribution with a mean weight of 5 pounds, and a standard deviation of $2\sqrt{2.5} = 3.162$ ounces.

If a much larger number of coins were employed and the amount of flour added or subtracted per head or tail were made very

small, the distribution of weights would become practically continuous and would form a normal curve.

The two examples just given suggest "laboratory" experiments by which a symmetrical binomial distribution and, in the second case, a normal curve might be produced by real events. Certain conditions in everyday life that appear to parallel these laboratory experiments also produce symmetrical binomial distributions and normal curves. Among a number of animals, for example, the biological conditions appear to be such that the chance of male offspring is equal to the chance of female offspring. Distributions of the number of males per families of given size therefore closely approximate the symmetrical binomial distribution. Again, in many cases of physical measurement, there are a host of forces tending to cause slight positive and negative errors, with the result that the net error of measurement is generally distributed like a normal frequency curve.¹ The heights of adult males of the same race, the heights of adult females of the same race, grades of students, the durability of electric-light bulbs, and many other biological, psychological, and physical variables are likewise normally distributed.

Summary of Conditions Leading to Symmetrical Distribution.

The foregoing analysis suggests that, whenever the following conditions exist in real life, the data generated by these conditions will tend to be distributed in the form of a symmetrical binomial distribution and, if certain other conditions are also present, in the form of a normal curve.

A. The conditions giving rise to the symmetrical binomial distribution may be stated as follows:

1. In the absence of certain "causes" of variation or in the event of a perfect balancing of their effects, the data assume a fixed central value (the 5 pounds of the flour illustration).

2. Deviations from this central value result from certain causes of variation, the effect of any cause being either to add a fixed quantity to the data or to subtract the same quantity (to add or subtract 1 ounce of flour).

3. The probability of a cause of variation producing a positive effect equals the probability of its producing a negative effect,

¹ For a more extended discussion see Smith and Duncan, *Elementary Statistics and Applications*, pp. 294-295.

that is, $P(+) = P(-) = \frac{1}{2}$ (the probability of a head equals the probability of a tail).

4. The effects of all contributory causes of variation are of equal magnitude (each adds or subtracts 1 ounce of flour).

5. The contributory causes are independent in their action. That is, the probability of a positive or negative contribution by any causal factor is independent of the previous contributions of other causal factors; the sets of deviations generated by the various causes are all independent.

6. The total deviation of any element from its central value is the algebraic sum of the positive and negative contributions of the individual causal factors (the total amount of flour added or subtracted from a bag is the sum of the ounces added for each head tossed minus the ounces subtracted for each tail tossed).

B. If, in addition to the above conditions, the following also exist, then the resulting distribution will tend to conform to the normal curve:

7. The number of contributory causes is very large (a large number of coins, instead of only 10 coins, are tossed).

8. The positive and negative contributions of each cause is very small (if .01 ounce is added or subtracted instead of 1 ounce).

It is to be noted that, so far as the normal curve is concerned, not all these conditions are necessary for its generation. The foregoing conditions will produce it, but it can be shown that the normal curve may also occur when some of these conditions are absent. The normal curve will still be produced if conditions 2 and 3 are relaxed so that a causal factor may affect the data in varying degree and with varying probabilities and also if condition 4 is only approximately and not exactly true.¹ Under certain conditions the requirement of independence (condition 5) may also be relaxed.

The most important conditions for the normal curve are 6, 7, and 8, and condition 4 in an approximate form. For example, in the case of the flour illustration, the resulting weights of the bags of flour would still tend to be normally distributed even if biased dice instead of unbiased coins were used and if the amount of flour added or subtracted varied with the result of the throw (say, .001 ounce for the occurrence of a one, —.002 ounce for the occurrence of a two, .003 ounce for the occurrence of a three,

¹ See pp. 137–142.

— .004 for the occurrence of a four, etc.), provided that the number of dice thrown were very large and the amount added or subtracted per die were very small and of about the same order of magnitude from die to die. The normal curve is thus a more general phenomenon than the symmetrical binomial distribution. Mathematically, the normal curve can be derived from a great variety of different assumptions.¹

¹ Cf. CZUBER, EMANUEL; *Theorie der Beobachtungsfehler*, B. G. Teubner, Leipzig, 1891.

CHAPTER IV

THE PEARSONIAN SYSTEM OF FREQUENCY CURVES

In the preceding chapter the conditions that would produce the symmetrical binomial distribution and the normal curve were outlined and briefly discussed. It was pointed out, however, that not all the conditions laid down are necessary for the production of the normal curve, and further consideration was postponed until this chapter. This more extensive examination will now be undertaken.

ASYMMETRICAL BINOMIAL DISTRIBUTION

Derivation. Beginning in this section some of the more important conditions will be examined that lead to nonnormality in homogeneous data. The discussion will again begin with some problems in the realm of combinatorial analysis.

Suppose there are 10 prisms, each of which has three faces marked with an H and the fourth marked with a T. The probability of an H on each prism is thus $\frac{3}{4}$, and the probability of a T is $\frac{1}{4}$. This differs from the previous coin problem in that the probabilities of the two attributes are not equal. As will be indicated later, it is an abandonment of condition 3 of the previously listed conditions for normality.

The particular problem that will be discussed is the determination of the probabilities of the various types of combinations that may be made from independent selection of the faces of these 10 prisms. The types of combinations will be the same as those of the coin problem (an H may be viewed as a "head" and a T as a "tail"), but the probabilities of each will be found to be different. For convenience let the various prisms be designated by the letters $A, B, C, \dots, J$.

Analysis of the problem begins, once again, with the determination of the probability of a particular combination. First consider the combination having no H's, viz.,

A	B	C	D	E	F	G	H	I	J
T	T	T	T	T	T	T	T	T	T

Since the probability of a T on each prism is $\frac{1}{4}$ and since the face of each prism is selected independently of the others, the probability of all prisms being T's is, by the multiplication theorem for independent probabilities,

$$\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \left(\frac{1}{4}\right)^{10}$$

Furthermore, the given combination is the only way in which no H's can occur. Hence the probability of no H's is just

$$\left(\frac{1}{4}\right)^{10} = \frac{1}{1,048,576}$$

Consider next the combination

$$\begin{array}{cccccccccc} A & B & C & D & E & F & G & H & I & J \\ H & T & T & T & T & T & T & T & T & T \end{array}$$

This is a combination having but one H. Since the probability of A being an H is $\frac{3}{4}$ and the probability of each of the other prisms being T's is $\frac{1}{4}$, the probability of this particular combination is

$$\frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^9$$

The same would be true of the combination

$$\begin{array}{cccccccccc} A & B & C & D & E & F & G & H & I & J \\ T & H & T & T & T & T & T & T & T & T \end{array}$$

or in fact of any combination having but one H. Since there are 10 such combinations altogether, the probability of any of these 10 combinations is, by the addition theorem,

$$10 \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^9 = \frac{(10)(3)}{1,048,576}$$

This, then, is the probability of one H.

The following is a combination having but two H's:

$$\begin{array}{cccccccccc} A & B & C & D & E & F & G & H & I & J \\ H & H & T & T & T & T & T & T & T & T \end{array}$$

The probability of this particular combination is

$$\frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^8$$

This is also the probability of any particular combination having

but two H's; and since there are 45 such combinations, the probability of any one of them is

$$45 \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^8 = \frac{(45)(9)}{1,048,576}$$

This is the probability of two H's.

In general, the probability of N_1 H's with 10 prisms is given by the equation

$$P(N_1) = \frac{10!}{N_1!(10 - N_1)!} \left(\frac{3}{4}\right)^{N_1} \left(\frac{1}{4}\right)^{10-N_1}$$

For $N_1 = 0$ to $N_1 = 10$, this equation yields the results shown in Table 6.

TABLE 6.—PROBABILITIES OF VARIOUS COMBINATIONS OF 10 PRISMS, EACH HAVING THREE SIDES MARKED WITH AN H AND ONE WITH A T

Combinations Having	Probability
0 H	$\frac{1}{1,048,576}$
1 H	$\frac{30}{1,048,576}$
2 H	$\frac{405}{1,048,576}$
3 H	$\frac{3,240}{1,048,576}$
4 H	$\frac{17,010}{1,048,576}$
5 H	$\frac{61,236}{1,048,576}$
6 H	$\frac{153,090}{1,048,576}$
7 H	$\frac{262,440}{1,048,576}$
8 H	$\frac{295,245}{1,048,576}$
9 H	$\frac{196,830}{1,048,576}$
10 H	$\frac{59,049}{1,048,576}$

It will be noted that the probabilities of Table 6 are the successive terms of the expansion of $(\frac{3}{4} + \frac{1}{4})^{10}$. The distribution obtained is thus a "binomial" distribution, but it is no longer symmetrical. If N prisms had been used, the probabilities obtained would have been the successive terms of the expansion

of $(\frac{3}{4} + \frac{1}{4})^N$ and the equation for any one term would have been

$$P(N_1) = \frac{N!}{N_1!(N - N_1)!} \left(\frac{3}{4}\right)^{N_1} \left(\frac{1}{4}\right)^{N - N_1}$$

or, if $N_2 = N - N_1$

$$P(N_1) = \frac{N!}{N_1!N_2!} \left(\frac{3}{4}\right)^{N_1} \left(\frac{1}{4}\right)^{N_2}$$

TABLE 7.—PROBABILITIES OF VARIOUS COMBINATIONS OF 10 PRISMS, EACH HAVING THREE SIDES MARKED WITH A T AND ONE WITH AN H

Combinations Having		Probability
0	H	<u>59,049</u>
		1,048,576
1	H	<u>196,830</u>
		1,048,576
2	H	<u>295,245</u>
		1,048,576
3	H	<u>262,440</u>
		1,048,576
4	H	<u>153,090</u>
		1,048,576
5	H	<u>61,236</u>
		1,048,576
6	H	<u>17,010</u>
		1,048,576
7	H	<u>3,240</u>
		1,048,576
8	H	<u>405</u>
		1,048,576
9	H	<u>30</u>
		1,048,576
10	H	<u>1</u>
		1,048,576

If each prism is replaced by a group of objects¹ within which the probability of an H is p_1 and that of a T is p_2 , if there are N such groups, and if combinations are composed so as to consist of one object from each group, then the probabilities of the various types of combinations among the set of all possible combinations will be given by the equation

$$P(N_1) = \frac{N!}{N_1!N_2!} p_1^{N_1} p_2^{N_2} \quad (1)$$

This is the most general equation for the binomial distribution. When $p_1 = p_2 = \frac{1}{2}$, it is the equation for the symmetrical bino-

¹ The number of objects in each group is not relevant to the argument.

mial distribution. When $p_1 \neq p_2$, it represents an asymmetrical binomial distribution.

Character of the Asymmetrical Binomial Distribution. A graph of the probabilities of Table 6 is shown in Fig. 9. The asym-

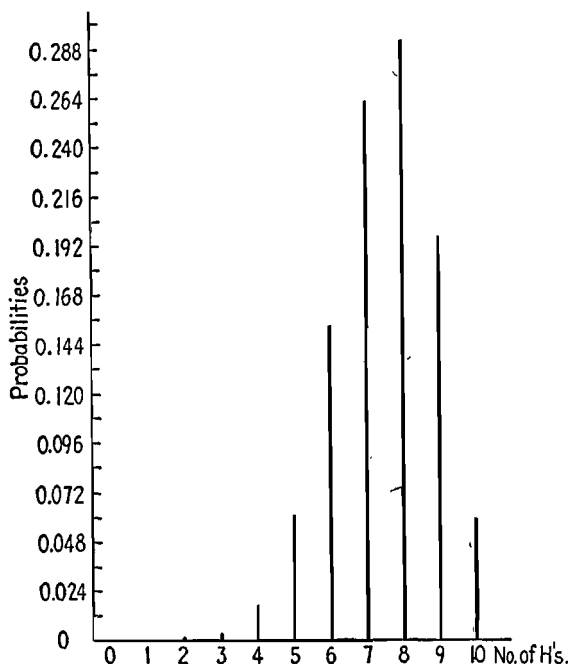


FIG. 9.—Asymmetrical binomial distribution, $N = 10$, $p_1 = \frac{2}{3}$, $p_2 = \frac{1}{3}$ (see Table 6).

metrical character of the distribution is obvious and needs no comment. Table 7 and Fig. 10 show what the distribution of the number of H's would have been if the probabilities of H's and T's had been reversed, *i.e.*, if the probability of an H had been $\frac{1}{3}$ and the probability of a T $\frac{2}{3}$. It will be noted that this alteration in the probabilities changes the skewness of the distribution from negative to positive. This is generally the case; if the probability of an H is greater than the probability of a T, the distribution of the number of H's will be negatively skewed; and if the probability of an H is less than the probability of a T, the distribution will be positively skewed.¹

¹ If the number of T's were taken as the attribute, the reverse would be true.

Mathematical analysis shows that in general the asymmetrical binomial distribution¹ has the following characteristics (the variable is the number of H's, N_1):

$$\left. \begin{aligned} \bar{X} &= Np_1 \\ Mo &= \text{integer between } Np_1 - p_2 \text{ and } Np_1 + p_1 \\ \sigma &= \sqrt{Np_1p_2} \\ \beta_1 &= \frac{(p_2 - p_1)^2}{Np_1p_2} \\ \beta_2 &= 3 + \frac{1 - 6p_1p_2}{Np_1p_2} \end{aligned} \right\} (2)$$

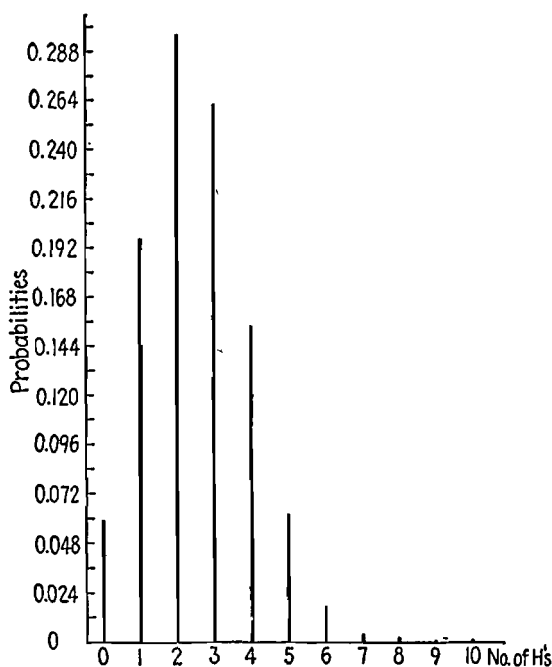


FIG. 10.—Asymmetrical binomial distribution, $N = 10$, $p_1 = \frac{1}{4}$, $p_2 = \frac{3}{4}$ (see Table 7).

As indicated above, the sign of $\sqrt{\beta_1}$ should be the sign of $p_2 - p_1$, so that it will be positive when the distribution is positively skewed and negative when it is negatively skewed. These equa-

¹ These, of course, also apply to the symmetrical binomial distribution in which case $p_1 = p_2 = \frac{1}{2}$.

tions are derived in the Appendix to this chapter.¹ It might be well for the nonmathematical student, however, to check them by calculating $\bar{X}$, σ , β_1 and β_2 of the distribution of Table 7 and then comparing the results given by the application of Eqs. (2).*

Asymmetrical Binomial Distribution and the Normal Curve.

If N is increased and if the horizontal scale is reduced in proportion to $\sqrt{N}$, the tops of the ordinates of the asymmetrical binomial distribution will tend to trace out a smooth curve, just as did the symmetrical binomial distribution. In general, the character of this curve will depend on the size of N and on the difference between p_1 and p_2 .

Equations (2) suggest that, if N is very large, any binomial distribution, asymmetrical or symmetrical, will be approximated fairly well by a normal frequency curve. For as N is increased, β_1 approaches 0 and β_2 approaches 3, which are the values of these coefficients for a normal curve. This implication is borne out by rigorous mathematical analysis.² It is to be concluded then that, if N is very large, an asymmetrical binomial distribution is approximated by a normal frequency curve, just as was the symmetrical binomial distribution.

The conclusion that an asymmetrical curve can be approximated by a symmetrical curve would seem at first glance to be paradoxical. It is to be noted, however, that this is true only when N is very large, and in that case the binomial distribution will not be very asymmetrical. For, as just indicated, the skewness of the asymmetrical binomial distribution diminishes as N increases. This follows directly from the fact that

$$\beta_1 = \frac{(p_2 - p_1)^2}{N p_1 p_2}.$$

It may also be noted from any graphic analysis

showing how the shape of a particular binomial distribution, *i.e.*, one with a given p_1 and a given p_2 , changes its shape as N is increased. Such a graphic comparison is presented in Fig. 11, for which the data are shown in Tables 8 and 9.

The size of N that must be attained before the asymmetrical binomial distribution becomes practically symmetrical and can

¹ See pp. 65-68.

* The direct calculation of these quantities is the same as that carried out in Smith and Duncan, *Elementary Statistics and Applications*, pp. 284-285, for the symmetrical binomial distribution.

² See Appendix to this chapter (pp. 68-74).

be approximated by the normal curve depends on the relative size of p_1 and p_2 . For as the formula for β_1 shows, if the difference between p_1 and p_2 is very great, then N must be so much the

TABLE 8.—PROBABILITIES OF VARIOUS COMBINATIONS OF FOUR PRISMS, EACH HAVING THREE SIDES MARKED WITH AN H AND ONE WITH A T
($N = 4$)

Combinations	Probability
0 H	.003
1 H	.047
2 H.	.211
2 H	.422
4 H	.316

TABLE 9.—PROBABILITIES OF VARIOUS COMBINATIONS OF 16 PRISMS, EACH HAVING THREE SIDES MARKED WITH AN H AND ONE WITH A T
($N = 16$)

Combinations	Probability
0 H	.0000000023
1 H	.000000012
2 H	.00000025
3 H	.0000035
4 H	.000034
5 H	.00025
6 H	.00135
7 H	.00583
8 H.	.01966
9 H	.05243
10 H	.11095
11 H	.18015
12 H	.22519
13 H	.20787
14 H	.13363
15 H	.05345
16 H	.01002

larger if β_1 is to be close to zero. On the other hand, if p_1 and p_2 are relatively close, then N need not be very large to make the distribution approximately symmetrical.

Asymmetrical Binomial Distribution and Pearson's Type III Curve. Although the asymmetrical binomial distribution approaches the normal curve when N is large relative to $p_2 - p_1$,

it is nevertheless of interest to find the type of curve that approximates this distribution in those cases when it is still markedly skewed. One approach to this problem is suggested by the previous analysis of the symmetrical binomial. The normal

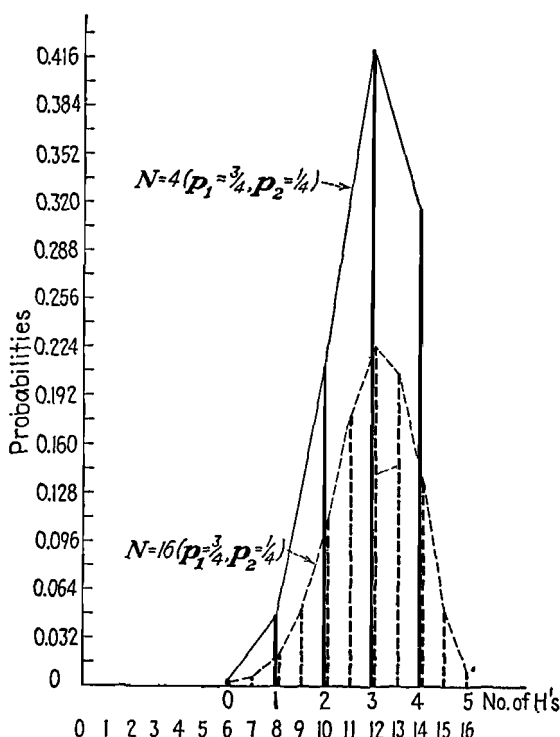


FIG. 11.—Comparison of two asymmetrical binomial distributions (see Table 9).

curve, it will be recalled,¹ was found to be the curve that had the same relative slope at various points as did the symmetrical binomial. This suggests that the curve that will give a good fit to the asymmetrical binomial will also be one for which the relative slope at various points is the same as the relative slope of the asymmetrical binomial distribution. Such a line of attack was adopted by Karl Pearson, and the curve that he found to have this property is the one whose logarithmic form is

¹ See pp. 34–35.

$$\log_{10} y = \log_{10} y_0 + k a \log_{10} \left(1 + \frac{x}{a} \right) - k x \log_{10} e \quad (3)$$

where x represents a deviation from the mode of the curve, y_0 the height of the curve at the mode, y the height of the curve at any point x , $\log_{10} e = .434294$, and a and k are constants depending

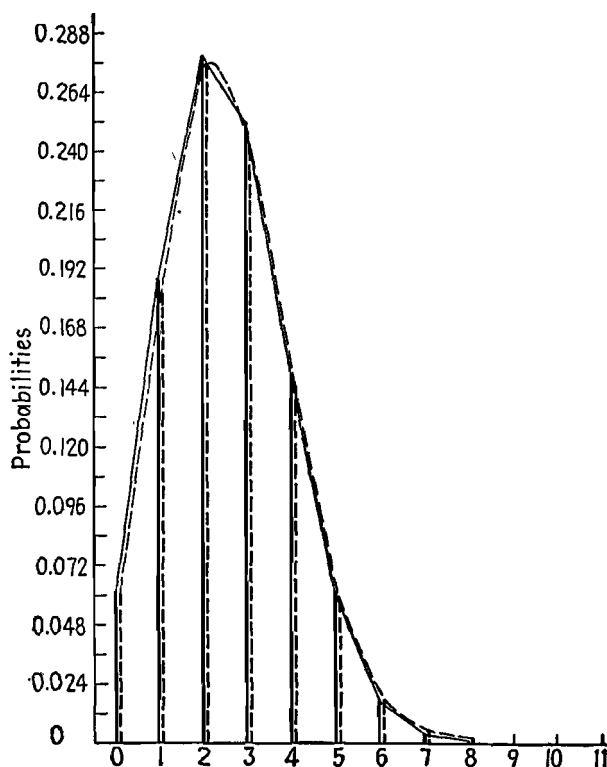


FIG. 12.—Comparison of asymmetrical (solid line) binomial with Pearson's type III curve (broken line) (see Table 10).

on the p_1 , p_2 , and N of the binomial distribution from which the curve is derived.¹ This curve has come to be known as Pearson's

¹ More specifically, $x = (N_1 + \frac{1}{2}) - Np_1 + p_2 - \frac{1}{2}$,

$\log_{10} y_0 = (ka + 1) \log_{10} ka - \log_{10} (ka)! - \log_{10} a - ka \log_{10} e$,
 $a = \frac{2p_1p_2(N+1)}{p_2 - p_1}$, and $k = \frac{2}{p_2 - p_1}$. See Appendix to this chapter (pp. 76-79).

type III curve.¹ The way in which it fits the asymmetrical binomial distribution of Table 10 is shown in Fig. 12.

TABLE 10.—COMPARISON OF ASYMMETRICAL BINOMIAL DISTRIBUTION WITH PEARSON'S TYPE III CURVE

X	Asymmetrical binomial distribution ¹ $P(X)$	Pearson's type III curve $P'(X)$
0	.056,314	.061,245
1	.187,712	.181,660
2	.281,568	.272,720
3	.250,282	.243,300
4	.145,998	.144,220
5	.058,399	.061,333
6	.016,222	.019,826
7	.003,090	.005,098
8	.000,386	.001,079
9	.000,029	.000,192
10	.000,001	.000,030

¹ Derived from Table 7 (p. 43), by converting the fractions into relative numbers.

HYPERGEOMETRICAL DISTRIBUTION

In the previous section it has been seen that, when the probabilities of the two attributes are not equal, the result is an asymmetrical instead of a symmetrical binomial distribution. The bearing of this upon the generation of nonnormal frequency distributions in real life will be discussed below.² Before turning to these broader aspects of the analysis, however, it is desirable to consider the effects of other modifications of the conditions for normality. It is of interest in particular to consider what happens when the condition of independence (condition 5)³ is removed. This will now be studied in some detail.

¹ The equation for this curve was first published by Karl Pearson in the *Proceedings of the Royal Society of London*, Vol. 54 (1893), p. 331. It was later discussed at some length in his fundamental paper on frequency curves, "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material." See the *Philosophical Transactions of the Royal Society of London*, Series A, Vol. 186 (1895), pp. 356-360. Also see W. P. Elderton, *Frequency Curves and Correlation*, Cambridge University Press, London (1927), pp. 45, 90-94.

² See pp. 60ff.

³ See p. 38.

Derivation of the Hypergeometrical Distribution. To show the effect of dependent probabilities on the form of a frequency distribution consider the following card problem. Suppose there are 10 packs of 52 cards, each containing at the start 13 spades, 13 hearts, 13 diamonds, and 13 clubs. Suppose further that all possible combinations are made by selecting one card from each pack in order, subject to the condition that if a card has already been selected from one pack the same card is not eligible for selection from subsequent packs. For example, if the ten of spades is selected from the first pack, then the ten of spades is not eligible for selection from any other pack. Again, if the ten of spades is selected from the first pack, the five of hearts from the second pack, and the ace of clubs from the third pack, then neither the ten of spades, nor the five of hearts, nor the ace of clubs is eligible for selection from packs 4 to 10. Given this restriction on the selection of cards, let the problem be to find the probabilities of combinations containing various numbers of spades in the whole set of combinations of 10 cards that may be formed in the way described. To facilitate the analysis let the cards that make up a combination be represented by letters from *A* to *J*.

Consider first the combination

$$\begin{array}{cccccccccc} A & B & C & D & E & F & G & H & I & J \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

where 0 stands for a card other than a spade. This is a combination containing no spades. The probability of the first card, *A*, in the combination being other than a spade is $\frac{39}{52}$, for there are 39 nonspades among the 52 cards open for selection from the first pack. The probability of the second card, *B*, being other than a spade is $\frac{38}{51}$, for after a nonspade has been selected from the first pack there are only 51 cards left for selection from the second pack and only 38 of these are nonspades. In like manner the probability of the third card, *C*, being other than a spade is $\frac{37}{50}$, etc.

By the multiplication theorem for dependent probabilities the probability of the combination containing no spades is therefore

$$\frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50} \cdot \frac{36}{49} \cdot \frac{35}{48} \cdot \frac{34}{47} \cdot \frac{33}{46} \cdot \frac{32}{45} \cdot \frac{31}{44} \cdot \frac{30}{43}$$

Consider next the combination

A	B	C	D	E	F	G	H	I	J
S	0	0	0	0	0	0	0	0	0

This is a combination containing only one spade. Its probability by the multiplication theorem for dependent probabilities is equal to the probability of a spade in the first pack (that is, $\frac{13}{52}$), multiplied by the probability of a nonspade in the second pack after a spade has been selected from the first pack (that is, $\frac{39}{51}$), times the probability of a nonspade in the third pack after both a spade and a nonspade have been selected from the first and second packs (that is, $\frac{38}{50}$), etc. In short, the probability of this particular combination is

$$\frac{13}{52} \cdot \frac{39}{51} \cdot \frac{38}{50} \cdot \frac{37}{49} \cdot \frac{36}{48} \cdot \frac{35}{47} \cdot \frac{34}{46} \cdot \frac{33}{45} \cdot \frac{32}{44} \cdot \frac{31}{43}$$

There are 9 other combinations, however, that contain only one spade, and in each case the probability can be shown to be the same as that just computed. Hence, by the addition theorem, the probability of any one of these 10 combinations, *i.e.*, the probability of a combination containing one spade, is

$$10 \cdot \frac{13}{52} \cdot \frac{39}{51} \cdot \frac{38}{50} \cdot \frac{37}{49} \cdot \frac{36}{48} \cdot \frac{35}{47} \cdot \frac{34}{46} \cdot \frac{33}{45} \cdot \frac{32}{44} \cdot \frac{31}{43} =$$

Consider next the following combination:

A	B	C	D	E	F	G	H	I	J
S	S	0	0	0	0	0	0	0	0

This is a combination containing two spades. The probability of this particular combination is equal to the probability of a spade in the first pack (that is, $\frac{13}{52}$), times the probability of a spade in a deck from which a spade has been drawn (that is, $\frac{12}{51}$), times the probability of a nonspade in a deck from which two spades have been drawn (that is, $\frac{39}{50}$), times the probability of a nonspade in a deck from which two spades and a nonspade have been drawn (that is, $\frac{38}{49}$), etc. In short, the probability of this particular combination is

$$\frac{13}{52} \cdot \frac{12}{51} \cdot \frac{39}{50} \cdot \frac{38}{49} \cdot \frac{37}{48} \cdot \frac{36}{47} \cdot \frac{35}{46} \cdot \frac{34}{45} \cdot \frac{33}{44} \cdot \frac{32}{43}$$

There are $C_2^{10} = 45$ different combinations, however, containing only two spades, and the probability of each can be shown to be the same as that just computed. Hence the probability of any

one of these 45 combinations, *i.e.*, the probability of two spades, is

$$45 \cdot \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{10}{50} \cdot \frac{9}{49} \cdot \frac{8}{48} \cdot \frac{7}{47} \cdot \frac{6}{46} \cdot \frac{5}{45} \cdot \frac{4}{44} \cdot \frac{3}{43}$$

A continuation of this same line of reasoning shows that in general the probability of N_1 spades out of 10 is

$$P(N_1) = C_{N_1}^{10} \cdot \frac{(13)(12) \cdots (13 - N_1 + 1)(39)(38) \cdots (39 - 10 + N_1 + 1)}{(52)(51)(50)(49)(48)(47)(46)(45)(44)(43)}$$

This yields the distribution shown in Table 11.

TABLE 11.—PROBABILITIES OF THE VARIOUS POSSIBLE NUMBERS OF SPADES AMONG ALL THE COMBINATIONS OF 10 CARDS EACH THAT MIGHT BE MADE OF THE CARDS IN AN ORDINARY PLAYING DECK

Combinations Containing	Probability
0 spade	<u>40,186</u> 1,000,000
1 spade	<u>174,140</u> 1,000,000
2 spades	<u>303,340</u> 1,000,000
3 spades	<u>278,070</u> 1,000,000
4 spades	<u>147,460</u> 1,000,000
5 spades	<u>46,840</u> 1,000,000
6 spades	<u>8,922</u> 1,000,000
7 spades	<u>991</u> 1,000,000
8 spades	<u>60</u> 1,000,000
9 spades	<u>2</u> 1,000,000
10 spades	<u>0.02</u> 1,000,000

If each of N packs contains S cards, $p_1 S$ of which are spades and $p_2 S$ other suits, $p_1 + p_2$ being equal to 1, and if all possible combinations of N cards each are formed by selecting a different card from each pack, the probability of N_1 spades will be given by the general equation

$$P(N_1) = C_{N_1}^N \frac{(p_1 S)(p_1 S - 1) \cdots (p_1 S - N_1 + 1)(p_2 S) \cdots (p_2 S - N + N_1 + 1)}{(S)(S - 1) \cdots (S - N + 1)} \quad (4)$$

This expression is a term in a hypergeometrical series,¹ and the distribution of Table 11 may be called a hypergeometrical distribution. A somewhat simpler equation that is equivalent to Eq. (3)* is as follows,

$$P(N_1) = \frac{(p_1 S)!(p_2 S)!(S - N)!N!}{(p_1 S - N_1)!(p_2 S - N + N_1)!S!N_1!(N - N_1)!} \quad (4')$$

In conclusion, it is to be noted that the percentages of 0, 1, 2, . . . , N_1 spades among combinations formed by selecting a different card from each of N packs are the same as the percentages of 0, 1, 2, . . . , N_1 spades among combinations of N cards from a single pack. That these two problems are the same and have the same solution may be shown as follows: In selecting the first card to form a combination, the situation is obviously the same for both problems, since pack 1 of the former problem and the original pack in the new problem are both complete packs. In selecting the second card, the situation continues to be the same for both problems. For, in the former problem, the cards eligible for selection from pack 2 are, according to the terms of the problem, all the cards except the card that matches the one drawn from pack 1; in the present problem the cards eligible for selection are by necessity only the cards left in the pack after the first card has been selected. Similarly, in selecting the third, fourth, fifth, or N th card, the situation is the same for both

¹ As N_1 goes from 0 to N , this equation generates the series

$$\frac{(p_2 S) \cdots (p_2 S - N + 1)}{(S)(S - 1) \cdots (S - N + 1)} \left[\left\{ 1 + \frac{(N)(p_1 S)}{(p_2 S - N + 1)} + \frac{N(N - 1)}{2!} \frac{(p_1 S)(p_1 S - 1)}{(p_2 S - N + 1)(p_2 S - N + 2)} + \cdots \right\} \right]$$

A general hypergeometrical series is of the form

$$1 + \frac{(a)(b)}{1(c)} x + \frac{(a)(a + 1)(b)(b + 1)}{2!(c)(c + 1)} x^2 + \cdots$$

The part of the former expression in brackets forms a hypergeometrical series in which $x = 1$, $a = -N$, $b = -p_1 S$, and $c = p_2 S - N + 1$.

* This is obtained from Eq. (3) by multiplying numerator and denominator by $(p_1 S - N_1)!(p_2 S - N + N_1)!(S - N)!(N - N_1)!$

problems. Hence it follows that the two problems are identical in all respects and therefore have the same solution.

In general, therefore, if all possible combinations of N cards are made from a deck of S cards, in which the percentage of spades is p_1 and the percentage of nonspades is p_2 , where $p_1 + p_2 = 1$,

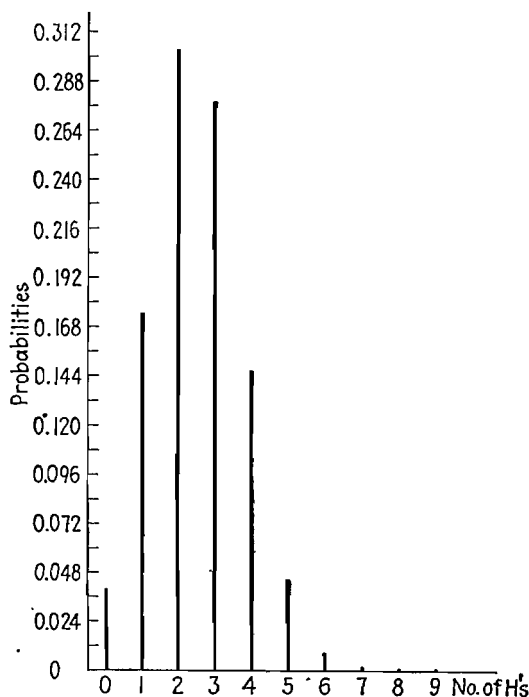


FIG. 13.—A hypergeometrical distribution, $S = 52$, $N = 10$, $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{2}$ (see Table 11).

the relative frequencies of combinations containing N_1 spades are those given by Eq. (4). This conclusion is of importance in certain types of sampling problems.¹

Character of the Hypergeometrical Distribution. It will be noted from Fig. 13 that the distribution of Table 11 is skewed. In general, the mean of the distribution is at Np_1 and the various moments about the mean are²

¹ See pp. 209-211.

² See PEARSON, KARL, "On the Curves Which Are Most Suitable for Describing the Frequency of Random Samples of a Population," *Biometrika*, Vol. 5 (1906-1907), pp. 172-175.

$$\left. \begin{aligned} u_2 &= N p_1 p_2 \left(1 - \frac{N-1}{S-1} \right) \\ u_3 &= N p_1 p_2 (p_2 - p_1) \left(1 - \frac{N-1}{S-1} \right) \left[1 - \frac{2(N-1)}{S-2} \right] \\ u_4 &= N p_1 p_2 \left(1 - \frac{N-1}{S-1} \right) \left\{ 1 - \frac{6(N-1)}{S-2} \left(1 - \frac{N-2}{S-3} \right) \right. \\ &\quad \left. + 3 p_1 p_2 (N-2) \left[1 - \frac{N-1}{S-2} \left(\frac{N-10}{N-2} + \frac{9}{S-3} \right) \right] \right\} \end{aligned} \right\} \quad (5)$$

It is to be noted, that the sign of u_3 is negative when $2N > S$ or $N/S > \frac{1}{2}$.

These equations show that, if p_1 does not equal p_2 , the hypergeometrical distribution is generally asymmetrical. If p_2 is greater than p_1 , it will be skewed positively; and if p_1 is greater than p_2 , it will be skewed negatively. If $p_2 = p_1$, the distribution will be symmetrical. The equations also show that the greater the difference between p_2 and p_1 and hence the smaller the product $p_1 p_2$, the larger¹ will be the coefficient of kurtosis $\beta_2 = u_4/u_2^2$.

If S is made indefinitely large relative to N , *i.e.*, if the number of cards in each pack is made indefinitely large relative to the number of cards selected to form a combination, then $p_1 S$, $p_1 S - 1$, . . . , $p_1 S - N_1$ all become practically equivalent to $p_1 S$; similarly, $p_2 S$, $p_2 S - 1$, . . . , $p_2 S - N + N_1$ become practically equivalent to $p_2 S$, and S , $S - 1$, $S - 2$, . . . , $S - N + 1$ become practically equivalent to S . Hence, as S is made indefinitely large relative to N , Eq. (3) becomes practically the same as Eq. (1) and the hypergeometrical distribution reduces to the asymmetrical binomial distribution (or to the symmetrical binomial distribution if $p_1 = p_2$).* This is also shown by the fact that when S is increased relative to N the equations for $\beta_1 = u_3^2/u_2^3$ and $\beta_2 = u_4/u_2^2$ of the hypergeometrical distribution (see Eq. 5) approach the equations for β_1 and β_2 of the asymmetrical binomial distribution [Eqs. (2)]. Such a

¹ For β_2 reduces to the form

$$\beta_2 = \frac{3r}{t} + \frac{1 - 6m - 6p_1 p_2 r}{N p_1 p_2 t}$$

where m , r , and t are constants depending only on N and S .

* This assumes that S becomes so large that not only S but also $p_1 S$ is very large relative to N .

relationship is logically to be expected; for if the size of each pack of cards is very large relative to the number of cards making up a combination, the probability of a spade in the pack is not much changed in the course of forming combinations. That is, the problem reduces to the one considered in the previous section. Viewed in this manner the binomial distribution is a special case of the hypergeometrical distribution.

Hypergeometrical Distribution and the Pearsonian System of Frequency Curves. The hypergeometrical distribution forms the basis of the Pearsonian system of frequency curves. Having noted that the normal curve had the same relative slope at various points as the symmetrical binomial distribution and that Pearson's type III curve had the same relative slope at various points as the symmetrical binomial distribution, Karl Pearson raised the question: What curve has the same relative slope at various points as the hypergeometrical distribution? Such a curve, he contended, would be a generalized frequency curve; for in the derivation of the hypergeometrical distribution no assumption was made as to the equality of the probabilities nor was any assumption made as to independence.

Now it can be shown¹ that the relative slope of the hypergeometrical distribution (*i.e.*, the relative slope of the frequency polygon that the distribution forms) is given for various mid-points, $X = N_1 + \frac{1}{2}$, by the expression

$$\text{Relative slope} = \frac{X + a}{b_0 + b_1X + b_2X^2} \quad (6)$$

where a , b_0 , b_1 , and b_2 depend on the values of S , N , p_1 , and p_2 of the hypergeometrical distribution. Equation (6), therefore, was taken by Pearson as a general equation capable of representing any frequency distribution.

The characteristics of a particular curve represented by this equation will depend on the values of a , b_0 , b_1 , and b_2 . The parameter a is of significance in determining the position of the curve. For the slope of the curve is zero at its peak (assuming the curve to be a smooth one), and the mode of the curve is therefore given by $X + a = 0$. That is, the mode comes at $X = -a$. The other parameters determine the general shape

¹ See Appendix to this chapter (p. 79).

of the curve. Thus, if the denominator of Eq. (6) is factored into the product of two expressions, viz.,

$$\left(X - \frac{-b_1 + \sqrt{b_1^2 - 4b_0b_2}}{2b_2}\right) \left(X - \frac{-b_1 - \sqrt{b_1^2 - 4b_0b_2}}{2b_2}\right)$$

it is seen that the form of the curve will depend on the value of $b_1^2 - 4b_0b_2$. This it will be noted, is the so-called "discriminant" of the equation $b_0 + b_1X + b_2X^2 = 0$, since its sign determines the nature of its roots. The discriminant may also be written

$4b_0b_2 \left(\frac{b_1^2}{4b_0b_2} - 1 \right)$. Accordingly, Karl Pearson took $b_1^2/4b_0b_2$ as

the criterion of curve form.¹ On the basis of this criterion, he distinguished three main types of curves. If the criterion is negative, i.e., if the roots of the equation $b_0 + b_1X + b_2X^2 = 0$ are real and of opposite sign,² the curve belongs to the class of curves distinguished by Pearson as type I. If the criterion is positive but less than unity, i.e., if the roots of the equation $b_0 + b_1X + b_2X^2 = 0$ are imaginary, the curve belongs to the class of curves distinguished by Pearson as type IV. Finally, if the criterion is positive but greater than unity, i.e., if the roots of the equation $b_0 + b_1X + b_2X^2 = 0$ are real and of the same sign,

¹ For the correct use of the Pearsonian criterion, X should be measured from the mean.

² Equation (6) is more general than the hypergeometrical distribution from which it was derived, for the constants of a hypergeometrical distribution will never give rise to a type I curve (see Appendix to this chapter, p. 81). Having been suggested by the hypergeometrical distribution, Eq. (6) was taken by Pearson as a general frequency equation in which the constants could have any values whatsoever, whether or not they satisfied the requirements of a hypergeometrical distribution. Other considerations suggested the reasonableness of this. For example, if the causes of variation in X remained the same over the whole range they would generate a normal distribution, for which the relative slope would be $-x/\sigma^2$. If, however, the causes of variation varied with X , then σ would become a function of X , and the relative slope might be written $-\frac{x}{\sigma(X)}$. If the function $\sigma(X)$ is expanded in a power series (Taylor's expansion) in the neighborhood of the mean of X and the first three terms of this expansion are taken as a good approximation to the function, the equation for the relative slope becomes

$-\frac{x}{b_0 + b_1x_1 + b_2x^2}$, which is essentially the same as Eq. (6). (Cf. PEARSON, KARL, "Das Fehlgengesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder," *Biometrika*, Vol. 4 (1905-1906), p. 204.)

the curve belongs to the class of curves distinguished by Pearson as type VI.*

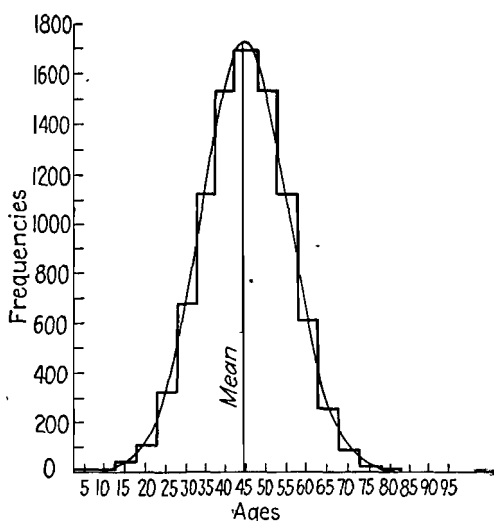


FIG. 14.—Type IV frequency curve. Number exposed to risk of sickness according to Sutton's sickness tables (males, all durations).

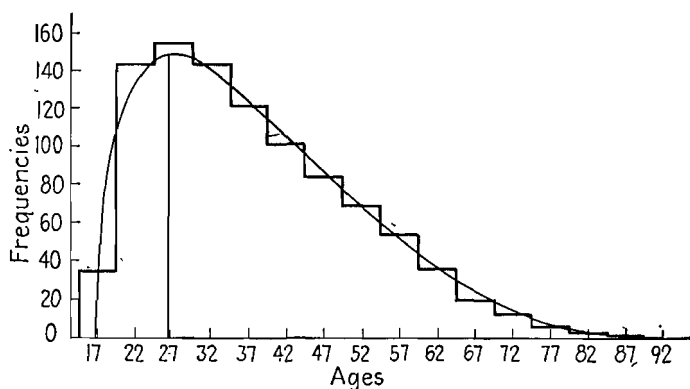


FIG. 15.—Type I frequency curve. Number exposed to risk of sickness. (According to Watson, *M.U. Tables*, p. 19.)

These are the three main classes of Pearsonian frequency curves. A type I curve is shown in Fig. 15. Curves of this type are limited in range. They are usually bellshaped, although skewed, but may under special circumstances be even U shaped,

* See ELDERTON, *op. cit.*, pp. 42-43. Figures 14, 15, and 16 are reproduced by permission of the author and publishers of this book, 2d ed. (1927), pp. 62, 69, 77.

J shaped, or twisted J shaped. Figure 14 shows a type IV curve. These curves are unlimited in range, bell shaped, and skewed. Finally, Fig. 16 shows a type VI curve. These are unlimited in range in one direction; they are usually bell shaped and skewed but may also be J shaped.¹

Various "transitional" types of curves have also been distinguished depending on special values of the b 's. For example,

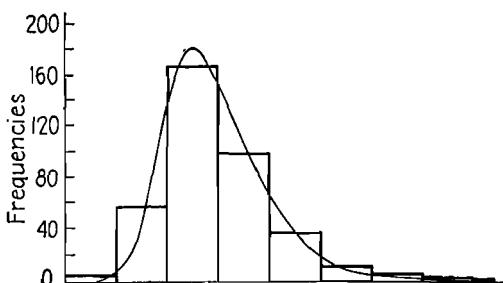


FIG. 16.—Type VI frequency curve. Number of entrants, limited-payment policies, 1863-1893; experience summed in groups of 10 years of age and divided by 100.

if $b_2 = 0$ and the criterion is accordingly infinite, the result is Pearson's type III curve discussed above.² If both b_2 and b_1 are zero, the result is the normal frequency curve. A table of criterion values and curve types is given on page 135.

General Significance of the Pearsonian System of Frequency Curves. *Explanation of Nonnormal Distributions.* The foregoing analysis is of considerable significance in explaining the conditions that give rise to various types of nonnormal frequency distributions. To illustrate this, consider once again the flour-bag experiment described in the previous chapter.³ Suppose as before that the experimenter has a large number of bags, each weighing exactly 5 pounds. Suppose, however, that the amount added or subtracted from each bag is now dependent upon the throwing of 10 prisms, each of which has three faces marked with a T and one with an H. As previously, the experimenter adds an ounce of flour to a bag for each H that faces down and subtracts an ounce for each T that faces down. Under the assumptions that the prisms are thrown in a random fashion,

¹ See *ibid.*, chart opposite p. 46.

² See pp. 47-50.

³ Cf. pp. 36-37. Also see SMITH and DUNCAN, *op. cit.*, pp. 292-293.

intuition suggests that a frequency distribution of the weights of a large number of bags will approximate the form of an asymmetrical binomial distribution for which $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$. It would be like the relative frequency distribution presented in Table 12.

TABLE 12.—FREQUENCY DISTRIBUTION OF A LARGE NUMBER OF BAGS OF FLOUR OF SPECIFIED WEIGHTS

Weight of Bag	Relative Frequency
	59,049
4 lb. 6 oz.	<u>1,048,576</u>
	196,830
4 lb. 8 oz.	<u>1,048,576</u>
	295,245
4 lb. 10 oz.	<u>1,048,576</u>
	262,440
4 lb. 12 oz.	<u>1,048,576</u>
	153,090
4 lb. 14 oz.	<u>1,048,576</u>
	61,236
5 lb. 0 oz.	<u>1,048,576</u>
	17,010
5 lb. 2 oz.	<u>1,048,576</u>
	3,240
5 lb. 4 oz.	<u>1,048,576</u>
	405
5 lb. 6 oz.	<u>1,048,576</u>
	30
5 lb. 8 oz.	<u>1,048,576</u>
	1
5 lb. 10 oz.	<u>1,048,576</u>

The general shape of this distribution can be represented by a Pearsonian type III curve. Owing in this particular case to the greater probability of subtracting an ounce than of adding an ounce, the mean weight of the bags will be less than the original "central value" of 5 pounds, and the distribution will be skewed positively. If the prisms had had more H than T faces, the probability of adding an ounce would have exceeded the probability of subtracting an ounce, the mean weight would have been greater than 5 pounds, and the distribution of weights would have been skewed negatively.

If, instead of adding or subtracting flour in accordance with the number of H's and T's appearing on the throws of 10 prisms,

the additions and subtractions had been made with reference to the number of spades and the number of other cards found among 10 cards drawn without replacement from a pack of 52 playing cards, then the distribution of weights would have tended to conform to a hypergeometrical distribution and it could have been represented by one of the Pearsonian curves given by Eq. (6). In the particular case in hand, the relative frequencies with which bags of 4 pounds 6 ounces, bags of 4 pounds 8 ounces, etc., would have tended to occur would have been those given in Table 11, and the general shape of the distribution would have been described by a Pearsonian curve of type IV.* Since the probability of adding the initial ounce (*i.e.*, the probability of a spade on the first draw) is again less than the probability of subtracting the initial ounce (*i.e.*, the probability of a nonspade on the first draw), the distribution of weights is again skewed positively and its mean is less than the central value of 5 pounds. If an ounce had been added when a nonspade was drawn and subtracted when a spade was drawn, then the distribution of weights would have been skewed negatively and the mean weight would have been greater than 5 pounds. If the probability of adding the initial ounce had been equal to the probability of subtracting it, as would have been the case if an ounce were added whenever a black card appeared and subtracted whenever a red card appeared, then the distribution of weights would have been symmetrical about the normal value of 5 pounds, but the kurtosis of the distribution would have been less than 3, that is, it would have been less peaked than a normal frequency distribution.¹

Thus, whenever the deviation of a variable from some central value is the algebraic sum of the contributions of a number of causal factors, the absolute size of each contribution being the same, then if (1) the contributory causes are independent of each other, but the probability of a positive contribution is greater or less than the probability of a negative contribution, or (2) the contributory causes are not independent of each other, the resulting distribution of the variable will tend to be skewed in the direction for which the average probability is the smaller and will tend to be more or less peaked than the normal curve.

* See Appendix to this chapter (p. 81).

¹ Cf. p. 56.

Although this conclusion as to the causes of skewness and kurtosis in homogeneous variates would appear to be generally applicable, one point needs to be further considered. The asymmetrical binomial distribution and the hypergeometrical distribution, upon which this conclusion was based, are both discrete distributions. The number of H's and the number of spades can vary only by integral units. The weights of the various bags of flour derived from either the binomial or the hypergeometrical distribution differed exactly by multiples of 2 ounces; there were no bags of fractional weights. Although the foregoing conclusion would therefore appear to be a valid explanation of the distributions of discrete variates, such as the number of veins in a leaf, the number of petals on a flower, and the number of individuals in a litter, its validity as an explanation of the distributions of continuous variates, *i.e.*, of frequency curves proper, has yet to be established. This will now be considered.

Difficulty in Explaining Nonnormal Curves. In the case of the symmetrical binomial distribution, it was assumed that if the number of contributory causes was made infinitely large, and if each contribution was made infinitesimally small, the resulting fluctuations in the variate would become practically continuous. In that instance, it was found that the distribution of the continuous variate was of the form of the normal frequency curve.

If the same method of attaining continuity, however, is applied to the asymmetrical binomial and to the hypergeometrical distributions, the analysis immediately runs into difficulties. For, as already pointed out,¹ if the number of prisms is made very large, the asymmetrical binomial distribution diminishes in skewness; and as N approaches infinity, the distribution approaches the symmetrical form. The same thing is true of the hypergeometrical distribution. If the number of cards in a pack is made larger and larger, the proportion of spades always remaining the same, and if the number of cards making up a combination is also increased, the hypergeometrical distribution likewise diminishes in skewness and its coefficient of kurtosis approaches 3.* The consequence is that, if continuity of a variate is the

¹ See p. 46.

* This is shown by Eq. (5). For $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ and if both S and N are

result of the number of contributory causes being infinitely large while their individual effects are infinitesimally small, the distribution of the variate is of the form of the normal frequency curve and the foregoing analysis fails to explain nonnormal curves.

Karl Pearson recognized this difficulty in his analysis and sought to meet it as follows: That distributions of continuous biological and physical variates were nonnormal was a fact that had been demonstrated by empirical investigation. "It is clear," argued Pearson, "that if such frequency curves . . . are to be treated as chance distributions at all, it would be idle to compare them to the limit of a symmetrical binomial. We are really quite ignorant as to the nature of the contributory 'causes' in biological, physical, or economic frequency curves. The continuity of such frequency curves may depend upon other features than the magnitude of n It may possibly be that continuity in biological or physical frequency curves may arise from a limited number of 'contributory causes' with a power of fractionizing the result."¹ For example, in the flour-bag experiment, the amount of flour added or subtracted in each case could be taken as a handful instead of an exact ounce. In such a case, the distribution of the weights of the bags of flour would tend to form a smooth curve even when the number of prisms used was small.²

Thus, in proceeding by argument from the discrete symmetrical binomial to the continuous normal curve, the element of discreteness is blurred and finally obliterated by making the number of causes infinitely large and decreasing the contribution of each cause. The same argument applied to the asymmetrical

increased, such terms as $\frac{N-1}{S-1}$ are little changed; but the N that appears in the denominator when μ_3^2 is divided by μ_1 is uncompensated for, and this causes β_1 to get smaller and smaller as N is increased. The same is true for $\beta_2 = \frac{\mu_4}{\mu_2^2}$; for as N is increased, all terms in the ratio except 3 reduce to 0.

¹ "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London*, Series A, Vol. 186 (1895), p. 358.

² Cf. PEARSON, KARL, "Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder," *Biometrika*, Vol. 4 (1905-1906), p. 206.

binomial or to the hypergeometric series will also blur the discreteness, but it will, in addition, cause the nonnormality to disappear. In order to retain the nonnormality, it seems necessary to keep the assumption of a small number of contributing causes; the blurring of the discreteness is accomplished by the assumption that the contributions of the contributing causes vary among themselves.

In the Pearsonian analysis of nonnormal frequency curves, then, it is necessary to suppose that nonnormal series in real life are created by the number of causes affecting variability being small and their contributions varying in magnitude. This supposition is not unreasonable when applied, for example, to such things as the distribution of income, wage rates, and other similar economic phenomena that are known to have nonnormal static variability. In these and other economic and social phenomena the factors causing static variability may reasonably be supposed to be not only finite but also comparatively few in number.

APPENDIX

MATHEMATICAL PROOFS

A. Proof That, for Any Binomial Distribution Represented by the Equation $P(N_1) = \frac{N!}{N_1!N_2!} p_1^{N_1} p_2^{N_2}$, the Mean = Np_1 ,

$$\sigma = \sqrt{Np_1p_2},$$

$\beta_1 = \frac{(p_2 - p_1)^2}{Np_1p_2}$, and $\beta_2 = 3 + \frac{1 - 6p_1p_2}{Np_1p_2}$ [Since Eqs. (3) of Chap. II are merely a special case of the foregoing for which $p_1 = p_2 = \frac{1}{2}$, the following is also a proof of them.]

By definition the mean of any distribution of probability is equal to $\sum P(X)X$ [Chap. 1, Eq. (2)]. Hence the mean of the binomial distribution is

$$\sum P(N_1)N_1 = \sum \frac{N!}{N_1!N_2!} p_1^{N_1} p_2^{N_2} N_1$$

where the summation is for N_1 from 0 to N .

If N_1 is canceled out of the numerator and out of $N_1!$ in the denominator, however, and if Np_1 is factored out of the resulting

expression, the value of the mean becomes

$$Np_1 \sum \frac{(N-1)!}{(N_1-1)!(N_2)!} p_1^{N_1-1} p_2^{N_2}$$

Also, by definition $N_2 = N - N_1$, which may be written $[N - 1 - (N_1 - 1)]$, and the sum may thus be put in the form

$$\sum \frac{(N-1)!}{(N_1-1)![N-1-(N_1-1)]!} p_1^{N_1-1} p_2^{[N-1-(N_1-1)]}$$

This is recognized,¹ however, to be the expansion of $(p_1 + p_2)^{N-1}$, which equals 1 since $p_1 + p_2 = 1$. Hence the mean of the binomial distribution reduces to Np_1 .

The second moment of the binomial distribution about the origin is, by definition, $\Sigma P(N_1)N_1^2$. But N_1^2 may be written $N_1^2 = N_1(N_1 - 1) + N_1$. Hence

$$\begin{aligned} \Sigma P(N_1)N_1^2 &= \Sigma P(N_1)[N_1(N_1 - 1) + N_1] \\ &= \Sigma P(N_1)N_1(N_1 - 1) + \Sigma P(N_1)N_1. \end{aligned}$$

The second term of this has just been seen to be equal to Np_1 , and the first, when written in full, is

$$\sum \frac{N!}{N_1!N_2!} p_1^{N_1} p_2^{N_2} N_1 (N_1 - 1)$$

If, as in the case of the mean, the terms $N_1(N_1 - 1)$ in the numerator are canceled against the first two factors of $N_1!$ in the denominator and if the quantity $N(N-1)p_1^2$ is factored out of the resulting expression, the above becomes

$$N(N-1)p_1^2 \sum \frac{(N-2)!}{(N_1-2)!N_2!} p_1^{N_1-2} p_2^{N_2}$$

The quantity N_2 , however, which equals $N - N_1$, may be written $[N - 2 - (N_1 - 2)]$; and when this is done, the sum is recognized as the expansion of $(p_1 + p_2)^{N-2}$, which equals 1. Hence the second moment about the origin is equal to $N(N-1)p_1^2 + Np_1$, and the second moment about the mean is equal to this minus $(Np_1)^2$ (by the short equation, $\sigma^2 = \Sigma P(X^2) - \bar{X}^2$). Therefore,

¹ Cf. p. 43.

$$\begin{aligned}\sigma^2 &= N(N-1)p_1^2 + Np_1 - N^2p_1^2 = Np_1 - Np_1^2 \\ &= Np_1(1-p_1) = Np_1p_2,\end{aligned}$$

and $\sigma = \sqrt{Np_1p_2}$.

Again the third moment about the origin is equal to $\Sigma P(N_1)N_1^3$. But N_1^3 may be written

$$N_1^3 \equiv N_1(N_1-1)(N_1-2) + 3N_1(N_1-1) + N_1,$$

so that

$$\begin{aligned}\Sigma P(N_1)N_1^3 &= \Sigma P(N_1)N_1(N_1-1)(N_1-2) \\ &\quad + \Sigma 3P(N_1)N_1(N_1-1) + \Sigma P(N_1)N_1.\end{aligned}$$

The second and third terms of this have already been proved to be equal to $3N(N-1)p_1^2$ and Np_1 , respectively. Furthermore, by canceling $N_1(N_1-1)(N_1-2)$ out of the first term, and by factoring out $N(N-1)(N-2)p_1^3$, the sum part of this term is reduced to the equivalent of $(p_1 + p_2)^{N-3}$, which equals 1. The whole expression for the third moment about the origin thus becomes $\nu_3 = N(N-1)(N-2)p_1^3 + 3N(N-1)p_1^2 + Np_1$. By making use of Eq. (6), Chap. I, for converting moments about an arbitrary origin to moments about the mean, the third moment about the mean of the binomial distribution is found to be

$$\begin{aligned}\mu_3 &= N(N-1)(N-2)p_1^3 + 3N(N-1)p_1^2 + Np_1 - 3N^3p_1^3 \\ &\quad - 3N^2p_1^2(1-p_1) + 2N^3p_1^3 \\ &= Np_1(2p_1^2 - 3p_1 + 1) = Np_1(1-p_1)(1-2p_1) \\ &\quad = Np_1p_2(p_2 - p_1)\end{aligned}$$

Finally, $\beta_1 = \mu_3/\mu_2^3$ has the value

$$\beta_1 = \frac{(p_2 - p_1)^2}{Np_1p_2}$$

In precisely the same way the fourth moment about the mean can be shown to be equal to

$$\mu_4 = 3(p_1p_2N)^2 + p_1p_2N(1-6p_1p_2)$$

Since $\beta_2 = \mu_4/\mu_2^2$, it follows that

$$\beta_2 = \frac{3(Np_1p_2)^2 + Np_1p_2(1-6p_1p_2)}{(p_1p_2N)^2}$$

or

$$\beta_2 = 3 + \frac{1-6p_1p_2}{Np_1p_2}$$

B. Proof That the Mode of a Binomial Distribution Lies between $Np_1 - p_2$ and $Np_1 + p_1$. If the ordinate of the binomial distribution at N_1 is to be the modal ordinate (*i.e.*, the highest) it must satisfy the following criterion:

$$Y_{N_1-1} \leq Y_{N_1} \geq Y_{N_1+1}$$

or

$$\frac{N! p_1^{N_1-1} p_2^{N-N_1+1}}{(N_1-1)!(N-N_1+1)!} \leq \frac{N! p_1^{N_1} p_2^{N-N_1}}{N_1!(N-N_1)!} \\ \geq \frac{N! p_1^{N_1+1} p_2^{N-N_1-1}}{(N_1+1)!(N-N_1-1)!}$$

On taking out the common factor

$$\frac{N! p_1^{N-1} p_2^{N-N_1-1}}{(N_1-1)!(N-N_1-1)!}$$

these inequalities become

$$\frac{p_2^2}{(N-N_1+1)(N-N_1)} \leq \frac{p_1 p_2}{N_1(N-N_1)} \geq \frac{p_1^2}{(N_1+1)(N_1)}$$

From the first inequality it follows that

$$N_1 p_2 \leq (N - N_1 + 1) p_1$$

or that $N_1(p_1 + p_2) \leq Np_1 + p_1$, or, since $p_1 + p_2 = 1$, that $N_1 \leq Np_1 + p_1$. From the second inequality it follows that $p_2(N_1 + 1) \geq p_1(N - N_1)$ or that $N_1(p_1 + p_2) + p_2 \geq Np_1$ or that $N_1 \geq Np_1 - p_2$. Hence, in summary, if N_1 is to be the mode it must be the integer lying between $Np_1 - p_2$ and $Np_1 + p_1$.

C. Proof That the Binomial Distribution Is Approximated by the Normal Curve. The general equation for the binomial distribution is

$$P(N_1) = \frac{N!}{N_1! N_2!} p_1^{N_1} p_2^{N_2} \quad (1)$$

where $N_2 = N - N_1$ and $p_1 + p_2 = 1$ (*cf.* page 43). The equation for the symmetrical binomial distribution (*cf.* page 33) is the special case for which $p_1 = p_2 = \frac{1}{2}$. The following analysis shows that the general binomial equation can be approximately represented by the equation for the normal curve, the degree of approximation depending on the size of N . The analysis thus includes the special case of the symmetrical binomial distribution as well as the more general asymmetrical binomial distribution.

If the variable N_1 is replaced by $x = N_1 - Np_1$, that is, by the deviation of N_1 from its mean, the general equation (1) becomes

$$P(x) = \frac{N!}{(Np_1 + x)!(Np_2 - x)!} p_1^{Np_1+x} p_2^{Np_2-x} \quad (2)$$

This gives the probability of x cases more or less than the mean number. The mean¹ of this distribution of x is, of course, 0, but its standard deviation² is the same as that of N_1 , viz., $\sqrt{Np_1p_2}$, its β_1 is likewise $\frac{(p_2 - p_1)^2}{Np_1p_2}$, and its $\beta_2 = 3 + \frac{1 - 6p_1p_2}{Np_1p_2}$.

For simplicity assume that Np_1 and Np_2 , and hence x , are integers. This assumption does not materially affect the results since the error involved is of order $1/N$ and the subsequent argument will assume that N is so large that terms of order $1/N$ or higher order may reasonably be neglected. In other words, this assumption is good enough for the degree of approximation given by the final equation.

Expression (2) may be written

$$P(x) = P(0) \frac{(Np_1)!(Np_2)!}{(Np_1 + x)!(Np_2 - x)!} p_1^x p_2^{-x} \quad (3)$$

where

$$P(0) = \frac{N!}{(Np_1)!(Np_2)!} p_1^{Np_1} p_2^{Np_2}$$

or the value of $P(x)$ when $x = 0$, its mean value. To Eq. (3)

¹ Since $x = N_1 - Np_1$, the mean of x is

$$\Sigma P(x)x = \Sigma P(N_1)N_1 - Np_1 \Sigma P(N_1).$$

But $\Sigma P(N_1)N_1 = Np_1$, the mean of N_1 (cf. p. 66), and $\Sigma P(N_1) = 1$. Hence the mean of x is $Np_1 - Np_1 = 0$.

² Since the mean of x is zero, the variance of x , that is, the square of its standard deviation, equals

$$\begin{aligned} \Sigma x^2 P(x) &= \Sigma (N_1 - Np_1)^2 P(N_1) \\ &= \Sigma N_1^2 P(N_1) - 2Np_1 \Sigma N_1 P(N_1) + N^2 p_1^2 \Sigma P(N_1). \end{aligned}$$

But $\Sigma N_1 P(N_1)$ is the mean of N_1 and equals Np_1 . Also, $\Sigma P(N_1) = 1$. Hence $\Sigma N_1^2 P(N_1) - 2Np_1 \Sigma N_1 P(N_1) + N^2 p_1^2 \Sigma P(N_1) = \Sigma N_1^2 P(N_1) - N^2 p_1^2$, which by the short equation (see p. 67) is the variance of N_1 . Therefore the standard deviation of x is the same as that of N_1 .

A similar argument can be used to show that the third and fourth moments of x about its mean are the same as the third and fourth moments of N_1 about its mean and hence that it has the same β_1 and β_2 .

Stirling's approximation for factorials may be applied. This is¹

$$a! \doteq a^a e^{-a} \sqrt{2\pi a}$$

which is correct to $1/a$ or as good an approximation as is being sought. Making use of this, Eq. (3) becomes

$$\frac{P(x)}{P(0)} = \frac{(N\mathbf{p}_1)^{N\mathbf{p}_1} e^{-N\mathbf{p}_1} \sqrt{2\pi N\mathbf{p}_1} (N\mathbf{p}_2)^{N\mathbf{p}_2} e^{-N\mathbf{p}_2} \sqrt{2\pi N\mathbf{p}_2} \mathbf{p}_1^x \mathbf{p}_2^{-x}}{(N\mathbf{p}_1 + x)^{N\mathbf{p}_1 + x} e^{-N\mathbf{p}_1 - x} \sqrt{2\pi(N\mathbf{p}_1 + x)} (N\mathbf{p}_2 - x)^{N\mathbf{p}_2 - x} e^{-N\mathbf{p}_2 + x} \sqrt{2\pi(N\mathbf{p}_2 - x)}}$$

which reduces to²

$$\frac{P(x)}{P(0)} = \frac{1}{\left(1 + \frac{x}{N\mathbf{p}_1}\right)^{N\mathbf{p}_1 + x + \frac{1}{2}} \left(1 - \frac{x}{N\mathbf{p}_2}\right)^{N\mathbf{p}_2 - x + \frac{1}{2}}} \quad (4)$$

A final step is to take logarithms and then expand these in a power series.³ Thus,

¹ More exactly Stirling's formula says that, as a approaches infinity,

$$a^a e^{-a} \sqrt{2\pi a} < a! < a^a e^{-a} \sqrt{2\pi a} \left(1 + \frac{1}{4a}\right)$$

Accordingly, $1 + \frac{1}{4a}$ gives an estimate of the degree of accuracy of the approximation. This formula may be demonstrated briefly by evaluating the area under the curve $y = \log x$ between ordinates $x = 1$ and $x = n$, which gives, by integration,

$$\text{Area} = \int_1^n \log x \, dx = x \log x - x \Big|_1^n = n \log n - n + 1$$

An approximate estimate of this same area is obtained, however, by the trapezoid formula [cf. R. Courant, *Differential and Integral Calculus* (1940), Vol. I, p. 343], erecting ordinates at $x = 1, x = 2, x = 3, \dots, x = n$. This approximate estimate gives the

$$\begin{aligned} \text{Area} &\doteq \frac{1}{2} \log 1 + \log 2 + \log 3 + \dots + \log(n-1) + \frac{1}{2} \log n \\ &= \log n! - \frac{1}{2} \log n, \text{ since } \log 1 = 0. \end{aligned}$$

Consequently, $\log n! - \frac{1}{2} \log n \doteq n \log n - n + 1$, and hence

$$\log n! \doteq \left(n + \frac{1}{2}\right) \log n - n + 1;$$

so that $n! \doteq n^{n+\frac{1}{2}} e^{-n+1} \doteq n^n e^{-n} 2.71 \sqrt{n}$, which is very close to Stirling's formula, $n^n e^{-n} \sqrt{2\pi} \sqrt{n}$. For a more precise derivation of Stirling's formula, see *ibid.*, Vol. I, pp. 361-364.

² Accomplished by dividing both numerator and denominator by $(N\mathbf{p}_1)^{N\mathbf{p}_1 + x + \frac{1}{2}} (N\mathbf{p}_2)^{N\mathbf{p}_2 - x + \frac{1}{2}}$ and canceling out like terms in numerator and denominator. Note that $\mathbf{p}_1 + \mathbf{p}_2 = 1$.

³ The value of $\log_e(1+a)$ is given approximately by

$$\log_e(1+a) \doteq a - \frac{a^2}{2} + \frac{a^3}{3} - \frac{a^4}{4} + \dots$$

$$\begin{aligned}
\log_e \frac{P(x)}{P(0)} &= - \left(N\mathbf{p}_1 + x + \frac{1}{2} \right) \log_e \left(1 + \frac{x}{N\mathbf{p}_1} \right) \\
&\quad - \left(N\mathbf{p}_2 - x + \frac{1}{2} \right) \log_e \left(1 - \frac{x}{N\mathbf{p}_2} \right) \\
&= - \left(N\mathbf{p}_1 + x + \frac{1}{2} \right) \left[\frac{x}{N\mathbf{p}_1} - \frac{1}{2} \left(\frac{x}{N\mathbf{p}_1} \right)^2 + \frac{1}{3} \left(\frac{x}{N\mathbf{p}_1} \right)^3 - \dots \right] \\
&\quad - \left(N\mathbf{p}_2 - x + \frac{1}{2} \right) \left[- \frac{x}{N\mathbf{p}_2} - \frac{1}{2} \left(\frac{x}{N\mathbf{p}_2} \right)^2 - \frac{1}{3} \left(\frac{x}{N\mathbf{p}_2} \right)^3 - \dots \right]
\end{aligned}$$

Upon multiplying out and arranging in ascending powers of $1/N$ this becomes

$$\begin{aligned}
\log_e \frac{P(x)}{P(0)} &= - \frac{[x^2 + x(\mathbf{p}_2 - \mathbf{p}_1)]}{2N\mathbf{p}_1\mathbf{p}_2} + \frac{2x^3(\mathbf{p}_2^2 - \mathbf{p}_1^2) + 3x^2(\mathbf{p}_2^2 + \mathbf{p}_1^2)}{12N^2\mathbf{p}_1^2\mathbf{p}_2^2} \\
&\quad + \text{terms of order } \frac{1}{N^3\mathbf{p}_1^3\mathbf{p}_2^3} \text{ and terms of higher order} \quad (5)
\end{aligned}$$

Now the equation for the standard deviation of x , *viz.*,

$$\sigma = \sqrt{N\mathbf{p}_1\mathbf{p}_2},$$

shows that variation in x is of order $\sqrt{N}$. In other words, as N increases, the variation in x increases as $\sqrt{N}$. Hence, terms such as $x/N\mathbf{p}_1\mathbf{p}_2$ are of order $\sqrt{N}/N$ or $1/\sqrt{N}$ and terms

for $-1 < a \leq 1$.

To demonstrate this, note that by the integral calculus

$$\log_e (1 + a) = \int_0^a \frac{dt}{1+t}$$

If the division $\frac{1}{1+t}$ is carried out to n terms and the integration carried out term by term, it follows that

$$\log_e (1+a) = a - \frac{a^2}{2} + \frac{a^3}{3} - \frac{a^4}{4} + \dots + (-1)^{n-1} \frac{a^n}{n} + (-1)^n \int_0^a \frac{t^n}{1+t} dt$$

For $a \geq 0$, the integral term is less than $a^{n+1}/(n+1)$ which approaches 0 as n approaches ∞ . If $-1 < a \leq 0$, the absolute value of the integral term is less than or equal to $\frac{|a|^{n+1}}{(1+a)(n+1)}$. (Cf. R. Courant, *op. cit.*, Vol. I, pp. 315-317.) Since the standard deviation of x equals $\sqrt{N\mathbf{p}_1\mathbf{p}_2}$, it follows that x is of order $\sqrt{N}$. Hence $\frac{x}{N\mathbf{p}_1}$ and $\frac{x}{N\mathbf{p}_2}$ are of order $1/\sqrt{N}$, and if N is taken large enough, $|x/N\mathbf{p}_1|$ and $|x/N\mathbf{p}_2|$ will for some value of N become less than 1 for the important part of the distribution lying between $\frac{x}{\sigma} = -3$ and $\frac{x}{\sigma} = 3$, say.

such as $x^2/N^2 p_1^2 p_2^2$ are of order N/N^2 or $1/N$. If in Eq. (5) terms of order $1/N$ and those of higher order are neglected, then

$$\log_e \frac{P(x)}{P(0)} = -\frac{x^2}{2N p_1 p_2} - \frac{x}{2N p_1 p_2} (p_2 - p_1) + \frac{x^3}{6N^2 p_1^2 p_2^2} (p_2^2 - p_1^2)$$

or since $p_2^2 - p_1^2 = (p_2 - p_1)(p_2 + p_1)$ and $p_2 + p_1 = 1$,

$$\log_e \frac{P(x)}{P(0)} = -\frac{x^2}{2N p_1 p_2} - \left(\frac{x}{N p_1 p_2} - \frac{x^3}{3N^2 p_1^2 p_2^2} \right) \frac{p_2 - p_1}{2}$$

On replacing $N p_1 p_2$ by δ^2 , etc., this becomes

$$\log_e \frac{P(x)}{P(0)} = -\frac{x^2}{2\delta^2} - \left(\frac{x}{\delta} - \frac{x^3}{3\delta^3} \right) \frac{p_2 - p_1}{2\delta}$$

or, on taking antilogarithms,

$$P(x) = P(0) e^{-\frac{x^2}{2\delta^2}} e^{-\left(\frac{p_2 - p_1}{2\delta}\right) \left(\frac{x}{\delta} - \frac{x^3}{3\delta^3}\right)} \quad (6)$$

To complete this equation, $P(0)$ must be evaluated. This may be done as follows: By definition, $P(0)$ equals

$$\frac{N!}{(N p_1)! (N p_2)!} p_1^{N p_1} p_2^{N p_2}.$$

Using Stirling's formula,

$$P(0) \doteq \frac{N^N e^{-N} \sqrt{2\pi N} p_1^{N p_1} p_2^{N p_2}}{(N p_1)^{N p_1} e^{-N p_1} \sqrt{2\pi N p_1} (N p_2)^{N p_2} e^{-N p_2} \sqrt{2\pi N p_2}}$$

which reduces to

$$P(0) \doteq \frac{1}{\sqrt{2\pi N p_1 p_2}} = \frac{1}{\delta \sqrt{2\pi}}$$

all within an error of order $1/N$ or $1/\delta^2$.

Equation (6) may thus be rewritten

$$P(x) \doteq \frac{1}{\delta \sqrt{2\pi}} e^{-\frac{x^2}{2\delta^2}} e^{-\frac{p_2 - p_1}{2\delta} \left(\frac{x}{\delta} - \frac{x^3}{3\delta^3}\right)} \quad (7)$$

This gives the value of $P(x)$ within a margin of error of order $1/\delta^2$ or $1/N$. If N is sufficiently large so that terms of order $1/\delta$ or $1/\sqrt{N}$ can reasonably be neglected, Eq. (7) reduces to

$$P(x) = \frac{1}{\delta \sqrt{2\pi}} e^{-\frac{x^2}{2\delta^2}} \quad (8)$$

This is the equation for the normal distribution.

It will be noted that, when $p_1 = p_2$, Eqs. (7) and (8) become identical. The normal curve is thus a better approximation to the symmetrical than to the asymmetrical binomial distribution, as common sense suggests. Equation (8) however, will also give results very close to those of (7) wherever the difference $p_2 - p_1$ is sufficiently small relative to σ . That is, if p_1 and p_2 differ little absolutely and hence the binomial distribution is only slightly skewed even for small values of N , or if N is so large that $\frac{p_1 - p_2}{\sigma} = \frac{p_1 - p_2}{\sqrt{N p_1 p_2}}$ is small, then the normal distribution becomes

a satisfactory "first" approximation even to the asymmetrical binomial distribution. But if N is not very large and if p_1 differs radically from p_2 , then the binomial distribution is rather skewed and the second approximation (7) had better be used. Of course, if N is very small, then it is preferable to use the binomial formula itself, for in this case neglect of terms of order $1/N$ may lead to serious error.

Equation (8) shows that the ordinates of the binomial distribution may be approximated by the ordinates of the normal curve. It will be noted that, as N increases, the ordinates of the normal curve, as well as the ordinates of the binomial distribution (cf. page 34), tend to get smaller and smaller and the curve becomes more spread out. This is because N enters into the formula for σ (that is, $\sigma = \sqrt{N p_1 p_2}$). The larger the value of N , therefore, the larger the standard deviation and, according to Eq. (8), the smaller any particular ordinate (for σ enters into the denominator of this equation). If, however, the scales on which the curve is graphed are varied in proportion to σ , that is, if the vertical scale is lengthened and the horizontal scale is shortened in proportion to σ (that is, $\sqrt{N}$), then the normal curve retains a constant shape, namely, that of the standard normal curve,

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

where $z = x/\sigma$. This is the basis for the statement on page 34 of the text, that if the scales are adjusted in this way then the limit of the binomial distribution as N is increased is the standard normal curve.

The above shows that a binomial ordinate can be estimated by an ordinate of the normal curve and that a sum of binomial

ordinates can be estimated from the sum of certain ordinates of the normal curve. What is done in practice, however, is to estimate a sum of binomial ordinates over a given range from the area under the normal curve for that range, and it remains to be shown that an area under the normal curve is the approximate equivalent of a sum of normal curve ordinates. This is demonstrated as follows:

An ordinate of a normal curve

$$y = \frac{1}{\delta \sqrt{2\pi}} e^{\frac{-x^2}{2\delta^2}}$$

can always be represented by a rectangle of height $\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2\delta^2}}$ and a base $1/\delta$. Furthermore, if the x -scale is measured in standard deviation units, a succession of ordinates will be $1/\delta$ units apart and the succession of rectangles representing them will all touch each other or, in other words, be contiguous. As N is increased and $1/\delta = 1/\sqrt{Np_1p_2}$ is decreased, rectangles over any range will be thinner and their area will approach the area of the standard normal curve for that range. Hence the area under the standard normal curve can be used as an estimate of the sum of a series of normal ordinates and thus of the corresponding series of binomial ordinates.

D. Proof That the Relative Slope of the Symmetrical Binomial Polygon Is the Same at Any Abscissa Mid-point as That of a Normal Curve with the Same Mean as the Binomial Distribution and a Variance Equal to $\frac{N+1}{N}$ Times the Binomial Variance.¹

The ordinate of the symmetrical binomial distribution at any abscissa point N_1 is

$$y_{N_1} = \frac{N!}{N_1!(N - N_1)!} \left(\frac{1}{2}\right)^N$$

and the ordinate at the next abscissa point $N_1 + 1$ is

$$y_{N_1+1} = \frac{N!}{(N_1 + 1)!(N - N_1 - 1)!} \left(\frac{1}{2}\right)^N$$

¹ This and the next two proofs are based upon Karl Pearson's analysis in the *Philosophical Transactions of the Royal Society of London*, Series A, Vol. 186 (1895), pp. 355 ff.

The difference between these ordinates is

$$\Delta y_{N_1} = y_{N_1+1} - y_{N_1} = \frac{N!}{N_1!(N - N_1 - 1)!} \left(\frac{1}{N_1 + 1} - \frac{1}{N - N_1} \right)$$

and since the abscissa interval is $N_1 + 1 - N_1 = 1$, the absolute slope of the side of the polygon joining the tops of these two ordinates is this difference Δy_{N_1} (see Fig. 3, page 4).

The ordinate of the polygon at the abscissa mid-point $N_1 + \frac{1}{2}$ is the average of the two ordinates at the abscissa points N_1 and $N_1 + 1$. Hence the ordinate at this mid-point is

$$\frac{1}{2} (y_{N_1+1} + y_{N_1}) = \frac{N!}{2N_1!(N - N_1 - 1)!} \left(\frac{1}{N_1 + 1} + \frac{1}{N - N_1} \right)$$

The relative slope at the abscissa mid-point $N_1 + \frac{1}{2}$ is defined as the ratio of the absolute slope to the ordinate at that mid-point. Hence the relative slope of the polygon at the abscissa mid-point $N_1 + \frac{1}{2}$ is

$$\begin{aligned} \frac{\Delta y_{N_1}}{y_{N_1+\frac{1}{2}}} &= \frac{y_{N_1+1} - y_{N_1}}{\frac{1}{2}(y_{N_1+1} + y_{N_1})} = \frac{N - N_1 - N_1 - 1}{\frac{1}{2}(N - N_1 + N_1 + 1)} \\ &= \frac{N - 2N_1 - 1}{\frac{1}{2}(N + 1)} = -\frac{2(N_1 - N/2 + \frac{1}{2})}{\frac{1}{2}(N + 1)} \end{aligned}$$

If x is set equal to $N_1 + \frac{1}{2} - \frac{N}{2}$ that is, to $N_1 + \frac{1}{2}$ minus the mean of N_1 , and if k^2 is set equal to $\frac{N+1}{4}$, this expression for the relative slope at $N_1 + \frac{1}{2}$ becomes

$$\frac{\Delta y_{N_1}}{y_{N_1+\frac{1}{2}}} = \frac{-2x}{2k^2}$$

The relative slope of the normal curve, $y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$, at any abscissa point $x = X - \bar{X}$ is

$$\frac{1}{y} \frac{dy}{dx} = \frac{d \log y}{dx} = \frac{d(-\frac{x^2}{2\sigma^2} - \log \sigma \sqrt{2\pi})}{dx} = -\frac{2x}{2\sigma^2}$$

Hence, $-\frac{2x}{2k^2}$ is the relative slope of a normal curve at the point x whose standard deviation $= k = \sqrt{\frac{(N+1)}{4}}$. Since the

standard deviation of the symmetrical binomial distribution is $\sqrt{N/4}$, the normal curve that has the same relative slope as the polygon of a symmetrical binomial distribution at any abscissa mid-point is the one that has the same mean as the mean of the binomial distribution and a variance equal to $\frac{N+1}{N}$ times the variance of the binomial distribution. If N is large, the two variances are practically the same, for then $\frac{N+1}{N}$ is practically equal to 1.

E. Proof That the Relative Slope of the General Binomial Polygon Is the Same at Any Abscissa Mid-point as That of a Pearsonian Type III Curve. The ordinate of the general binomial distribution at any abscissa point N_1 is

$$y_{N_1} = \frac{N!}{N_1!(N - N_1)!} p_1^{N_1} p_2^{N - N_1}$$

and the ordinate at the abscissa point $N_1 + 1$ is

$$y_{N_1+1} = \frac{N!}{(N_1 + 1)!(N - N_1 - 1)!} p_1^{N_1+1} p_2^{N - N_1 - 1}$$

The difference between these ordinates is

$$\Delta y_{N_1} = y_{N_1+1} - y_{N_1} = \frac{N! p_1^{N_1} p_2^{N - N_1 - 1}}{N_1!(N - N_1 - 1)!} \left(\frac{p_1}{N_1 + 1} - \frac{p_2}{N - N_1} \right)$$

and since the distance between the abscissa points $N_1 + 1$ and N_1 is 1, the absolute slope of the side of the polygon joining the tops of these two ordinates is Δy_{N_1} .

The ordinate of the polygon at the abscissa mid-point $N_1 + \frac{1}{2}$ is the average of the two ordinates at the abscissa points N_1 and $N_1 + 1$, that is,

$$y_{N_1+\frac{1}{2}} = \frac{1}{2}(y_{N_1+1} + y_{N_1}) = \frac{N! p_1^{N_1} p_2^{N - N_1 - 1}}{2N_1!(N - N_1 - 1)!} \left(\frac{p_1}{N_1 + 1} + \frac{p_2}{N - N_1} \right)$$

The relative slope at the abscissa mid-point $N_1 + \frac{1}{2}$ is the ratio of the absolute slope to the ordinate at that mid-point. This has the value

$$\begin{aligned}\frac{\Delta y_{N_1}}{y_{N_1+\frac{1}{2}}} &= \frac{y_{N_1+1} - y_{N_1}}{\frac{1}{2}(y_{N_1+1} + y_{N_1})} = \frac{2[(N - N_1)p_1 - (N_1 + 1)p_2]}{(N - N_1)p_1 + (N_1 + 1)p_2} \\ &= \frac{2(Np_1 - N_1 - p_2)}{Np_1 + (p_2 - p_1)N_1 + p_2}\end{aligned}$$

which may be put in terms of the abscissa mid-point $N_1 + \frac{1}{2}$ by adding and subtracting $\frac{1}{2}$ to the numerator and adding and subtracting $\frac{1}{2}(p_2 - p_1)$ in the denominator, as follows:

$$\frac{2[Np_1 - (N_1 + \frac{1}{2}) - p_2 + \frac{1}{2}]}{Np_1 + (p_2 - p_1)(N_1 + \frac{1}{2}) + p_2 - \frac{1}{2}(p_2 - p_1)}$$

If now x is set equal to $(N_1 + \frac{1}{2}) - Np_1 + p_2 - \frac{1}{2}$, that is, to the deviation of $N_1 + \frac{1}{2}$ from what is practically the mode of the binomial distribution (see page 68), the foregoing expression for the relative slope can be put in the form

$$\text{Relative slope} = \frac{-2x}{Np_1 + (p_2 - p_1)(x + Np_1 - p_2 + \frac{1}{2}) + p_2 - \frac{1}{2}(p_2 - p_1)}$$

which, upon making use of the fact that $p_1 + p_2 = 1$, reduces to

$$\begin{aligned}\text{Relative slope} &= \frac{-2x}{2Np_1p_2 + 2p_1p_2 + (p_2 - p_1)x} \\ &= \frac{-\frac{2}{p_2 - p_1}x}{\frac{2p_1p_2(N+1)}{p_2 - p_1} + x} \\ &= \frac{-kx}{a + x}\end{aligned}$$

where¹

$$k = \frac{2}{p_2 - p_1} \quad \text{and} \quad a = \frac{2p_1p_2(N+1)}{p_2 - p_1}$$

To find the curve whose relative slope at the point x is that of the binomial polygon, it is necessary merely to integrate. This is done as follows:

¹ The k and a used in this proof are the same as the boldface k and a used in the body of the text. Italic k and a are used here to simplify the printing of the mathematical derivation; they do not signify sample values but represent population values just as k and a do in the text.

Set

$$\frac{1}{y} \frac{dy}{dx} = \frac{-kx}{a+x}$$

But

$$\frac{1}{y} \frac{dy}{dx} = \frac{d \log y}{dx}$$

and, by division,

$$\frac{-kx}{a+x} = -k + \frac{ka}{a+x}$$

Hence,

$$\frac{d \log y}{dx} = -k + \frac{ka}{a+x}$$

and, upon integrating,

$$\log_e y = -kx + ka \log_e (a+x) + \log_e c$$

Therefore,

$$y = ce^{-kx}(a+x)^{ka} = c' \left(1 + \frac{x}{a}\right)^{ka} e^{-kx}$$

where c is the constant of integration and $c' = ca^{ka}$.

The value of c' is determined from the condition that the area under the curve must equal 1. In doing this, it is to be noted that negative frequencies are inadmissible so that x cannot be less than $-a$. It is therefore the area under that part of the curve from $x = -a$ to ∞ that is to be equal to 1, that is, $\int_{-a}^{\infty} y dx = 1$.

To carry out this integration, multiply y by $\frac{e^{ka}e^{-ka}(ka)^{ka}}{(ka)^{ka}}$, which, of course, equals 1, which puts the integral in the form

$$\int_{-a}^{\infty} y dx = \int_{-a}^{\infty} c' \frac{e^{ka}}{(ka)^{ka}} [k(a+x)]^{ka} e^{-k(a+x)} dx$$

If z is set equal to $k(a+x)$, this becomes

$$\int_{-a}^{\infty} y dx = \frac{c'ae^{ka}}{(ka)^{ka+1}} \int_0^{\infty} z^{ka} e^{-z} dz$$

(Note that $dz = k dx$). But $\int_0^{\infty} z^{ka} e^{-z} dz$ is by definition $\Gamma(ka+1)$, called gamma of $ka+1$, and equals $(ka)!$ * Consequently, since

* Integration of $\int_0^{\infty} z^{m-1} e^{-z} dz$ by parts gives

$\int_{-a}^{\infty} y \, dx$ must equal 1,

$$c' = \frac{(ka)^{ka+1}}{ae^{ka}(ka)!}$$

Since $y = c'$ when $x = 0$ and since $x = 0$ is the mode of the distribution (for $dy/dx = 0$ when $x = 0$) c' is designated as y_0 and equals the height of the curve at the mode. Therefore, the formula for the curve is that given in the text (page 49), viz.,

$$y = y_0 \left(1 + \frac{x}{a}\right)^{ka} e^{-kx}, \text{ the origin being the mode of the distribu-}$$

tion. This is Pearson's type III curve. Taking logarithms of both sides yields

$$\log_{10} y = \log_{10} y_0 + ka \log_{10} \left(1 + \frac{x}{a}\right) - kx \log_{10} e,$$

which is the form given on page 49.

F. Proof That the Relative Slope of the Hypergeometrical Polygon Is Given at Any Abscissa Mid-point $X = N_1 + \frac{1}{2}$ by a Formula of the Type Relative Slope = $\frac{X + a}{b_0 + b_1 X + b_2 X^2}$. By Eq. (4) of Chap. IV (see page 54) the ordinate of the hypergeometrical distribution at any abscissa, point N_1 is

$$y_{N_1} = \frac{(p_1 S)!(p_2 S)!(S - N)!N!}{(p_1 S - N_1)!(p_2 S - N + N_1)!S!N_1!(N - N_1)!} \cdot \frac{1}{-e^{-z}z^{m-1} \Big|_0^{\infty} + (m-1) \int_0^{\infty} z^{m-2}e^{-z} \, dz}.$$

But the first term is zero both for $z = 0$ and $z = \infty$, and therefore

$$\int_0^{\infty} z^{m-1}e^{-z} \, dz = (m-1) \int_0^{\infty} z^{m-2}e^{-z} \, dz.$$

If m is a positive integer, repetition gives

$$\int_0^{\infty} z^{m-1}e^{-z} \, dz = (m-1)(m-2) \cdots \int_0^{\infty} e^{-z} \, dz$$

or since $\int_0^{\infty} e^{-z} \, dz = -e^{-z} \Big|_0^{\infty} = 1$

$$\int_0^{\infty} z^{m-1}e^{-z} \, dz = (m-1)!$$

When m is not a positive integer, $\int_0^{\infty} z^{m-1}e^{-z} \, dz$ is taken as the definition of

$(m-1)!$ The function $\int_0^{\infty} z^{m-1}e^{-z} \, dz$ is called the gamma function of m and is written $\Gamma(m)$. Thus $\Gamma(m) = (m-1)!$, $\Gamma(m+1) = m!$, etc. (see p. 254).

and its ordinate at the abscissa point $N_1 + 1$ is

$$y_{N_1+1} = \frac{(p_1 S)!(p_2 S)!(S - N)!N!}{(p_1 S - N_1 - 1)!(p_2 S - N + N_1 + 1)!S!(N_1 + 1)!(N - N_1 - 1)!}$$

The difference between these two ordinates is

$$\begin{aligned} \Delta y_{N_1} &= y_{N_1+1} - y_{N_1} \\ &= \frac{(p_1 S)!(p_2 S)!(S - N)!N!}{(p_1 S - N_1 - 1)!(p_2 S - N + N_1)!S!N_1!(N - N_1 - 1)!} \\ &\quad \left[\frac{1}{(p_2 S - N + N_1 + 1)(N_1 + 1)} - \frac{1}{(p_1 S - N_1)(N - N_1)} \right] \end{aligned}$$

Since the distance between the abscissa points is 1, this difference is the value of the absolute slope of the side of the polygon joining the two ordinates.

The ordinate at the abscissa mid-point $N_1 + \frac{1}{2}$ is equal to $\frac{1}{2}(y_{N_1+1} + y_{N_1})$, which equals

$$\begin{aligned} &\frac{(p_1 S)!(p_2 S)!(S - N)!N!}{2(p_1 S - N_1 + 1)!(p_2 S - N + N_1)!S!N_1!(N - N_1 - 1)!} \\ &\quad \left[\frac{1}{(p_2 S - N + N_1 + 1)(N_1 + 1)} + \frac{1}{(p_1 S - N_1)(N - N_1)} \right] \end{aligned}$$

and the relative slope is the ratio of the absolute slope to the value of this mid-ordinate. That is,

$$\text{Relative slope} = \frac{2[(p_1 S - N_1)(N - N_1) - (p_2 S - N + N_1 + 1)(N_1 + 1)]}{(p_1 S - N_1)(N - N_1) + (p_2 S - N + N_1 + 1)(N_1 + 1)}$$

which reduces to

$$\text{Relative slope} = \frac{2[N + N p_1 S - p_2 S - 1 - N_1(S + 2)]}{N p_1 S + p_2 S + 1 - N + N_1[(p_2 - p_1)S - 2N + 2] + 2N_1^2}$$

If now X is set equal to $N_1 + \frac{1}{2}$, the foregoing expression for the relative slope may be put in the form

$$\begin{aligned} \text{Relative slope} &= \frac{X + \left[\frac{1}{2} - \frac{(N + 1)(1 + p_1 S)}{S + 2} \right]}{\frac{2N p_1 S + S + 1}{4(S + 2)} + \frac{[(p_2 - p_1)S - 2N]X}{2(S + 2)}} \\ &\quad + \frac{X^2}{S + 2} \end{aligned}$$

which is equivalent to

$$\text{Relative slope} = \frac{X + a}{b_0 + b_1 X + b_2 X^2}$$

where

$$\begin{aligned} a &= \frac{1}{2} - \frac{(N+1)(1+p_1S)}{S+2} \\ b_0 &= \frac{2Np_1S + S + 1}{4(S+2)} \\ b_1 &= \frac{(p_2 - p_1)S - 2N}{2(S+2)} \\ b_2 &= \frac{1}{S+2} \end{aligned}$$

G. The Criterion for the Hypergeometrical Distribution. The character of the hypergeometrical distribution will depend on the value of the discriminant $b_1^2 - 4b_0b_2$. From the results of the previous section, this is seen to be

$$\frac{[(p_2 - p_1)S - 2N]^2 - 4(2Np_1S + S + 1)}{4(S+2)^2}$$

The value of the discriminant is thus related to the values of p_1 , p_2 , N , and S . It is positive if N/S lies outside the limits $\frac{1}{2} \pm \sqrt{\left(p_1 + \frac{1}{S}\right)\left(p_2 + \frac{1}{S}\right)}$, and it is negative if N/S lies within these limits. In the first case, the hypergeometrical distribution is approximated by a type VI curve (see pages 60, 135); in the second case, by a type IV curve.

In the example employed in the text, $p_1 = \frac{1}{4}$, $p_2 = \frac{3}{4}$, $S = 52$, and $N = 10$. Hence, $N/S = \frac{1}{5.2}$. On the other hand,

$$\begin{aligned} \frac{1}{2} \pm \sqrt{\left(p_1 + \frac{1}{S}\right)\left(p_2 + \frac{1}{S}\right)} &= \frac{1}{2} \pm \sqrt{\left(\frac{1}{4} + \frac{1}{52}\right)\left(\frac{3}{4} + \frac{1}{52}\right)} \\ &= \frac{1}{2} \pm \sqrt{\left(\frac{14}{52}\right)\left(\frac{40}{52}\right)} = \frac{1}{2} \pm \frac{1}{52} \sqrt{560} = \frac{49.7}{52} \text{ or } \frac{2.3}{52} \end{aligned}$$

Since $N/S (= \frac{1}{5.2})$ lies within these limits, the hypergeometrical distribution given by these values of p_1 , p_2 , S , and N is approximated by a type IV curve.

It will be noted that N , p_1 , and S are all positive quantities so that the second term of the discriminant (without the minus sign) must be positive. That is, the roots of the equation $b_0 + b_1x + b_2x^2 = 0$ cannot be real and of opposite sign. Consequently, Pearsonian curves of type I cannot be derived from a hypergeometrical distribution (see footnote to page 58).

CHAPTER V

THE GRAM-CHARLIER SYSTEM OF FREQUENCY CURVES

The introduction of a possible variation in the contribution of a causal factor at the end rather than at the beginning of the argument may be considered a weak point of the Pearsonian analysis. It is therefore of interest to consider another approach to the theory of frequency curves that assumes variable contributions from the start. Such an approach is that which gives rise to the Gram-Charlier system of frequency curves.¹

Derivation of the Gram-Charlier Formula. The assumptions on which the Gram-Charlier system of frequency curves is based are similar in many respects to those from which the skew binomial and Pearson's type III curve were derived. The first fundamental assumption is that the fluctuations in a given variable X are the algebraic sum of the contributions of a number of causal factors. Thus the deviations of X from some central value are assumed to be equal to $\epsilon_1 + \epsilon_2 + \dots + \epsilon_N$ where the ϵ 's may take on either positive or negative values. The significance of this assumption is that the contributions of the various causes are additive rather than multiplicative or related in some more complex way. The second principal assumption is that the contribution of each cause is independent of the contributions of other causes.

The two assumptions above are essentially the same as those made by Pearson in the derivation of his type III curve; but a third fundamental assumption differs from Pearson's. In the case of the binomial distribution it was assumed that each con-

¹ So called because J. P. Gram and C. V. L. Charlier were principally responsible for its development. See, for example, J. P. Gram, *Om Rækkedeviklinger* (Copenhagen, 1879) (Doctor's Dissertation); and "Über die Entwicklung reeller Functionen in Reihen mittelst die Methode der kleinsten Quadrate," *Journal für Mathematik*, Vol. 94 (1883), pp. 41-73. Also see C. V. L. Charlier, "Über das Fehlergesetz," *Arkiv för Matematik, Astronomi och Fysik*, Vol. 2, No. 8 (1905), pp. 1-9; and "Über die Darstellung willkürlicher Functionen," *Arkiv för Matematik, Astronomi och Fysik*, Vol. 2, No. 20 (1905), pp. 1-35.

tributory cause could add or subtract only a fixed amount to the variable (only an H or a T was possible). Under the Gram-Charlier system, it is assumed that the contributions of each causal factor can take on various values, the probability of any given value being given by a distribution the form of which is fixed, but not necessarily known. These distributions of probability are allowed to vary from cause to cause; they may be discrete or continuous and are not subject to any restrictions other than that the probability of very large positive or negative contributions shall be practically zero, *i.e.*, that the distributions taper off to zero in both directions.

On the basis of these assumptions the Gram-Charlier analysis¹ shows that, if the number of contributory causes is relatively large, then the distribution of the resultant variable X bears certain approximate relationships to the distributions of the individual contributory elements ϵ . In setting up these relationships it makes use of certain quantities, called "cumulants" or "semi-invariants," that are directly related to the moments of a distribution. Thus, if μ_1, μ_2, μ_3 , and μ_4 are the first four moments of a distribution about its mean, then $k_1 = \mu_1, k_2 = \mu_2, k_3 = \mu_3$, and $k_4 = \mu_4 - 3\mu_2^2$, where the k 's are the first four cumulants of the distribution. Both k_1 and μ_1 are 0, of course, when the moments are measured about the mean. It is obvious that $k_2 = \sigma^2$, that k_3 is related to the skewness of the distribution, and that k_4 is related to its kurtosis.* When k_3 is 0, the distribution is symmetrical; when k_4 is 0, its kurtosis is 3.

By adopting certain mathematical approximations, the Gram-Charlier analysis shows that the cumulants of the distribution of X are the sum of the cumulants of the distributions of the individual contributory elements. That is, if k_{12} is the second cumulant of the distribution of the first contributory element ϵ_1 , k_{22} the second cumulant of the second contributory element ϵ_2 , k_{23} the third cumulant of the second contributory element, etc., and if K_2, K_3 , and K_4 are the cumulants of the distribution of X , then

$$K_2 = k_{12} + k_{22} + k_{32} + \dots + k_{N2}$$

$$K_3 = k_{13} + k_{23} + k_{33} + \dots + k_{N3}$$

$$K_4 = k_{14} + k_{24} + k_{34} + \dots + k_{N4}$$

¹ For a fuller discussion of the Gram-Charlier analysis see the Appendix, pp. 92-99.

These relationships form the basis of the Gram-Charlier analysis. Suppose, it is argued, that the variance K_2 of the resultant variable X is finite, as it is in most practical cases, and suppose that the variances, k_2 's, of the individual contributory elements are all of about the same order of magnitude. Then, since there are N contributory elements, the variance k_2 of each of them is of about the size of K_2/N , or, as the mathematicians say, of the order of $1/N$, and its standard deviation $\sqrt{k_2}$ is of the order of $1/\sqrt{N}$. This means that the average variation in each contributory element is of order $1/\sqrt{N}$. Hence the third and fourth moments (and therefore the third and fourth cumulants), which involve the third and fourth powers of this variation, will be of order $1/N^3$ and $1/N^2$, respectively. Consequently, the third and fourth cumulants of X , K_3 and K_4 , which are equal, respectively, to the sum of the third and fourth cumulants of the N contributory elements, will be of order $1/\sqrt{N}$ and $1/N$.

It therefore follows that, if N is large, the third and fourth cumulants of X will be approximately 0. This means that the distribution of X will be practically symmetrical and its kurtosis will be close to 3. The Gram-Charlier analysis shows further that under these circumstances the form of the distribution of X will be that of the normal curve.

In summary, then, the Gram-Charlier analysis shows that if variation in a quantity X is the sum of a large number of elementary independent variations ϵ , and if these elementary variations are all of about the same order of magnitude, then the distribution of X will be approximately normal in form.

If N is not large enough to make terms of order $1/\sqrt{N}$ or $1/N$ negligible but is still large enough to make terms of higher order (such as terms of order $1/N^2$) practically zero, then the Gram-Charlier analysis shows that the form of the distribution of X will be given approximately by

$$y = \frac{1}{\sigma \sqrt{2\pi}} \left[1 + \frac{A}{\sigma^3} \left(\frac{3x}{\sigma} - \frac{x^3}{\sigma^3} \right) + \frac{B}{\sigma^4} \left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right) \right] e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

where

$$x = X - \bar{X}, \quad \sigma^2 = K_2, \quad A = \frac{-K_3}{3!} = \frac{\mu_3}{3!}, \quad \text{and}$$

$$B = \frac{K_4}{4!} = \frac{\mu_4 - 3\mu_2^2}{4!}.$$

This is the more general Gram-Charlier formula for a frequency curve. The normal curve is the special case for which A and B both equal 0; all other curves are nonnormal.

Components of a Gram-Charlier Curve. The effect of the additional A and B terms on the shape of the distribution of X is illustrated in Fig. 17. Here the ordinary normal curve is represented by curve I. If A is negative, the effect of the A

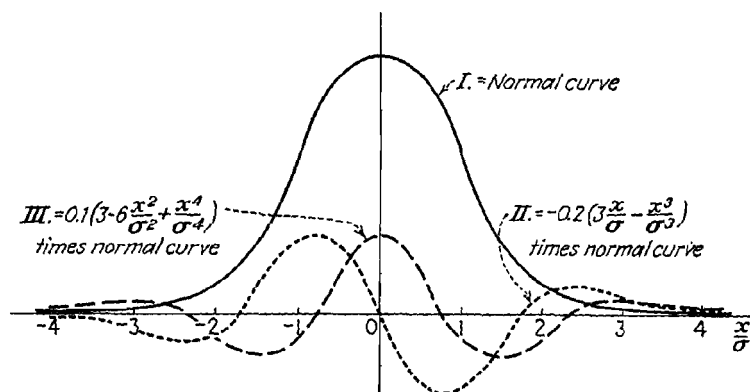


FIG. 17.—Components of a Gram-Charlier frequency curve.

term is to subtract from the normal ordinate for positive values of $\frac{X - \bar{X}}{\sigma}$ an increasing percentage of that ordinate up to $\frac{X - \bar{X}}{\sigma} = 1$

and then a decreasing percentage up to $\frac{X - \bar{X}}{\sigma} = \sqrt{3}$ and there-

after to add a small percentage. The opposite is true for negative values of $\frac{X - \bar{X}}{\sigma}$. The fluctuations in the amounts added

and subtracted are illustrated by curve II of Fig. 17, for which $A/\sigma^3 = -.2$. Since, when A is negative, the effect of the A term is to subtract from the right of the mean and to add to the left of the mean (at least in its immediate neighborhood), the result is a transformation of the normal curve into a positively skewed curve (*cf.* Fig. 18). If A is positive, the effect of the A term is to add to the right of the mean and to subtract from the left (*cf.* Fig. 19), at least within the immediate neighborhood, and thus to transform the normal curve into a negatively skewed curve (*cf.* Fig. 20).

This dependence of the skewness of the distribution of X on the sign of A was to be expected. For, as indicated above, $A = -K_3/3!$ and K_3 itself is a sum of the cumulants k_{13} , k_{23} ,

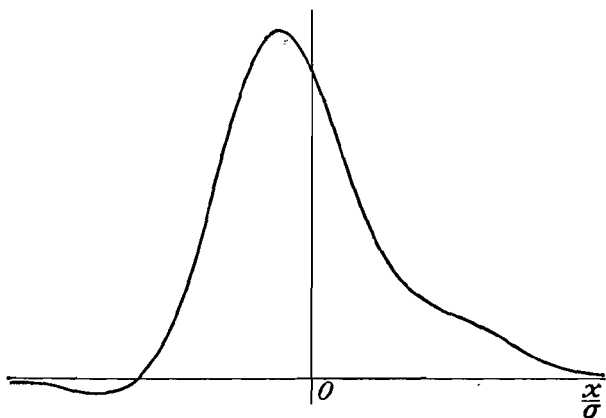


FIG. 18.—Combination of curves I and II shown in Fig. 17.

etc., which are the third moments of the distributions of the individual contributions. Consequently, if the distributions of the individual contributions are all positively skewed, their

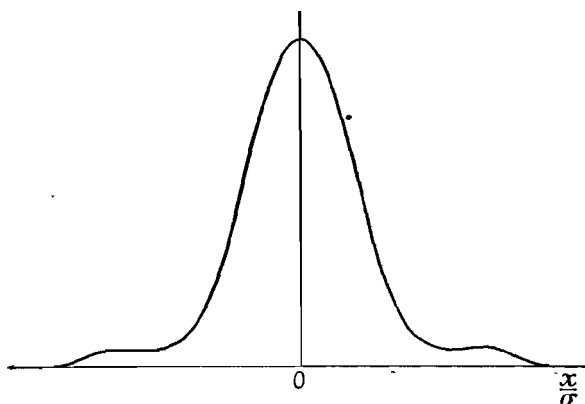


FIG. 19.—Combination of curves I and III shown in Fig. 17.

third moments, and hence k_{13} , k_{23} , etc., will all be positive; K_3 , which equals $k_{13} + k_{23} + k_{33} + \dots + k_{N3}$, will be positive; A will be negative; and the distribution of X will be positively skewed. That is, if the distributions of the individual con-

tributions, ϵ , are all positively skewed, then $\bar{X}$ itself will be positively skewed. Just the opposite is true if the individual contributions are all negatively skewed. If some of the individual contributions are positively skewed and others are

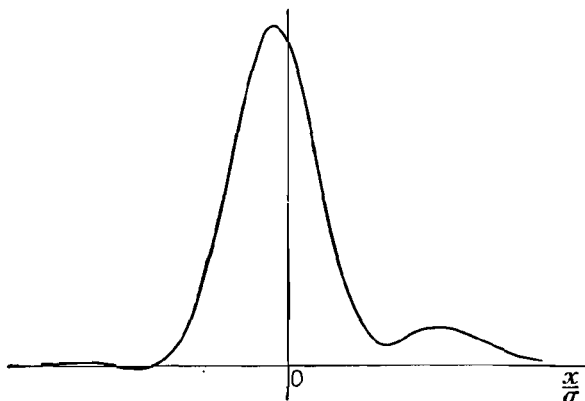


FIG. 20.—Combination of curves I, II, and III shown in Fig. 17.

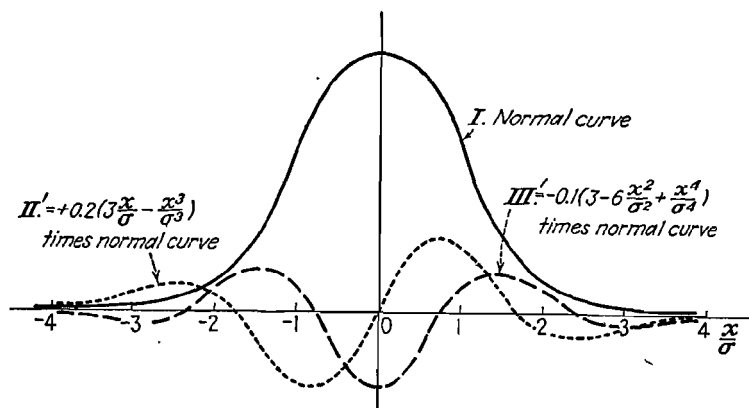


FIG. 21.—Effect upon normal curve of additions and subtractions type II' and III'.

negatively skewed, then the value of K_3 , and hence the skewness of X , will depend on whether the positive or negative influences are predominant. Since the contributions are presumed to be of approximately the same order of magnitude, the result will depend primarily on the relative number of positively skewed and negatively skewed contributions.

The effect of the B term on the distribution of X is shown graphically in Figs. 17, 19, 21, and 23. If B is positive, the effect of the B term is to add a maximum percentage to the ordinate of the normal curve at $\frac{X - \bar{X}}{\sigma} = 0$. This percentage decreases in size until $\frac{X - \bar{X}}{\sigma}$ equals $-.74$ and $+.74$, respectively, from which points an increasing percentage is subtracted from the

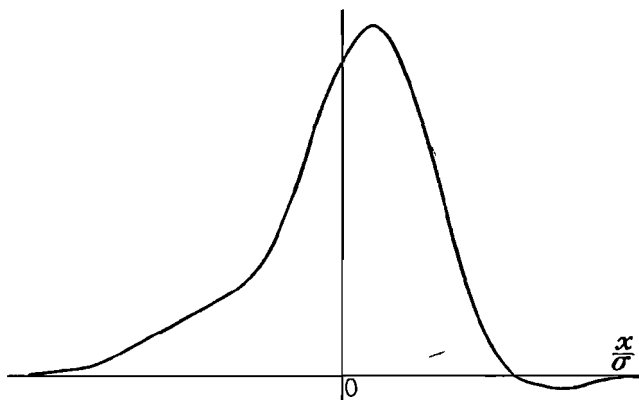


FIG. 22.—Combination of curves I and II' shown in Fig. 21.

normal ordinates. The maximum percentage subtraction is attained at $\frac{X - \bar{X}}{\sigma} = -\sqrt{3}$ and $+\sqrt{3}$, after which a decreasing percentage is subtracted until $\frac{X - \bar{X}}{\sigma} = -2.33$ and $+2.33$. From those points on, a small percentage is added to the normal ordinates.

These fluctuations in the percentages added and subtracted are illustrated by curve III of Fig. 17, for which B/σ^4 is taken equal to $+10$. When B is positive, the effect of the B term is to add to the normal curve in the immediate vicinity of the mean, to subtract from it at intermediate distances from the mean, and then to add to it again on the tails of the distribution; the result is to give the final curve a greater than normal peakedness. That is, if B is positive, the distribution of X will have a coefficient of kurtosis greater than 3 (*cf.* Fig. 19). Just the opposite is true if B is negative. In this case subtractions are made

from the normal curve within the immediate vicinity of the mean, additions are made at intermediate distances, and subtractions are again made out on the tails (*cf.* Fig. 21). The result is to produce less than normal peakedness (*cf.* Fig. 23). If B is negative, therefore, the distribution of X will have a coefficient of kurtosis less than 3.

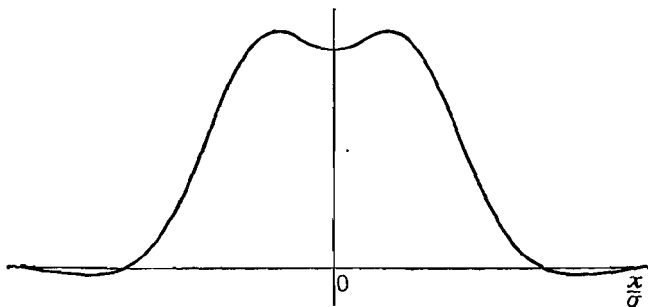


FIG. 23.—Combination of curves I and III' shown in Fig. 21.

Again this result was to be expected. For the value of B depends on the value of K_4 , and this in turn depends on the values of k_{14} , k_{24} , etc., which represent the excess above 3 or the deficit below 3 of the coefficients of kurtosis of the distributions of the individual contributions. Thus if the distributions of the

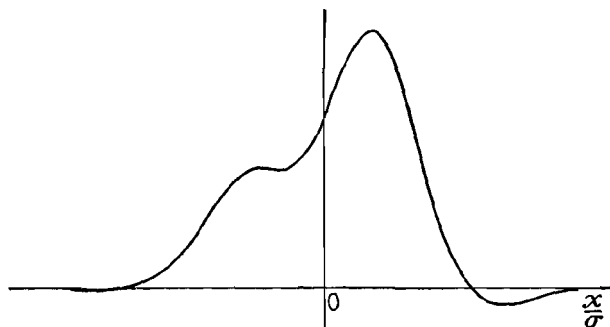


FIG. 24.—Combination of curves I, II', and III' shown in Fig. 21.

individual contributions, ϵ , are all more peaked than normal, *i.e.*, if the k_4 's are all positive, then K_4 , which equals

$$k_{14} + k_{24} + k_{34} + \dots + k_{N4},$$

will also be positive, B will be positive, and the distribution of X will be more peaked than normal. The opposite is true if the

distributions of the individual contributions are all less peaked than normal. If some of them are more peaked than normal and others are less peaked, then the peakedness of the distribution of X will depend upon the predominance of excess or deficit influences; and since the individual contributions are all assumed to be of approximately the same order of magnitude, this predominance will be primarily determined by the relative number of excess and deficit influences.

Usually the A and B terms are the only additional terms deemed worthy of consideration. For unless the number of causal factors, N , is relatively small, higher-order terms will be of negligible importance in determining the distribution of X , continuing to assume, of course, that the contributions of the various causal factors are approximately of the same order of magnitude.¹ In the practical work of fitting a Gram-Charlier curve, it also becomes difficult to determine with any degree of accuracy the values of these higher-order terms.² For these reasons, Eq. (1) is usually considered a sufficiently general equation. In special cases of very skewed distributions, somewhat better results may be obtained by a type of mathematical approximation that gives rise to a different equation³ from Eq. (1). Gram-Charlier distributions of this latter kind are called "type B distributions" to distinguish them from those given by Eq. (1), which are called 'type A distributions.'

GENERAL SIGNIFICANCE OF THE GRAM-CHARLIER ANALYSIS

The principal significance of the Gram-Charlier analysis for the theory of frequency curves is the explanation it affords of nonnormal frequency curves when the number of contributory causes is limited. The Pearsonian analysis, it will be recalled,⁴

¹ Failure to include higher-order terms, however, may sometimes produce negative frequencies in certain parts of a Gram-Charlier curve, which is of course a practical impossibility. Such, for instance, is seen to be the case in Figs. 15 and 20.

² The difficulty is that the sampling fluctuations in these higher-order terms are especially great.

³ For a discussion of this special variation in the analysis, see C. V. L. Charlier, "Über die Darstellung willkürlicher Functionen," *Arkiv för Matematik, Astronomi och Fysik*, Vol. 2, No. 20 (1905), pp. 1-35. An elementary discussion in English is to be found in H. L. Reitz, *Mathematical Statistics*, Chap. VII.

⁴ See p. 63.

was strictly valid for nonnormal distributions of discrete variables, but its explanation of continuous nonnormal frequency curves depended upon a tardily introduced assumption of "lumpiness" in the contributions of the causal factors—a procedure that was not entirely satisfactory since the analysis was originally based upon the assumption of definitely fixed contributions.

The Gram-Charlier analysis assumes at the very beginning that the contributions of a causal factor themselves vary continuously, the probabilities of various possible values being given by a distribution of a definite but unknown form. It then goes on to show that if a given variable X is an algebraic sum of the contributions ϵ of a number of such causal factors, if the contributions are approximately of the same order of magnitude (*i.e.*, if the standard deviations of the contributions are about equal), and if they are independent of each other, then the form of the distribution of X will depend partly on the forms of the distributions of the individual contributions and partly on the number of such contributory causes. Thus, if the number of causal factors is very large, the distribution of X will be approximately normal, whatever the forms of the distributions of the individual contributions. If the number of causal factors is only moderately large, however, but their contributions are still of the same order of magnitude, the skewness and kurtosis of the distributions of the individual contributions will tend to produce a skewness and kurtosis in the distribution of X . In special cases the skewness of the individual contributions may be compensatory. Thus it is possible for the distribution of X to be symmetrical, not only when the distributions of the individual contributions are themselves all symmetrical, but also when the skewness of one or more contributions is offset by a contrary skewness of one or more other contributions. Similar remarks may be made with respect to the kurtosis of X .

These conclusions are much the same as those drawn from the Pearsonian analysis pertaining to the asymmetrical binomial distribution. In fact, this part of the Pearsonian analysis may be considered a very special case of the Gram-Charlier analysis. Suppose, for example, that, in the latter, distributions of the individual contributions are all assumed to be of a very special sort. Suppose that in each case only one specified positive

contribution and one specified negative contribution are possible and that these have the same absolute value. Let the probability with which the contribution takes on its positive value differ from that with which it takes on its negative value. Furthermore, let these distributions of the individual contributions be all alike, both as to the amounts of the positive and negative values and as to the probabilities of each. Then, under these very special assumptions, the Gram-Charlier analysis will yield the asymmetrical binomial distribution. The latter is thus a special case of a Gram-Charlier distribution.

Finally, it is to be noted that the Gram-Charlier analysis does not, as does the Pearsonian analysis, relax the assumption of independence of the contributory causes. The latter, it will be recalled, laid aside this assumption and permitted the contribution of a causal factor to be dependent on the contributions of other factors. The skewness and kurtosis yielded by the general Pearsonian formula were in part to be attributed to this assumption of dependence. The Gram-Charlier analysis assumes throughout that the contribution of any causal factor is independent of the contributions of other causal factors. It is restricted in this respect in the explanation that it affords of actual frequency distributions.

APPENDIX

DERIVATION OF THE GRAM-CHARLIER FORMULA FOR A TYPE A FREQUENCY DISTRIBUTION

The following is a more detailed discussion of the Gram-Charlier system of frequency curves than was given in the main body of the text. Its principal purpose is to outline the argument by which the formula for a type A frequency distribution [Eq. (1)] is derived. For the sake of those not well acquainted with higher mathematics, the essential steps are presented in a more or less nonmathematical manner, while the mathematical basis for each is sketched in a series of footnotes. For a fuller mathematical analysis, the reader is referred to the original works of Gram and Charlier (see footnote to page 82) or to the English accounts of them given by Arne Fisher's *Mathematical Theory of Probabilities* or H. L. Reitz's *Mathematical Statistics*.¹ The

¹ Also see THIELE, T. N., *Theory of Observations*, first published in 1903, (C. and E. Layton, London) and recently reprinted in the *Annals of Mathematical Statistics*, Vol. 3 (1933), pp. 165-308.

present discussion follows closely that outlined in E. T. Whittaker and G. Robinson, *The Calculus of Observations*,¹ pages 168 to 174.

The cornerstone of the Gram-Charlier analysis is the Fourier integral theorem. This, so far as the argument in question is concerned, states that if a certain function of a given variable, called its "moment generating function," is known, then the frequency distribution of the variable itself may be determined and conversely if the distribution is known its moment generating function may be found.² This moment generating function is

of the form $G = \sum_{-\infty}^{\infty} e^{i\theta x} f(x) dx$, where $x = X - \bar{X}$, $f(x) dx$ is the

distribution of x , e is the constant 2.718+, $i = \sqrt{-1}$, and θ is an arbitrary variable.³ G is called the moment generating function because, if certain operations are performed on it with reference to $i\theta$ and then θ is given the value of 0, the results are the various moments of the distribution of x .*

The significance of the moment generating function for the present analysis is that if a variable X is the sum of a number of other variables, the values of any one of which are independent of the values assumed by the others, then the moment generating function of the composite variable X is simply the product of the moment generating functions of the independent variables.⁴ Thus, with reference to the given problem, if variations in X are the sum of the contributions of certain causal factors, all acting independently of each other, the moment generating function of the frequency distribution of X , the form of which

¹ Blackie & Son, Ltd., Glasgow, 1925, 2d ed.

² More exactly, if the moment-generating function is defined as

$$G(\theta) = \int_{-\infty}^{\infty} f(x)e^{i\theta x} dx,$$

where $i = \sqrt{-1}$ and $f(x) dx$ represents the frequency distribution of x , then $f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\theta)e^{-i\theta x} d\theta$, and vice versa.

³ When x is distributed in the form of a smooth continuous curve, G is defined more exactly by the formula of the previous footnote.

* Thus, if G is differentiated with respect to $i\theta$ and θ is set equal to 0, the result is the first moment of $f(x)$. If G is differentiated twice and then θ is set equal to 0, the result is the second moment of $f(x)$, etc. This is frequently the simplest method of obtaining equations for the moments of a known distribution.

⁴ The reason is that $e^{a+b} = e^a e^b$. Cf. WHITTAKER and ROBINSON, *op. cit.*, p. 170.

is being sought, is the product of the moment generating functions of the frequency (*i.e.*, probability) distributions of the individual contributory factors. Symbolically,

$$G_X = G_1 G_2 \cdots G_N$$

This relationship between the generating function of the sum and that of the individual contributions is of fundamental importance, but it does not of itself advance the argument. For there is nothing in the given assumptions that provides any knowledge about the generating functions of the individual contributory factors. Hence the relationship of the latter to the generating function of X can of itself give no information about the nature of this resultant generating function.

The next step in the argument, however, yields more tangible results. This is to note that the logarithm of a moment generating function can be represented by an infinite series, the terms of which consist of rising powers of $i\theta$ and certain quantities that depend on the distribution to which the generating function relates. Thus if $f(\epsilon_1) d\epsilon_1$ represents the probability distribution of cause 1 and $G_1 = \sum_{-\infty}^{\infty} f(\epsilon_1) e^{i\theta \epsilon_1} d\epsilon_1$ is its moment generating function, then the logarithm of G_1 can be put in the form

$$\log G_1 = i\theta k_{11} - \frac{\theta^2}{2!} k_{12} + \frac{(i\theta)^3}{3!} k_{13} + \frac{\theta^4}{4!} k_{14} - \cdots \quad (1)$$

where k_{11} , k_{12} , etc., are quantities that are called "semivariants" or "cumulants" and, if ϵ_1 is measured from its mean, are related to the moments of $f(\epsilon_1) d\epsilon_1$ as follows:¹

¹ The moment-generating function $G_1 = \sum_{-\infty}^{\infty} e^{i\theta \epsilon_1} f(\epsilon_1) d\epsilon_1$ is a function of $i\theta$ and may be written $G_1(i\theta)$. Let $\log G_1(i\theta)$ be called $K_1(i\theta)$; then $K_1(i\theta)$ may be expanded in a power series in $i\theta$ (Taylor's expansion), in the neighborhood of $i\theta = 0$, as follows:

$$K_1(i\theta) = K_1(0) + K_1'(0)i\theta + K_1''(0)\frac{(i\theta)^2}{2!} + \cdots$$

where K_1' , K_1'' , etc., represent derivatives with respect to $i\theta$. The problem is to find $K_1(0)$, $K_1'(0)$, $K_1''(0)$, $\cdots$.

Since $G_1(i\theta) = 1$ when $\theta = 0$ [note that the sum from $-\infty$ to $+\infty$ of the probabilities of $f(\epsilon_1) d\epsilon_1 = 1$], it follows that $\log G_1(0) = 0$, *i.e.*, $K_1(0) = 0$. Again

$$\frac{d \log G_1(i\theta)}{d(i\theta)} = \frac{1}{G_1(i\theta)} \frac{dG_1(i\theta)}{d(i\theta)}$$

or

$$k_{11} = \mu_1 = 0, \quad k_{12} = \mu_2, \quad k_{13} = \mu_3, \quad k_{14} = \mu_4 - 3\mu_2^2, \text{ etc.}$$

The same can be done for the logarithms of the generating functions of the other contributions, and since the logarithm of a product is the sum of the logarithms of the various factors, the logarithm of the generating function of X can be represented as the sum of the N infinite series representing the logarithms of the

$$K'(i\theta) = \frac{1}{G_1(i\theta)} G_1'(i\theta)$$

or

$$G_1(i\theta)K'(i\theta) = G_1'(i\theta)$$

Furthermore, by expanding $e^{i\theta\epsilon_1}$ in a power series, it follows that

$$\begin{aligned} G_1(i\theta) &= \sum_{-\infty}^{\infty} f(\epsilon_1) d\epsilon_1 + i\theta \sum_{-\infty}^{\infty} \epsilon_1 f(\epsilon_1) d\epsilon_1 \\ &\quad + \frac{(i\theta)^2}{2!} \sum_{-\infty}^{\infty} \epsilon_1^2 f(\epsilon_1) d\epsilon_1 + \frac{(i\theta)^3}{3!} \sum_{-\infty}^{\infty} \epsilon_1^3 f(\epsilon_1) d\epsilon_1 + \dots \\ &= 1 + i\theta\mu_1 + \frac{(i\theta)^2}{2!} \mu_2 + \frac{(i\theta)^3}{3!} \mu_3 + \dots \end{aligned}$$

$$\text{because by definition } \mu_1 = \sum_{-\infty}^{\infty} \epsilon_1 f(\epsilon_1) d\epsilon_1, \mu_2 = \sum_{-\infty}^{\infty} \epsilon_1^2 f(\epsilon_1) d\epsilon_1, \text{ etc.}$$

Thus

$$G_1'(i\theta) = \mu_1 + i\theta\mu_2 + \frac{(i\theta)^2}{2!} \mu_3 + \dots$$

and likewise from Taylor's expansion for $K_1(i\theta)$, it follows that

$$K_1'(i\theta) = K_1'(0) + K_1''(0)i\theta + K_1'''(0)\frac{(i\theta)^2}{2!} + \dots$$

Hence the above identity may be written

$$\begin{aligned} \left[1 + i\theta\mu_1 + \frac{(i\theta)^2}{2!} \mu_2 + \dots \right] &\left[K_1'(0) + K_1''(0)i\theta + K_1'''(0)\frac{(i\theta)^2}{2!} + \dots \right] \\ &= \mu_1 + i\theta\mu_2 + \frac{(i\theta)^2}{2!} \mu_3 + \dots \end{aligned}$$

If the left side of this identity is multiplied out and if the coefficients of various powers of $i\theta$ on the left are equated to the coefficients of the same powers of $i\theta$ on the right, it follows that when ϵ_1 is measured from its mean

$$\begin{aligned} K_1'(0) &= \mu_1 = 0 \\ K_1''(0) &= \mu_2 \\ K_1'''(0) &= \mu_3 \end{aligned}$$

and

$$\frac{\mu_2}{2!} K_1''(0) + \frac{K_1^{IV}(0)}{3!} = \frac{\mu_4}{3!}$$

or

$$K_1^{IV}(0) = \mu_4 - 3\mu_2^2$$

If k_{11} is set equal to $K_1'(0)$, k_{12} to $K_1''(0)$, k_{13} to $K_1'''(0)$, k_{14} to $K_1^{IV}(0)$, etc., the equation for $\log G_1(i\theta) = K_1(i\theta)$ is seen to be equation (1) of the text.

generating functions of the N individual contributions. Thus, if the first subscript of a cumulant $\mathbf{k}$ represents the contribution to which it pertains and the second the order of the cumulant, it follows that:¹

$$\begin{aligned} \log G_X = & -(\mathbf{k}_{13} + \mathbf{k}_{22} + \mathbf{k}_{32} + \cdots + \mathbf{k}_{N2}) \frac{\theta^2}{2!} \\ & + (\mathbf{k}_{13} + \mathbf{k}_{23} + \mathbf{k}_{33} + \cdots + \mathbf{k}_{N3}) \frac{(i\theta)^3}{3!} \\ & + (\mathbf{k}_{14} + \mathbf{k}_{24} + \mathbf{k}_{34} + \cdots + \mathbf{k}_{N4}) \frac{\theta^4}{4!} + \cdots \quad (2) \end{aligned}$$

Comparison of the coefficients of Eq. (2) with those of (1) indicates that the cumulants of the generating function of the distribution of X are the sums of the corresponding cumulants of the generating functions of the individual contributions. That is, if the cumulants of X are indicated by $\mathbf{K}_2, \mathbf{K}_3$, etc., it follows that

$$\left. \begin{aligned} \mathbf{K}_2 &= \mathbf{k}_{12} + \mathbf{k}_{22} + \cdots + \mathbf{k}_{N2} \\ \mathbf{K}_3 &= \mathbf{k}_{13} + \mathbf{k}_{23} + \cdots + \mathbf{k}_{N3} \\ \mathbf{K}_4 &= \mathbf{k}_{14} + \mathbf{k}_{24} + \cdots + \mathbf{k}_{N4}, \text{ etc.} \end{aligned} \right\} \quad (3)$$

It is this last expression that permits certain inferences regarding the moment generating function of X and hence the distribution of X . Thus suppose that two additional assumptions are made regarding the distributions of the contributions of the individual causal factors, *viz.*: (1) that the number of these causal factors, N , is very large and (2) that the variances $\sigma^2 = \mu_2$ of the individual contributions are all approximately of the same order of magnitude. The $\mathbf{k}_{12}$'s, it will be recalled, are equal to the variances of the individual contributions, and expression (3) shows that their sum gives the variance $\mathbf{K}_2$ of X . The variances of physical, biological, and economic variables are generally of finite size so that for practical purposes it may be assumed that the variance of X is finite. Since it is equal to the sum of N variances, all of about the same order of magnitude, it can be concluded that each of the contributory variances is of the order of $1/N$ and that each contribution is of order $1/\sqrt{N}$.

From this it follows that the higher cumulants and moments are of order greater than $1/N$, for example, that $\mathbf{k}_{13}, \mathbf{k}_{23}$, etc.,

¹ For, it is to be remembered, $\mathbf{k}_{11} = 0$ on the assumption that ϵ_1 is measured from its mean, and the same is true for $\mathbf{k}_{21}, \mathbf{k}_{31}$, etc., on the assumption that all the ϵ 's are measured from their means.

are of order $1/\sqrt{N^3}$ and that k_{14}, k_{24} , etc., are of order $1/N^2$. Since K_3 is equal to the sum of the N cumulants k_{13}, k_{23} , etc., and K_4 is equal to the sum of the N cumulants k_{14}, k_{24} , etc., the magnitude of these higher cumulants of X is therefore of the order of $1/\sqrt{N}$ and $1/N$, respectively. Consequently, if the variance of X is finite and if N is very large, as is assumed, the size of K_3, K_4 and that of the higher cumulants of X will be very small. A first approximation to the generating function of the distribution of X is thus given by $\log G_x = -K_2 \frac{\theta^2}{2!}$ or $G = \exp \left[\frac{-K_2 \theta^2}{2!} \right]$; and from this, by means of the Fourier integral theorem mentioned above, it may be shown¹ that a first approximation to the distribution of X itself is $f(x) = \frac{1}{\sqrt{2\pi K_2}} \exp \left[-\frac{x^2}{2K_2} \right]$ which is the formula for the normal curve. The Gram-Charlier analysis thus shows that if the number of the contributory causes is very large and if their contributions are of about the same order of magnitude, *i.e.*, if the variances of the contributions of the individual causal factors are roughly the same, the distribution of the resultant variable X will approximate the normal form.

¹ Since, according to the Fourier integral theorem, the frequency distribution of X is $f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(\theta) e^{-i\theta x} d\theta$ when its moment generating function is $G(\theta) = \int_{-\infty}^{\infty} f(x) e^{i\theta x} dx$, it follows that, when $G(\theta) = \exp \left[\frac{-K_2 \theta^2}{2} \right]$, then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left[\frac{-K_2 \theta^2}{2} - i\theta x \right] d\theta$$

But if the square of the exponential is completed, it makes

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left[-\left(\frac{K_2 \theta^2}{2} + i\theta x + \frac{(ix)^2}{2K_2} \right) \right] \exp \left[\frac{(ix)^2}{2K_2} \right] d\theta \\ &= \frac{1}{2\pi} \exp \left[\frac{-x^2}{2K_2} \right] \int_{-\infty}^{\infty} \sqrt{\frac{2}{K_2}} \exp \left[-\left(\sqrt{\frac{K_2}{2}} \theta + \frac{ix}{\sqrt{2K_2}} \right)^2 \right] d \left(\sqrt{\frac{K_2}{2}} \theta + \frac{ix}{\sqrt{2K_2}} \right) \end{aligned}$$

But the integral is of the form $\int_{-\infty}^{\infty} \sqrt{\frac{2}{K_2}} e^{-z^2} dz$, which equals $\sqrt{\frac{2\pi}{K_2}}$.

Hence,

$$f(x) = \frac{1}{\sqrt{2\pi K_2}} \exp \left[-\frac{x^2}{2K_2} \right].$$

If the variances of the individual contributions are about the same (that is, if k_{12} , k_{22} , etc., are approximately equal) but the number of causal factors, N , is not large enough to warrant the dropping of terms of the order of $1/\sqrt{N}$ and $1/N$, although sufficiently large to warrant the dropping of terms of higher order (*i.e.*, terms of order $1/\sqrt{N}$, $1/N^2$, etc.), then the moment generating function of the distribution of X will be given approximately¹ by $\log G_X = -K_2 \frac{\theta^2}{2!} + K_3 \frac{(i\theta)^3}{3!} + K_4 \frac{\theta^4}{4!}$. From this generating function, the distribution of X is found to be²

¹ For K_5 , K_6 , and the higher cumulants of the generating function will all be of higher order than $1/N$; and, according to the assumption, terms of that order may be considered as negligible.

² If $\log G_X = -K_2 \frac{\theta^2}{2!} + K_3 \frac{(i\theta)^3}{3!} + K_4 \frac{\theta^4}{4!}$ and

$$G = \exp \left[\frac{-K_2 \theta^2}{2!} + \frac{K_3 (i\theta)^3}{3!} + \frac{K_4 \theta^4}{4!} \right],$$

then the distribution of x is given by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left[\frac{-K_2 \theta^2}{2!} + \frac{K_3 (i\theta)^3}{3!} + \frac{K_4 \theta^4}{4!} - i\theta x \right] d\theta$$

This may also be written, to within terms of order $1/N$,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[1 + \frac{K_3 (i\theta)^3}{3!} + \frac{K_4 \theta^4}{4!} \right] \exp \left[\frac{-K_2 \theta^2}{2!} - i\theta x \right] d\theta$$

But as noted in the footnote to page 97

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left[\frac{-K_2 \theta^2}{2!} - i\theta x \right] d\theta = \frac{1}{\sqrt{2\pi K_2}} \exp \left[\frac{-x^2}{2K_2} \right]$$

and successive differentiation with respect to x gives

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} (i\theta)^3 \exp \left[\frac{-K_2 \theta^2}{2!} - i\theta x \right] d\theta = \frac{d^3}{dx^3} \frac{1}{\sqrt{2\pi K_2}} \exp \left[\frac{-x^2}{2K_2} \right]$$

and

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \theta^4 \exp \left[\frac{-K_2 \theta^2}{2!} - i\theta x \right] d\theta = \frac{d^4}{dx^4} \left(\frac{1}{\sqrt{2\pi K_2}} \exp \left[\frac{-x^2}{2K_2} \right] \right)$$

Hence $f(x)$ may be written in operational form,

$$f(x) = \left[1 - \frac{K_3}{3!} \left(\frac{d}{dx} \right)^3 + \frac{K_4}{4!} \left(\frac{d}{dx} \right)^4 \right] \frac{1}{\sqrt{2\pi K_2}} \exp \left[\frac{-x^2}{2K_2} \right]$$

or, actually working out the differentiation,

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \left[1 + \frac{A}{\sigma^3} \left(\frac{3x}{\sigma} - \frac{x^3}{\sigma^3} \right) + \frac{B}{\sigma^4} \left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right) \right] \exp \left[-\frac{x^2}{2\sigma^2} \right] \quad (4)$$

where $x = X - \bar{X}$, σ is the standard deviation of $x = \sqrt{K_2}$, $A = -K_3/3!$, and $B = K_4/4!$

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \left[1 + \frac{A}{\sigma^3} \left(\frac{3x}{\sigma} - \frac{x^3}{\sigma^3} \right) + \frac{B}{\sigma^4} \left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right) \right] \exp \left[-\frac{x^2}{2\sigma^2} \right]$$

where $\sigma^2 = K_2$, $A = -K_3/3!$, and $B = K_4/4!$.

CHAPTER VI

SUMMARY OF THE THEORY OF FREQUENCY CURVES, AND SOME EXAMPLES

At the end of Chap. II a summary of the conditions leading to a normal curve was given. These will now be reviewed, and the conditions leading to nonnormality will be summarized.

Summary of Conditions Leading to Normality. The Pearsonian and Gram-Charlier analyses suggest that a variable will tend to be normally distributed if (1) its fluctuations are the sum of the fluctuations in a number of contributory causes; (2) the fluctuations in the contributory causes are all of about the same order of magnitude; (3) the contributory causes act independently of each other; and (4) the number of contributory causes is large, while the contribution of each is relatively small. These conditions, it will be presumed, are sufficient to produce a normally distributed variable. It is not to be presumed, however, that they are all necessary. It is possible, for example, that the presence of (4) may under some conditions make (3) unnecessary.

Summary of Conditions Giving Rise to Nonnormality. The conditions that are likely to give rise to nonnormal frequency curves are in general the negation of the conditions that give rise to normal frequency curves. Thus, distribution of a variable X may fail to be normal for any one of the following reasons:

1. If the deviations of X from some central value are simply an algebraic sum of the contributions of a number of causal factors, if these contributions are of about the same order of magnitude, and if they are independent of each other, the distribution of X may not be normal (even approximately) if the distributions of the individual contributions are themselves nonnormal and if the number of causal factors is not exceptionally large.

2. If the deviations of X from some central value are an algebraic sum of the contributions of a number of causal factors and if these act independently of each other, the distribution of X will not be normal if the contributions of a few of the causal

factors are of outstanding importance as compared with the others and if the distributions of these contributions are themselves not normal. If the magnitude of one contribution, for example, overshadowed that of all the others, then the distribution of X would tend to conform to the distribution of that particular contribution.

In this connection it is to be noted that, if one or more contributing factors are of outstanding importance compared with other causes of variation, the conditions producing variation in X might possibly be considered as nonhomogeneous. This is

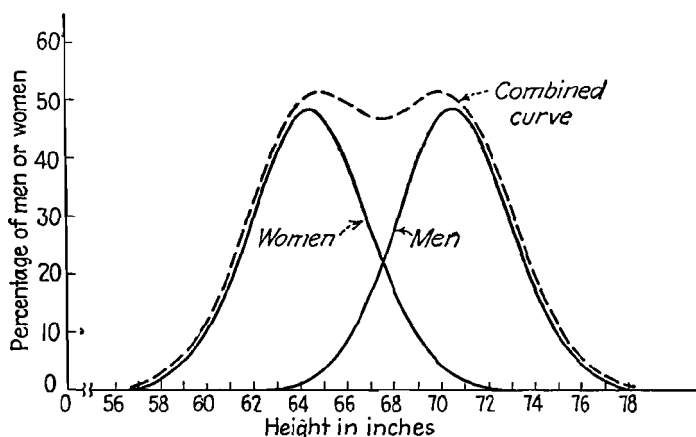


FIG. 25.—Combination of two normal distributions to form a nonnormal distribution (see Table 13).

especially likely to be true if the factors in question are qualitative in nature.

Consider, for example, the heights of adult human beings. Of all the single factors affecting mature height, sex and race are among the most important. Their effects are so outstanding that unless they are removed, by statistical classification, they overshadow the effects of most other causes of variation among mature persons. Furthermore, both sex and race are qualitative factors that make a contribution to height of one amount when one sex or race is present and a contribution of a distinctly different amount when another sex or race is present. The net effect is to produce a variable, *viz.*, adult height, that is not normally distributed.

The effects of sex and race being of this sort, it is the usual practice to view height in general as a heterogeneous variable whose frequency distribution has little significance. For when adult human beings are segregated according to sex and race, their heights form a distinctly normal distribution. • This serves to illustrate how a distribution may be nonnormal because the variable in question is nonhomogeneous.

TABLE 13.—AN ILLUSTRATION OF HOW THE COMBINATION OF TWO NORMAL DISTRIBUTIONS PRODUCED A NONNORMAL DISTRIBUTION

Height, in.	(1) Hypothetical distribution of heights of 300 males, frequency ¹	(2) Hypothetical distribution of heights of 300 females, frequency ²	(1) + (2), frequency
56—		.27	.27
57—		.91	.91
58—		2.59	2.59
59—		6.43	6.43
60—		13.69	13.69
61—		23.87	23.87
62—	.27	35.19	35.46
63—	.91	44.91	45.82
64—	2.59	48.45	51.04
65—	6.43	44.36	50.79
66—	13.69	34.62	48.31
67—	23.87	23.02	46.89
68—	35.19	12.83	48.02
69—	44.91	6.05	50.96
70—	48.45	2.47	50.92
71—	44.36	.86	45.22
72—	34.62	.24	34.86
73—	23.02		23.02
74—	12.83		12.83
75—	6.05		6.05
76—	2.47		2.47
77—	.86		.86
78—	.24		.24

¹ These are the ordinates of the normal curve fitted to the heights of the 300 Princeton freshmen of Table 20, Smith and Duncan, *Elementary Statistics and Application*, p. 301. They are approximately equal to the frequency.

² Assumed to be the same as (1), except that the mean height is 6 in. shorter.

Figure 25 and Table 13 offer an example of the production of a nonnormal distribution through the combination of two normal distributions with different mean values. This is what

the distribution of the heights of white adults would be like. It is of little significance because the important effect of the sex factor has not been separated from the minor effects of the many other factors influencing height.

3. If the deviations of X from some central value are an algebraic sum of the contributions of a number of causal factors and if these are all of about the same order of magnitude, the distribution of X will not be normal if the amount of the contribution of a causal factor is dependent on the contributions of other causal factors. It is possible, however, that even under these conditions the distribution of X will be approximately normal if the number of contributory causes is very large.¹

4. If the deviations in X from some normal value are not a simple algebraic sum of the contributions of a number of causal factors but are related to them in a more complex manner, it is likely that the distribution of X will not be normal. Thus suppose that the deviations in X are equal to the cube of the sum of the individual contributions instead of the sum itself. Then, although the sum may be normally distributed, X will be distributed in a skewed manner.²

This suggests that where certain "linear" measurements of a set of "individuals" are normally distributed, other "nonlinear" measurements of the same individuals will not be normally distributed. For example, the heights of adult males of the white race make up a set of homogeneous linear measurements, while the weights of these individuals compose a set of nonlinear measurements in the sense that the weight of an individual is highly correlated with his "volume," a quantity that is likely to vary with the cube of height rather than the height itself. In actuality, the heights of adult white males are found to be normally distributed, while their weights are definitely skewed, thus giving an empirical illustration of the relationship between the distribution of a sum and the distribution of its cube.

5. In some cases data that are selected from a larger normal group of data are themselves nonnormally distributed because

¹ Cf. p. 63. Also see MARKOV, A., *Wahrscheinlichkeitsrechnung* (B. G. Teubner, Leipzig, 1912), Appendices II and III.

² Cf. REITZ, H. L., *Mathematical Statistics*, pp. 72-74, and "Frequency Distributions Obtained by Certain Transformations of Normally Distributed Variates," *Annals of Mathematics*, Vol. 23 (1922), pp. 292-300.

of the special way in which they are selected. Suppose, for example, that the United States Army refuses to take adult males whose heights are less than 64 or greater than 74 inches. The distribution of the heights of Army men would then be a "truncated" normal distribution, such as that pictured in Fig. 26. The extension of the curve by a dotted line recognizes the possibility that a few exceptionally qualified men would be accepted in spite of the rule. Again, suppose that the students who select courses in higher mathematics are only the better

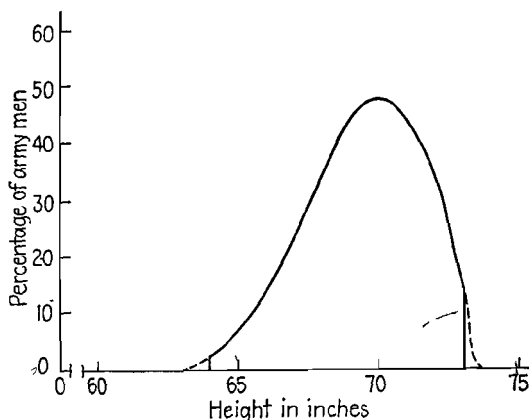


FIG. 26.—Nonnormal distribution of heights resulting from special selection.

students, say the upper third. Although the intelligence quotients of all students would be normally distributed, those of mathematics students would be more or less like the upper part of a normal curve. This is illustrated in Fig. 27, the rounded tail below 80 indicating that a few below that score would be allowed to enter the mathematics course despite the general rule. On the other hand, the distribution of I.Q.'s of students taking courses in a notoriously easy field of study would be more or less like the lower part of a normal curve, such as that pictured in Fig. 28. In all these cases no radical departure from the analysis of the previous chapter is required to explain the nonnormal character of the data. Nonnormality of the subgroup results only from special selection from a larger normal group.¹

¹ Note that the nonnormality of the subgroup arises from its selection, not at random, but with definite reference to the attribute of its members.

These are the principal reasons, it would appear, for the occurrence of nonnormal frequency distributions.

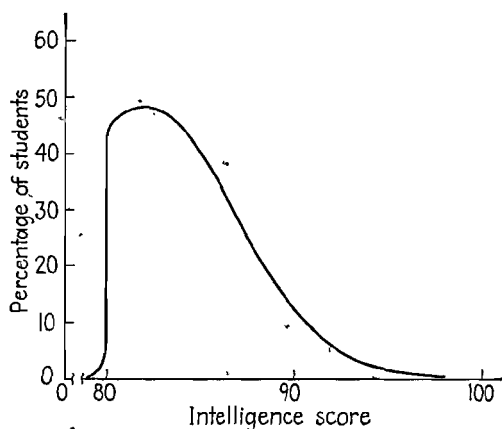


FIG. 27.—Nonnormal distribution of grades resulting from special selection—positively skewed.

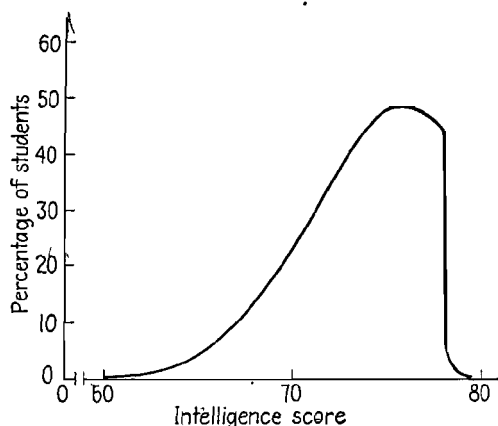


FIG. 28.—Nonnormal distribution of grades resulting from special selection—negatively skewed.

EXAMPLES OF NONNORMAL FREQUENCY DISTRIBUTIONS

Examples from Everyday Life. Figure 29 shows the distribution of weights of 300 Princeton freshmen of the class of 1943. Its general shape and the value of $\beta_1 = .36795$ and $\beta_2 = 4.6057$

A nonnormal subgroup may be obtained from a larger normal group in this way even though the latter is homogeneous.

(corrected for grouping) indicate a definite departure from normality. Since height is a linear measurement and since weight is related to volume, which is a cubical measurement, it is possible, as suggested in the previous section, that the departure of weights from normality is due to the variation in weight being equal to the cube of a sum of elementary variations rather than to the sum itself.

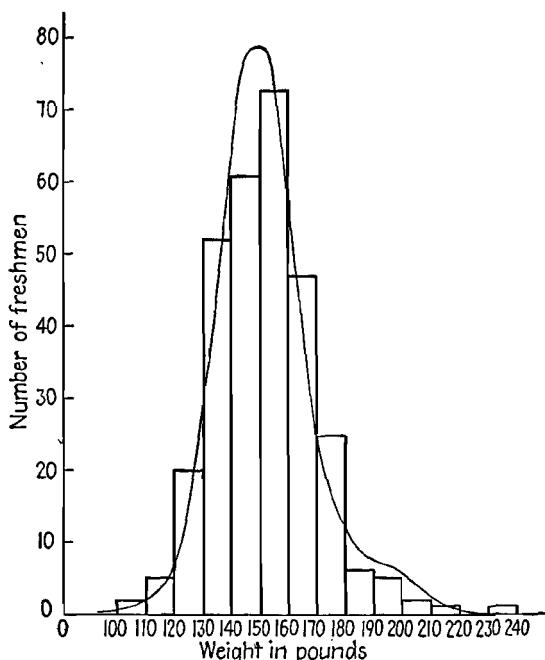


FIG. 29.—Histogram and fitted Gram-Charlier curve, distribution of weights of 300 Princeton freshmen.

The distribution of family incomes is a distinctly nonnormal distribution. This is illustrated by the distribution of family incomes in the United States in 1935–1936, as shown in Fig. 30. There are various causes for this departure from normality. Possibly one important cause is that, the more money a family has, the easier it is to get still more money. That is, it is likely that a principal cause of the nonnormality of the distribution of incomes is the lack of independence of the factors contributing to variation.

Examples from Sampling Analysis. Some of the better known forms of nonnormal frequency distributions are produced by the process of random sampling. The theory of sampling that provides the mathematical models for various types of sampling problems is discussed at some length in Parts II and III of this volume. It will be sufficient here to call attention to certain phases of sampling theory that are closely related to the discus-

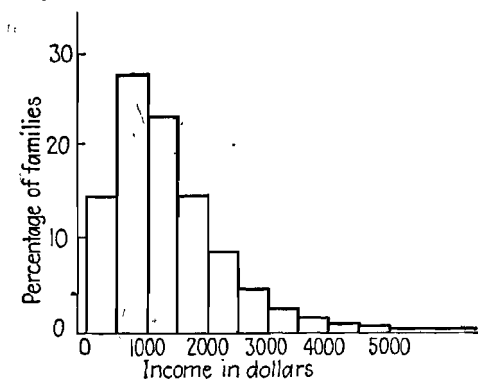


FIG. 30.—Distribution of family incomes in the United States, 1935-1936. (*National Resources Committee, Consumer Income in the United States*, p. 3.)

sion of the previous chapters and serve to illustrate the generation of nonnormal frequency curves. It is to be noted that some of the sampling distributions here discussed are nonnormal only when the samples are relatively small and tend toward normality as the size of the sample is increased. This tendency will be noted in the discussion so that the conditions producing normality and nonnormality will be clearly contrasted.

Sampling Distribution of the Mean for Any Population. The theoretical discussion of the previous sections offers considerable information about the distribution of the means of samples drawn at random from a large (theoretically infinite) population. For it will be noted that the mean is merely $1/N$ times the sum of the individual cases. The variations in the individual cases from sample to sample are thus the "contributory causes" (the ϵ 's of the Gram-Charlier analysis) of the sampling variation in the mean. Since the individual cases all come from the same population,¹ their individual distributions are all alike [that is,

¹ The population is assumed to be so large that the successive withdrawals

$f(\epsilon_1) = f(\epsilon_2) = f(\epsilon_3)$, etc.]. Hence the cumulants of the distribution of the mean will be $1/N^{k-1}$ times the cumulants of the population from which the samples are taken.¹

Since the second cumulant equals the variance, it follows that the variance of the distribution of the mean will equal $1/N$ times the variance of the population. Likewise, the third cumulant of the distribution of the mean (which equals the third moment of the mean) equals $1/N^2$ times the third cumulant, (i.e., moment) of the population, and the fourth cumulant of the distribution of the mean equals $1/N^3$ times the fourth cumulant of the population. Therefore, the distribution of sample means will have the same general form as the population from which the samples were drawn. Its variance, however, will be less, its skewness much less, and its kurtosis very much less than that of the population. In fact, if the sample is large, the skewness and kurtosis will be practically nonexistent and the distribution of sample means will become practically normal.

Sampling Distribution of Any Linear Function. What is true of the mean is also true of any statistic that is a linear function of the individual variables. For example, the regression coefficient $b_{12} = \Sigma x_1 x_2 / \Sigma x_2^2$. If samples are drawn from a bivariate population in such a manner that the x_2 values are always the same and only the x_1 values change, then b_{12} becomes merely a linear function of the sample x_1 's, viz.,

$$b_{12} = A_1 x_1^{[1]} + A_2 x_1^{[2]} + \dots + A_n x_1^{[n]}$$

where the A 's are dependent on the given values of the x_2 's and the numbers in the brackets differentiate the various sample values of x_1 . The cumulants of the sampling distribution of b_{12} are accordingly related to the cumulants of the individual x_1 's as follows:

$$\begin{aligned} K_2 &= A_1^2 k_{12} + A_2^2 k_{22} + \dots + A_n^2 k_{n2} \\ K_3 &= A_1^3 k_{13} + A_2^3 k_{23} + \dots + A_n^3 k_{n3} \\ K_4 &= A_1^4 k_{14} + A_2^4 k_{24} + \dots + A_n^4 k_{n4} \end{aligned}$$

of the members of a sample do not materially affect the distribution of probabilities in the population.

¹ The cumulants of the sum of the cases would be N times the cumulants of the population, and therefore the cumulants of the mean (which equals $1/N$ th of the sum) would be $1/N^{k-1}$ times the cumulants of the population.

Note that the k th cumulant of $f(X/N)$ is $1/N^k$ times the k th cumulant of $f(X)$.

in which k_{12} means the second cumulant of the first case in the sample of x_1 's, that is, $x_1^{[1]}$; k_{13} means the third cumulant of $x_1^{[1]}$; k_{n4} means the fourth cumulant of the n th case in the sample of x_1 's, that is, $x_1^{[n]}$; etc.

The variance of the sampling distribution of b_{12} will consequently be directly related to the variance of the individual x_1 's, being merely their weighted sum. Similarly, the skewness and kurtosis of the sampling distribution of b_{12} will be a weighted average of the skewness and kurtosis of the distributions of the individual x_1 's. If the distributions of the x_1 's for each x_2 are all alike in form (but not necessarily having the same mean, variance, etc.), the sampling distribution of b_{12} will have exactly the same sort (but not the same degree) of skewness and kurtosis as the individual x_1 's.

It is to be noted, however, that the "weights" that enter into the relationships between the cumulants of b_{12} and those of the individual x_1 's (i.e., the $A_1, A_2, \dots, A_n$) are equal to

$$\frac{x_2^{[1]}}{N\sigma_2^2}, \quad \frac{x_2^{[2]}}{N\sigma_2^2}, \quad \dots, \quad \frac{x_2^{[n]}}{N\sigma_2^2} \quad (\text{for } \Sigma x_2^2 = N\sigma_2^2)$$

So, if the variances of the individual x_1 's were identical or were about the same order of magnitude, the variance of b_{12} would be $1/N$ times the order of magnitude of the variances of the individual x_1 's and the third and fourth cumulants would be $1/N^2$ and $1/N^3$ times the order of magnitude of the variances of the individual x_1 's. Consequently, as in the case of the mean, the sampling distribution of the regression coefficient b_{12} will have much less skewness and much less kurtosis than the original distributions of the individual x_1 's. In fact, if the size of the sample is large, the distribution of b_{12} will be practically normal. This will be true whether the distributions of the individual x_1 's for various x_2 's are or are not normal or are or are not alike.

The t Distribution. A number of sample statistics from normal populations have sampling distributions that are nonnormal because they are not linear functions of the cases, although the distributions all approach normality as the size of the sample is increased. Three important nonnormal sampling distributions are the so-called " t distribution," the " χ^2 distribution," and the " F distribution."¹ The occasions on which these distributions

¹ The t distribution is also referred to as "Student's distribution," after

arise will be discussed in subsequent chapters. It is the purpose here merely to describe these three important nonnormal sampling distributions.

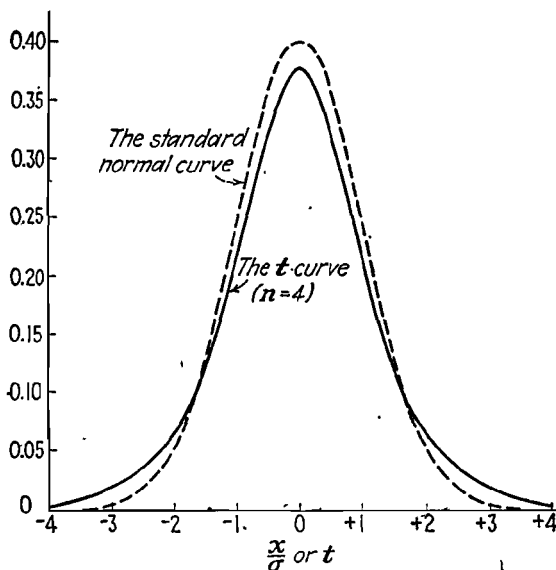


FIG. 31.—The standard normal curve compared to the curve of Student's distribution.

A graph of the t distribution is shown in Fig. 31¹. It will be noticed that the curve is symmetrical about the mean and looks very much like the normal curve. It has larger tails than the

the pseudonym of the man who first called attention to it. Sometimes the variable $z = \frac{1}{2} \log_e F$ is used instead of F . The distribution of $z = \frac{1}{2} \log_e F$ is shown on page 114.

¹ It will be noted that the vertical scales of figures showing discrete probabilities are labeled "Probabilities," or " $P(x)$," " $P(H)$," which mean "probability of an x " and "probability of an H ," etc. Vertical scales of figures showing probability curves cannot properly be so labeled; for in the graph of a probability curve the vertical distance is merely an ordinate. The probability of a case falling in a given range is the area under the curve for that range. If the vertical scale of the graph of a probability curve is labeled $f(x)$, then the probability, *i.e.*, $P(x)$, of a case falling in the infinitesimal range dx is given by $P(x) = f(x) dx$. For a finite range, $x_1 - x_2$, $P(x) = \int_{x_1}^{x_2} f(x) dx$. In the text (see page 116) y_1 , y_2 , y_{x^2} , etc., are used as equivalents of $f(t)$, $f(F)$, $f(x^2)$, etc. In this book, the vertical scales of probability curves are, generally speaking, not labeled.

normal curve, however, as can be seen from a comparison of the two curves.

The formula for the t distribution is

$$dP = \frac{\left(\frac{n-1}{2}\right)!}{\sqrt{n\pi} \left(\frac{n-2}{2}\right)! \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} dt \quad (1)$$

where n is a constant that determines the shape of the curve, just as $\bar{X}$ and σ do in the case of the normal curve.

In general, the formula says that the infinitesimal portion of the total area cut off by the infinitesimal class interval t to $t + dt$ is equal approximately to the area of the rectangle whose

height is $\frac{\left(\frac{n-1}{2}\right)!}{\sqrt{n\pi} \left(\frac{n-2}{2}\right)! \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}}$ and whose base is dt . The

t curve has its mode at 0 and tapers off symmetrically in both directions as t goes to $+\infty$ and $-\infty$. Its mean is 0 and its variance is $\frac{n}{n-2}$. Its skewness is zero, and its kurtosis is greater than 3 but approaches 3 as n increases. In general, the t curve approaches the normal curve as n increases; in fact, the standard normal curve is a good approximation to the t curve for values of $n > 30$.¹

The χ^2 Distribution. The χ^2 distribution is a positively skewed distribution, a picture of which is given in Fig. 32. The mathematical formula for the curve is

$$dP = \frac{1}{(2)^{\frac{n}{2}} \left(\frac{n-2}{2}\right)!} e^{\frac{-\chi^2}{2}} (\chi^2)^{\frac{n-2}{2}} d(\chi^2) \quad (2)$$

where n is a quantity that determines the shape and position of the curve. In general, the formula says that the infinitesimal portion of the total area cut off by the infinitesimal class interval

¹ A better approximation for values between 30 and 100, say, is given by taking $\sigma_t^2 = \frac{n}{n-2}$ instead of 1.

χ^2 to $\chi^2 + d\chi^2$ is equal approximately to the area of a rectangle whose height is $\frac{e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n-2}{2}}}{(2)^{\frac{n}{2}} \left(\frac{n-2}{2}\right)!}$ and whose base is $d(\chi^2)$.

The χ^2 curve begins at 0, rises to a peak at $\chi^2 = n - 2$, and falls again to zero as χ^2 goes to infinity. The mean of the curve

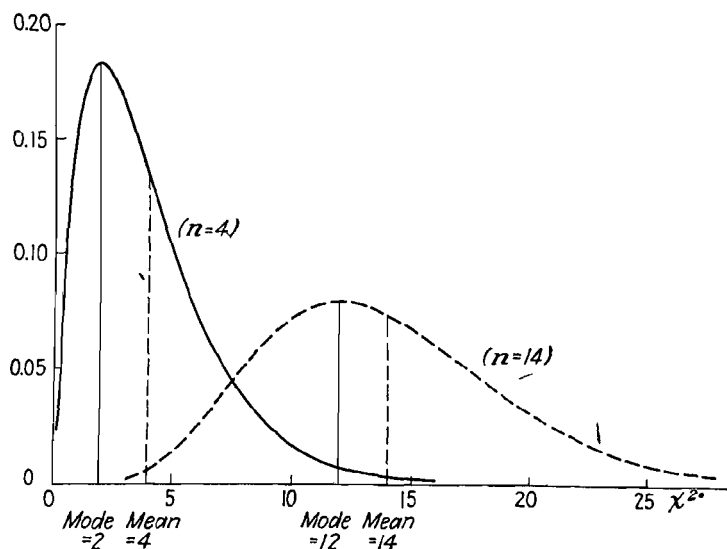


FIG. 32.—The χ^2 curves for $n = 4$ and for $n = 14$.

is at n , and its standard deviation is $\sqrt{2n}$. The skewness of the curve, as measured by $\frac{\bar{X}_{\chi^2} - \text{Mo}_{\chi^2}}{d_{\chi^2}}$, is $\sqrt{2/n}$. There are thus different χ^2 curves for different values of n . As n varies, the curve changes both its position and its shape. For larger values of n , the curve is located further along the χ^2 axis and is more spread out and more symmetrical. Figure 32 pictures two different χ^2 curves, one for $n = 4$, the other for $n = 14$. In general, as n gets larger, the χ^2 curve approaches the normal curve. The normal curve is, in fact, a special case of the χ^2 curve.

*The F Distribution.*¹ The F distribution is positively skewed like the χ^2 distribution. A picture of the F distribution is shown

¹ The reader is warned that the F of the F distribution and the F curve

in Fig. 33. The mathematical formula for the curve is as follows:

$$dP = \frac{\left(\frac{n_1 + n_2 - 2}{2}\right)! (n_1)^{\frac{n_1}{2}} (n_2)^{\frac{n_2}{2}} (F)^{\frac{n_1-2}{2}}}{\left(\frac{n_1-2}{2}\right)! \left(\frac{n_2-2}{2}\right)! (n_1 F + n_2)^{\frac{n_1+n_2}{2}}} dF \quad (3)$$

in which n_1 and n_2 are constants, such as n of the t curve and the χ^2 curve, which determine the character of the curve.

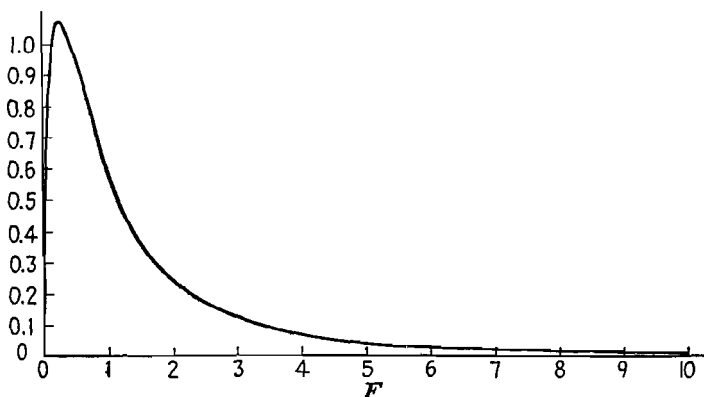


FIG. 33.—The F curve for $n_1 = 4$, $n_2 = 3$.

The formula for the F curve says that the infinitesimal portion of the total area cut off by the infinitesimal class interval F to $F + dF$ is equal to the area of an infinitesimal rectangle whose

height is $\frac{\left(\frac{n_1 + n_2 - 2}{2}\right)! (n_1)^{\frac{n_1}{2}} (n_2)^{\frac{n_2}{2}} (F)^{\frac{n_1-2}{2}}}{\left(\frac{n_1-2}{2}\right)! \left(\frac{n_2-2}{2}\right)! (n_1 F + n_2)^{\frac{n_1+n_2}{2}}}$ and whose base is dF .

The F curve starts at zero, rises to a peak at $\frac{n_2(n_1-2)}{n_1(n_2+2)}$ and falls again to zero as F goes to infinity. Its mean is $\frac{n_2}{n_2-2}$ and its standard deviation is $\frac{n_2}{n_2-2} \sqrt{\frac{2(n_2+n_1-2)}{n_1(n_2-4)}}$. The curve thus varies with n_1 and n_2 . As n_1 and n_2 get larger, the F curve

formula has no relationship to the F symbol employed to represent frequencies.

tends to become more symmetrical; and as n_1 and n_2 both approach infinity, the F curve approaches the normal curve. If one of the n 's approaches infinity, while the other remains small, the F curve approaches the χ^2 curve. If $n_1 = 1$ and n_2 approaches infinity, the distribution of $\sqrt{F}$ approaches the t distribution. Thus the normal curve, the t curve, and the χ^2 curve are all special cases of the F curve.¹

¹ As already pointed out on p. 109*n.*, the variable $z = \frac{1}{2} \log_e F$ is sometimes used instead of F . It is this z distribution that R. A. Fisher has called attention to as the one general sampling distribution. Cf. Paul R. Rider, *An Introduction to Modern Statistical Methods* (1939), who cites J. O. Irwin, "Mathematical Theorems Involved in the Analysis of Variance," *Journal of the Royal Statistical Society*, Vol. 94 (1931), pp. 287*ff.*; and R. A. Fisher, "On the Distribution Yielding the Error Functions of Several Well-known Statistics," *Proceedings of the International Mathematical Congress* (Toronto, 1924), pp. 805-813. The accompanying figure, numbered 34, is a picture of the z distribution for $n_1 = 4$, $n_2 = 3$. It will be noted that the logarithmic transformation produces a much more symmetrical distribution.

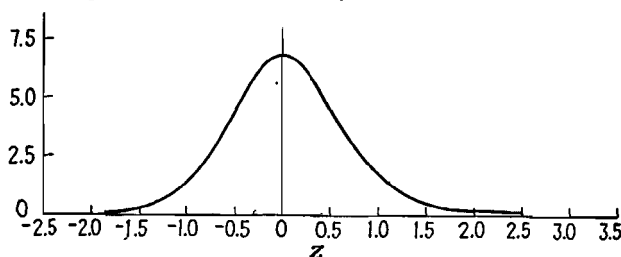


FIG. 34.—The z curve for $n_1 = 4$, $n_2 = 3$.

CHAPTER VII

NUMERICAL CALCULATIONS FOR FREQUENCY CURVES

Up to this point the discussion has been primarily concerned with the how and the why of frequency curves, *i.e.*, with theory per se. This chapter will show how frequency curves may be graphed, how frequencies and probabilities may be computed from frequency curves, how curves may be fitted to sample data, and how the goodness of fit may be tested. Attention will thus center in numerical rather than in abstract calculations.

GRAPHING FREQUENCY CURVES

It is sometimes desired for instruction purposes or for illustration to make graphs of some of the better known frequency curves. This section will indicate how such graphs may readily be made. The purpose is to show how a curve may be plotted after the constants of the curve have been determined. The problem of determining numerical values for the constants themselves will be discussed in the third section on curve fitting.

The Normal Curve. Probably the easiest curve to graph is the standard normal curve. Tables of the ordinates of this curve have been computed for various abscissa values. Such a table is Table VI in the Appendix. This gives values of

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

for various values of x/σ . To graph a normal curve it is necessary merely to plot these ordinates at selected values of x/σ and draw a curve through the tops of the ordinates. This has been done in Fig. 31. In making a quick freehand graph it may be noted that the curve is symmetrical, its mean and mode come at $x/\sigma = 0$, its points of inflection are at $x/\sigma = 1$, and its tails stretch to plus and minus infinity.

The standard normal curve is a normal curve whose mean is zero and whose standard deviation is 1 (since the abscissa

scale is in terms of x/δ units). To graph a normal curve with a given mean it is necessary merely to place the zero point of the standard curve at the mean and to adjust the abscissa scale for the difference in standard deviations. Further details and special adjustments are described in the section on testing goodness of fit of a frequency curve.¹

The t , χ^2 , and F Curves. For most other curves no tables of ordinates have been computed. Graphs of these curves must therefore be made directly from the equations for the curves. Most of the equations are such that it is easiest to find the logarithms of the ordinates first and then convert these to anti-logarithms. This method will now be illustrated for three important nonnormal sampling distributions, *viz.*, the t curve, the χ^2 curve, and the F curve.

The equations for these curves are

$$y_t = \frac{\left(\frac{n-1}{2}\right)!}{\sqrt{n\pi} \left(\frac{n-2}{2}\right)! \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} \quad (1)$$

$$y_{\chi^2} = \frac{e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n-2}{2}}}{(2)^{\frac{n}{2}} \left(\frac{n-2}{2}\right)!} \quad (2)$$

$$y_F = \frac{\left(\frac{n_1 + n_2 - 2}{2}\right)! (n_1)^{\frac{n_1}{2}} (n_2)^{\frac{n_2}{2}} F^{\frac{n_1-2}{2}}}{\left(\frac{n_1-2}{2}\right)! \left(\frac{n_2-2}{2}\right)! (n_1 F + n_2)^{\frac{n_1+2}{2}}} \quad (3)$$

If logarithms are taken to the base 10, these equations become

$$\log y_t = \log \left(\frac{n-1}{2}\right)! - \frac{1}{2} \log n - \frac{1}{2} \log \pi - \log \left(\frac{n-2}{2}\right)! - \frac{n+1}{2} \log \left(1 + \frac{t^2}{n}\right) \quad (1')$$

$$\log y_{\chi^2} = -\frac{\chi^2}{2} \log e + \frac{n-2}{2} \log \chi^2 - \frac{n}{2} \log 2 - \log \left(\frac{n-2}{2}\right)! \quad (2')$$

¹ See pp. 137-152.

$$\begin{aligned} \log y_F = & \log \left(\frac{n_1 + n_2 - 2}{2} \right)! + \frac{n_1}{2} \log n_1 + \frac{n_2}{2} \log n_2 \\ & + \frac{n_1 - 2}{2} \log F - \log \left(\frac{n_1 - 2}{2} \right)! - \log \left(\frac{n_2 - 2}{2} \right)! \\ & - \frac{n_1 + n_2}{2} \log (n_1 F + n_2) \quad (3') \end{aligned}$$

In using these equations it is to be noted that $\log_{10} 2 = .301030$, $\log_{10} e = .434295$, and $\log_{10} \pi = .49715-$; logarithms of factorials can be obtained from tables of the gamma function¹ given in *Tracts for Computers* (Cambridge University Press, London, 1921) No. IV. For small values of n and n_1 and n_2 , the factorials may easily be computed directly. For example, if $n = 8$, then $\frac{n-2}{2}! = \frac{8-2}{2}! = 3! = 3 \times 2 \times 1 = 6$. If $n = 7$, then $\frac{n-2}{2}! = \frac{5}{2} \times \frac{3}{2} \times \frac{1}{2} \sqrt{\pi}$ (see page 255). For selected values of n or n_1 and n_2 , Eqs. (1'), (2'), and (3') will give values of $\log y$ for various values of t , χ^2 , and F , and values of y can be computed by taking antilogarithms. Tables 14 to 16 illustrate this process for each of these three curves and Figs. 31 to 33 of Chap. VI show the final graphs.²

¹ By definition $(n-1)! = \Gamma(n)$ for both integral and fractional values of n . See pp. 79n., 133-134.

² Ordinates for the z distribution may be obtained from the F distribution as follows. For the values of F for which ordinates have been computed, the corresponding values of z may be obtained from the relationship

$$z = \frac{1}{2} \log_e F = 1.1513 \log_{10} F$$

The z ordinates for these values may be computed by multiplying the corresponding F ordinates by $2e^{2z}$, or the logarithms of the z ordinates may be obtained by adding $.4343 (\log_{10} 2 + \log_{10} F)$ to the logarithms of the F ordinates.

If z is to be computed directly, the equation to be used is

$$\begin{aligned} \log z = & \log 2 + \log \left(\frac{n_1 + n_2 - 2}{2} \right)! + \frac{n_1}{2} \log n_1 + \frac{n_2}{2} \log n_2 + n_1 z \log_{10} e \\ & - \log \left(\frac{n_1 - 2}{2} \right)! - \log \left(\frac{n_2 - 2}{2} \right)! - \frac{n_1 + n_2}{2} \log (n_1 e^{2z} + n_2) \end{aligned}$$

Values of e^{2z} can be found from tables of e^x given in most books of mathematical tables. This was how the figure in the footnote to p. 114 was derived.

Other Curves. The graphing of other frequency curves presents about the same sort of problems as those discussed above. For example, when the numerical equation for a Gram-Charlier curve has been determined,¹ its graphing is readily accomplished by the use of the tables of ordinates of the standard normal curve and ordinates of its derivatives. These are to be found in the Appendix, Table VI. Their use is discussed below when the testing of the goodness of fit of a Gram-Charlier curve is described.

TABLE 14.—CALCULATION OF ORDINATES OF THE t CURVE
($n = 4$)

(1) t	(2) t^2	(3) $1 + \frac{t^2}{n}$	(4) $\log (3)$	(5) $-\frac{n+1}{2} (4)$	(6) (5) in another form	(7) antilog (6)	(8)* $K_t \cdot (7)$
.0	.00	1.0000	.0000	.00000	.00000	1.0000	.375
.2	.04	1.0100	.00432	-.01080	9.98920 - 10	.9754	.366
.4	.16	1.0400	.01703	-.04258	9.95742 - 10	.9066	.340
.6	.36	1.0900	.03743	-.09358	9.90642 - 10	.8062	.302
.8	.64	1.1600	.06446	-.16115	9.83885 - 10	.6900	.259
1.0	1.00	1.2500	.09691	-.24228	9.75772 - 10	.5724	.215
1.2	1.44	1.3600	.13354	-.33385	9.66615 - 10	.4636	.174
1.4	1.96	1.4900	.17319	-.43297	9.56703 - 10	.3690	.138
1.7	2.89	1.7225	.23616	-.59040	9.40960 - 10	.2568	.096
2.0	4.00	2.0000	.30103	-.75258	9.24742 - 10	.1768	.066
2.5	6.25	2.5625	.40866	-1.02168	8.97832 - 10	.0951	.036
3.0	9.00	3.2500	.51188	-1.27970	8.72030 - 10	.0525	.020
4.0	16.00	5.0000	.69897	-1.74742	8.25258 - 10	.0179	.007
5.0	25.00	7.2500	.86034	-2.15085	7.84915 - 10	.0071	.003

$$* K_t = \frac{\left(\frac{n-1}{2}\right)!}{\left(\frac{n-2}{2}\right)! \sqrt{\pi}}, \text{ which, for } n = 4, \text{ equals } .375. \text{ The logarithm of } K_t \text{ could be}$$

added to (5) before taking antilogarithms, but if a calculating machine is available it is probably easier to follow the outline of the table.

When once the type of a Pearsonian curve has been determined and the numerical values of its constants have been computed,² the curve can usually be graphed in much the same way as a t curve, χ^2 curve, or F curve. The process usually consists of taking logarithms, setting up a table to compute the logarithms

¹ See pp. 133, and 142-144.

² See pp. 134, and 146-150.

of the ordinates, and then finding the antilogarithms. The process is illustrated below (pages 148-150).

TABLE 15.—CALCULATION OF THE ORDINATES OF THE χ^2 CURVE
($n = 14$)

(1) χ^2	(2) .2172 χ^2	(3) $\log \chi^2$	(4) $\frac{n-2}{2} \log \chi^2$	(5) $K\chi^2 - (2) + (4)^*$	(6) (5) in another form	(7) antilog (6)
4.0	.8688	.6021	3.6126	-2.2205	8.7795 - 10	.060
5.0	1.0860	.6990	4.1940	-1.8563	9.1437 - 10	.139
6.0	1.3032	.7782	4.6692	-1.5983	9.4017 - 10	.252
7.0	1.5204	.8451	5.0706	-1.4141	9.5859 - 10	.385
8.0	1.7376	.9031	5.4186	-1.2833	9.7167 - 10	.521
9.0	1.9548	.9542	5.7252	-1.1939	9.8061 - 10	.640
10.0	2.1720	1.0000	6.0000	-1.1363	9.8637 - 10	.731
11.0	2.3892	1.0414	6.2484	-1.1051	9.8949 - 10	.785
12.0	2.6064	1.0792	6.4752	-1.0955	9.9045 - 10	.803
13.0	2.8236	1.1139	6.6834	-1.1045	9.8955 - 10	.786
14.0	3.0408	1.1461	6.8766	-1.1285	9.8715 - 10	.744
15.0	3.2580	1.1761	7.0566	-1.1657	9.8343 - 10	.683
16.0	3.4752	1.2041	7.2246	-1.2149	9.7851 - 10	.610
17.0	3.6924	1.2304	7.3824	-1.2743	9.7257 - 10	.532
18.0	3.9096	1.2553	7.5318	-1.3421	9.6579 - 10	.455
19.0	4.1268	1.2788	7.6728	-1.4183	9.5817 - 10	.382
20.0	4.3440	1.3010	7.8060	-1.5023	9.4977 - 10	.315
21.0	4.5612	1.3222	7.9332	-1.5923	9.4077 - 10	.256
22.0	4.7784	1.3424	8.0544	-1.6883	9.3117 - 10	.205
23.0	4.9956	1.3617	8.1702	-1.7897	9.2103 - 10	.162
24.0	5.2128	1.3802	8.2812	-1.8959	9.1041 - 10	.127
25.0	5.4300	1.3979	8.3874	-2.0069	8.9931 - 10	.098
26.0	5.6472	1.4150	8.4900	-2.1215	8.8785 - 10	.076
27.0	5.8644	1.4314	8.5884	-2.2403	8.7597 - 10	.058

* $K\chi^2 = -.1505n - \log \left(\frac{n-2}{2} \right)!$, which, for $n = 14$, equals -4.9643 .

COMPUTATIONS OF PROBABILITIES

The computation of relative frequencies or probabilities for any frequency curve consists in finding the area under the curve for the stated range of values. This may be accomplished theoretically by a process of summation, or "integration." For the more important curves employed in sampling analysis these integrations have been worked out by the mathematicians and the results published in a series of tables.¹ The average student

¹ The integrations are generally complicated and involve a process of approximation through the use of rapidly convergent infinite series.

therefore need only know how to use these tables in order to compute the desired probabilities; it is not necessary for him to be an accomplished mathematician.

TABLE 16.—CALCULATION OF THE ORDINATES OF THE F DISTRIBUTION
($n_1 = 4, n_2 = 3$)

(1) F	(2) $\log F$	(3) $\frac{n_1 - 2}{2} \log F$	(4) $n_1 F + n_2$	(5) $\frac{\log}{(n_1 F + n_2)}$	(6) $\frac{n_1 + n_2}{2} \times (5)$	(7) $K_F + (3) - (6)^*$	(8) antilog (7)
.01	-2.000000	-2.000000	3.04	0.482874	1.690059	9.142335 - 10	.139
.02	-1.698970	-1.698970	3.08	.488551	1.709928	9.323596 - 10	.211
.05	-1.301030	-1.301030	3.20	.505150	1.768025	9.663239 - 10	.461
.10	-1.000000	-1.000000	3.40	.531479	1.860176	9.872218 - 10	.745
.20	-.698970	-.698970	3.80	.570784	2.029244	.004180	1.010
.30	-.522879	-.522879	4.20	.623249	2.176372	.033143	1.079
.40	-.397940	-.397940	4.60	.662758	2.319153	.015301	1.035
.50	-.301030	-.301030	5.00	.698970	2.446395	9.984969 - 10	.966
.60	-.221849	-.221849	5.40	.732394	2.566379	9.944100 - 10	.879
.70	-.154902	-.154902	5.80	.763428	2.671998	9.905496 - 10	.804
.80	-.096910	-.096910	6.20	.792392	2.773372	9.862112 - 10	.728
.90	-.045757	-.045757	6.60	.819544	2.867854	9.818783 - 10	.659
1.00	0	0	7.00	.845098	2.957843	9.774551 - 10	.595
1.10	.041393	.041393	7.40	.869232	3.042312	9.731475 - 10	.539
1.20	.079181	.079181	7.80	.892094	3.122329	9.689246 - 10	.489
1.30	.113943	.113943	8.20	.913814	3.198349	9.647988 - 10	.445
1.50	.176091	.176091	9.00	.954243	3.339851	9.568634 - 10	.370
1.70	.230449	.230449	9.80	.991226	3.469291	9.493552 - 10	.312
2.00	.301030	.301030	11.00	1.041393	3.644876	9.389647 - 10	.245
2.50	.397940	.397940	13.00	1.113943	3.898801	9.231533 - 10	.170
3.00	.477121	.477121	15.00	1.176091	4.116313	9.093202 - 10	.124
3.50	.540680	.540680	17.00	1.230449	4.306572	8.966502 - 10	.093
4.00	.602060	.602060	19.00	1.278754	4.475639	8.858815 - 10	.072
5.00	.698970	.698970	23.00	1.361728	4.766048	8.665415 - 10	.046
10.00	1.000000	1.000000	43.00	1.633468	5.717138	7.015256 - 10	.001

$$* K_F = \log \left[\frac{\left(\frac{n_1 + n_2 - 2}{2} \right)! n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}}}{\left(\frac{n_1 - 2}{2} \right)! \left(\frac{n_2 - 2}{2} \right)!} \right] \quad \text{which, when } n_1 = 4, n_2 = 3, \text{ is equal to}$$

2.732394.

The Normal Curve. Table VI in the Appendix gives the area under the normal curve between the mean point $x/\sigma = 0$ and selected values of x/σ . Since the curve is symmetrical, the areas are the same for plus deviations as for minus deviations.

To illustrate the use of the table consider a normal curve whose mean is 100 and whose standard deviation is 10. Since the mean of the standard normal curve is taken as its origin, 100 will in this case be the origin from which deviations will be measured.

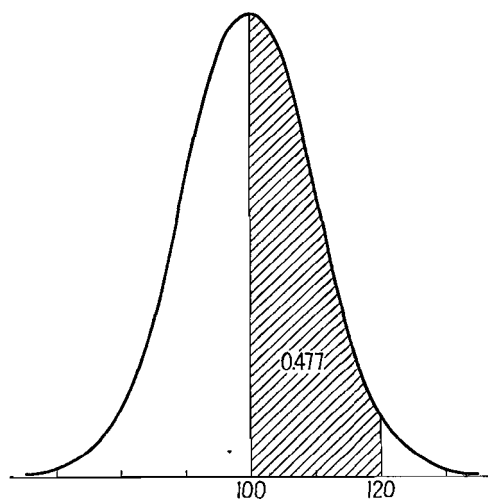


FIG. 35a.—Proportion of area under a normal curve between $\frac{x}{\sigma} = 0$ and $\frac{x}{\sigma} = 2$,
 $\bar{X} = 100$, $\sigma = 10$.

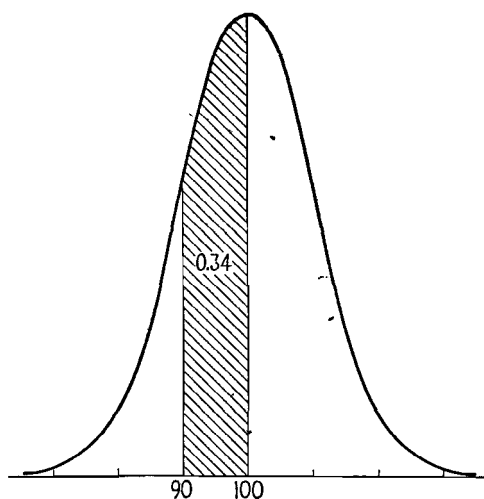


FIG. 35b.—Proportion of area under a normal curve between $\frac{x}{\sigma} = -1$ and $\frac{x}{\sigma} = 0$, $\bar{X} = 100$, $\sigma = 10$.

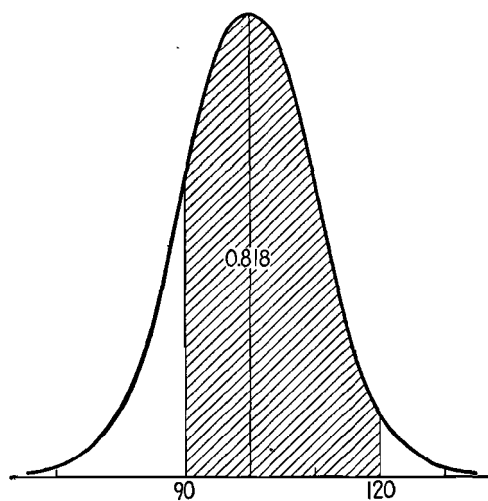


FIG. 35c.—Proportion of area under a normal curve between $\frac{x}{\sigma} = -1$ and $\frac{x}{\sigma} = 2$, $\bar{X} = 100$, $\sigma = 10$.

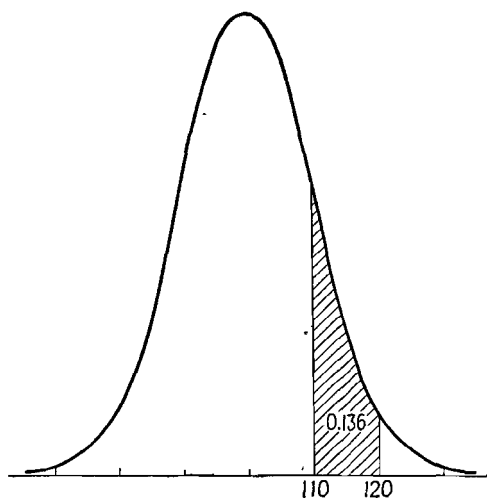


FIG. 35d.—Proportion of area under a normal curve between $\frac{x}{\sigma} = 1$ and $\frac{x}{\sigma} = 2$, $\bar{X} = 100$, $\sigma = 10$.

Since 10 is the standard deviation, the deviations from the mean will be measured in multiples of 10. To find the probability of a case falling in the interval 100 to 120, for example; set $\frac{x}{\sigma} = \frac{120 - 100}{10} = 2$ and note from Table VI of the Appendix that the probability of a normally distributed variate falling between 0 and 2 σ from the mean is .477. Hence the probability of a case falling between 100 and 120 is .477. To find the probability of a case falling between 90 and 100 set

$$\frac{x}{\sigma} = \frac{90 - 100}{10} = -1$$

and note that the probability of a normally distributed variate falling between 0 and -1 σ (same as between 0 and +1) is .341. Hence the probability of a case falling between 90 and 100 is .341. To find the probability of a case falling between 90 and 120, it is necessary merely to add the probability of a case falling between 90 and 100 to the probability of a case falling between 100 and 120. Thus the probability of a case falling between 90 and 120 is .477 + .341 = .818. To find the probability of a case lying between 110 and 120 it is necessary merely to subtract the probability of a case falling between 100 and 110 from the probability of a case lying between 100 and 120. Since the probability of a case falling between 100 and 110 is the same as the probability of a case falling between 90 and 100, it follows that the probability of a case falling between 110 and 120 is

$$.477 - .341 = .136.$$

These computations are pictured in Figs. 35a to 35d.

The t Curve. Normal curves may be drawn with different means and different standard deviations, but when the variable has been measured in standard deviation units and is measured from the mean as an origin all normal curves become one and the same curve, *i.e.*, the standard normal curve. It was therefore possible to construct a simple table which gave areas under the curve (*i.e.*, probabilities) for various x/σ deviations from the mean.

The t curve is inevitably used in the standard form. Even in that form, however, its shape depends on the parameter n . Hence areas under the curve, *i.e.*, the curve probabilities, will be

different for different values of n . A t table is thus a threefold table, listing the areas or probabilities that correspond to different deviations from the mean value for various values of n .

A typical t table will be found in the Appendix, Table VII. In contrast to the normal table, the deviations from the mean are given for selected probabilities, rather than the reverse. These selected probabilities are probabilities of an equal or greater absolute deviation, not probabilities of an equal or smaller deviation, as is true of the normal table. That is, the t table selects

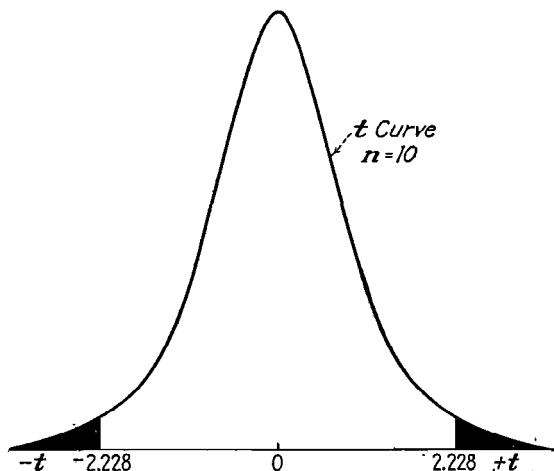


FIG. 36.—Equal probability areas of .025 in the tails of the t curve for $n = 10$. areas under the tail of the curve and gives for various values of n the absolute deviations from the mean that will yield these tail areas. Since the curve is symmetrical, the areas are equally divided between the two tails.

An example will illustrate the use of the table. Suppose the parameter n has the value 10, and it is desired to find the absolute deviation from the mean that will yield a tail area of .05. To find this deviation, locate the row $n = 10$, and proceed to the right until the column marked .05 is reached. The figure so located is the deviation desired. It will be seen to have the value 2.228. This means that a deviation of ± 2.228 will mark off an area of .025 on the upper tail and a deviation of -2.228 will mark off an area of .025 on the lower tail and the two deviations together will mark off an area of .025 on each tail, or a total area of .05. This is illustrated in Fig. 36.

The χ^2 Curve. The χ^2 table, Table VIII in the Appendix, is very much like the t table. It is again a threefold table; giving areas (probabilities) and deviations for various values of n . Also, like the t table, it lists deviations for selected areas or probabilities, and these areas refer to the portion of the curve beyond the deviation. That is, the areas represent probabilities of an equal or greater deviation. Unlike the t table the deviations are

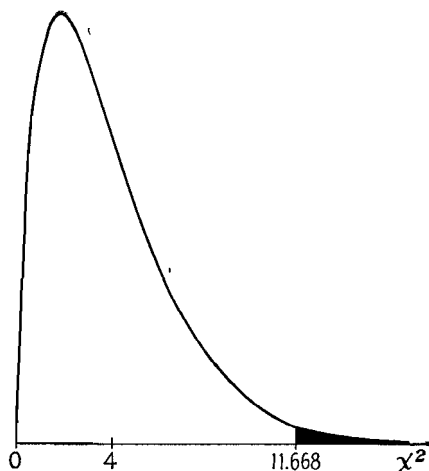


FIG. 37a.—The 2 per cent area in the upper tail of a χ^2 curve, $n = 4$ [.02 is the $P(\chi^2 \geq 11.668)$].

measured from the absolute origin of 0 and not from the mean of the curve.

To illustrate the use of the χ^2 table, let $n = 4$, and let the problem be to find the deviation from 0 for which the probability of an equal or greater deviation is .02. To find this deviation, locate the row $n = 4$, and proceed to the right until the column headed .02 is reached. The figure so located is the deviation desired. It will be seen to have the value 11.668 (see Fig. 37a):

Consider another problem. Suppose it is desired to find the deviation from 0 for which the area under the curve is just .02. That is, it is desired to find the deviation for which the probability of an equal or *smaller* value is just .02. Let n be 4 as before. To find this deviation, locate the row $n = 4$, and proceed to the column headed .98; the figure so located will be the desired deviation. It will be seen to have the value .711. That is, the deviation for which 98 per cent of the curve lies to the

right is also the deviation for which 2 per cent of the curve lies to the left (see Fig. 37b). It will be noted that the medians of

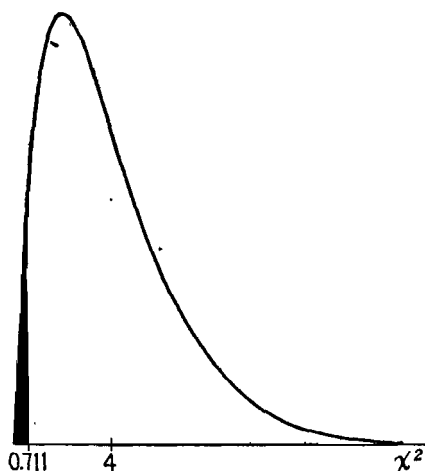


FIG. 37b.—The 2 per cent area in the lower tail of a χ^2 curve, $n = 4$ [.98 is the $P(\chi^2 \geq 0.711)$].

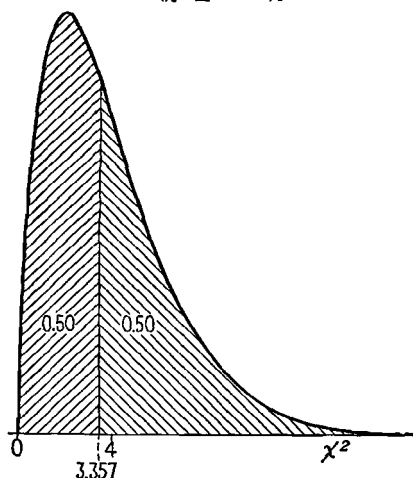


FIG. 37c.—The median point of a χ^2 curve, $n = 4$ [.50 is the $P(\chi^2 \leq 3.357) = P(\chi^2 \geq 3.357)$].

various χ^2 curves are the deviations in the column headed .50* (see Fig. 37c).

* The mean, it will be recalled, is n and the mode $n - 2$. See Chap. VI, pp. 111-112.

The F Curve. The F table (see Appendix, Table IX) is like the χ^2 table except that it takes account of two parameters, n_1 and n_2 , instead of one. It is thus a fourfold instead of a threefold table. Because of this greater complexity, it gives values for only two tail areas or probabilities, viz., the upper .05 tail and the upper .01 tail.

The use of the F table may be illustrated by the following problem: Let $n_1 = 4$ and $n_2 = 3$, and let it be desired to find the deviation for which the tail area (or probability of an equal or greater value) is .05. To find this deviation, first select the

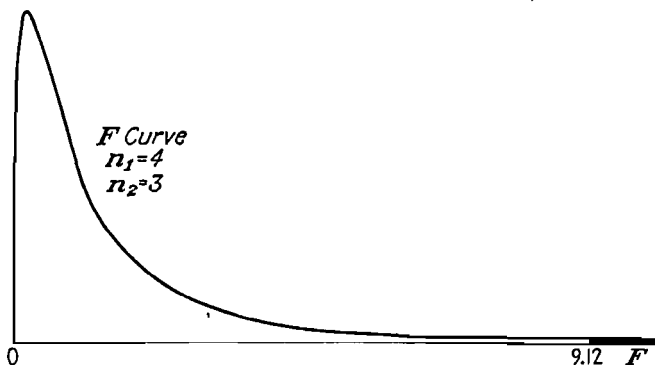


FIG. 38.—Probability area of .05 in the upper tail of an F curve, $n_1 = 4$, $n_2 = 3$ [.05 is the $P(F \geq 9.12)$].

row $n_2 = 3$, and proceed to the right until the column headed $n_1 = 4$ is reached. The lightface figure (the adjacent boldface figure is the “.01 point”) so located is the deviation desired. The “.05 point” is seen to be 9.12. The “.05 point” is shown in Fig. 38. If the problem were to find the deviation for which the tail area or probability of an equal or greater area were just .01, then the “.01 point” would have to be used; the rest of the procedure would be the same.

Other Curves. Special tables have been derived for the computation of areas and probabilities for curves belonging to the Gram-Charlier system. These may best be explained, however, in later sections¹ dealing with testing the goodness of fit of a Gram-Charlier curve. It can be remarked here that the tables are generally similar to the table of the normal curve.

The tables of the normal, t , χ^2 , and F curves, together with the tables used in computing probabilities for a Gram-Charlier

¹ See pp. 139-150.

curve, comprise the most commonly used tables of probabilities. When tables are lacking, the areas or probabilities must be computed directly or indirectly from the mathematical equation for the curve. If the equation is a simple one, the area may be found by the application of the integral calculus. Often, however, the equation for frequency curves make it difficult to apply the integral calculus, and some approximate method must be employed. If an integrator is available, the curve need only be plotted carefully and the integrator run around the desired area. If an integrator is not at hand, the area may be approximated from some "quadrature" equation that expresses the area of a given interval in terms of the ordinates of the curve for that interval and for the neighboring intervals.

Some of the more important quadrature equations given by W. P. Elderton are as follows:¹

$$\text{Area under the curve from } x = -\frac{1}{2} \text{ to } x = +\frac{1}{2} \\ \text{approximates } y_0 - \frac{1}{2}(\Delta y_{-1} - \Delta y_0) \quad (4)$$

or, if greater accuracy is desired, a nearer approximation is obtained by using

$$y_0 - \frac{291}{5760}(\Delta y_{-1} - \Delta y_0) + \frac{17}{5,760}(\Delta y_{-2} - \Delta y_1) \quad (5)$$

In these expressions, $x = 0$ is taken as the middle point of the interval for which the area is to be computed and $x = -\frac{1}{2}$ and $x = +\frac{1}{2}$ are the values of the lower and upper limits of the interval. The units are thus class-interval (d/i) units, and it is in these that the area is measured. The symbol y_0 stands for the ordinate of the curve at $x = 0$, that is, at the mid-point of the interval; Δy_0 means the difference between the ordinate y_1 at $x = 1$ (i.e., the ordinate at the mid-point of the next higher interval) and the ordinate y_0 at $x = 0$; Δy_1 means the difference between the ordinate y_2 at the mid-point of the second next higher interval and the ordinate y_1 at the mid-point of the next higher interval; Δy_{-1} means the difference between the ordinate at $x = 0$ and the ordinate at $x = -1$ (i.e., the ordinate at the mid-point of the next lower interval); and Δy_{-2} means the difference between the ordinate at the mid-point of the next lower interval and the ordinate at the mid-point of the second

¹ *Frequency Curves and Correlation*, pp. 25-26, 48.

lower interval. In short, if y_2, y_1, y_0, y_{-1} , and y_{-2} represent the ordinates at $x = 2, x = 1, x = 0, x = -1$, and $x = -2$, then $\Delta y_1 = y_2 - y_1$, $\Delta y_0 = y_1 - y_0$, $\Delta y_{-1} = y_0 - y_{-1}$, and $\Delta y_{-2} = y_{-1} - y_{-2}$.

To illustrate the use of these equations suppose that ordinates of a frequency curve for five values of the variable are as follows¹ (see Fig. 39):

X	y
105	1.21
115	3.40
125	16.10
135	47.57
145	78.87

Before proceeding further it is to be noted that the frequency curve from which these ordinates have been taken was drawn so that the curve would fit a histogram in which the area of an interval was represented by the heights of the rectangles and not the area. That is, the ordinates are actually one class interval (here the class interval is five) times greater than they should be if the area under the curve is to be correct. It may also be noted that in this case the curve gives the distribution of 300 cases and is drawn so that the total area is, not 1, but 300.

To find the area under the given curve for the interval whose mid-point is 125, set up the following table:

x	$x \left(= \frac{d}{i} \right)$	y	Δy
105	-2	1.21	$y_{-1} - y_{-2} = 3.40 - 1.21 = 2.19$
115	-1	3.40	$y_0 - y_{-1} = 16.10 - 3.40 = 12.70$
125	0	16.10	$y_1 - y_0 = 47.57 - 16.10 = 31.47$
135	1	47.57	$y_2 - y_1 = 78.90 - 47.57 = 31.33$
145	2	78.90	

This is shown graphically in Fig. 39.

By using Eq. (4) a first approximation to the area will be given by

$$16.10 - \frac{1}{2}(12.70 - 31.47) = 16.10 + .78 = 16.88$$

¹ Actually, these are the ordinates of a Gram-Charlier curve fitted to the weights of 300 Princeton freshmen (see pp. 143-145). It will be interesting to compare the area calculated here with that obtained from the tables of probabilities for a Gram-Charlier curve (see p. 147).

Equation (5) gives as a second approximation¹,

$$16.10 - \frac{291}{5,760} (12.70 - 31.47) + \frac{17}{5,760} (2.19 - 31.30) = 16.96$$

Since the original ordinates were one class interval times as large as they should be (if the area under the curve was to equal the total frequency), there is no need in this case to multiply the 16.88 or 16.96 by the class interval. Hence the final answers are 16.88 and 16.96 cases. If these are divided by 300 (the total number of cases), the results are .0563 and .0565, which are the first and second approximations to the probability (relative frequency) that

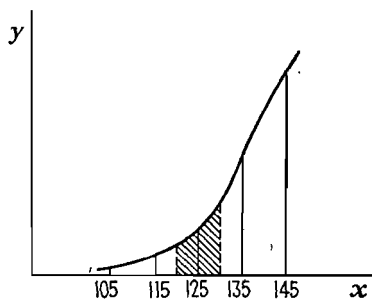


FIG. 39.—Five ordinates of the Gram-Charlier curve fitted to the weights of 300 Princeton freshmen.

a case falls in the interval 120–130.

It will be noted that the initial term in each of Eqs. (4) and (5) is the ordinate of the curve at the mid-point of the interval for which the area is to be calculated. This means that the area of the rectangle whose height is the ordinate of the curve at the mid-point of the interval and whose base is the class interval i may be taken as a first approximation to the area under the curve over that interval. That is, it is assumed that the curve can be roughly approximated for the given interval by the horizontal straight line $y = y_0$. The extra terms added in Eq. (4) imply that the curve can be better approximated by a second-degree parabola drawn through the ordinates of the curve at $x = 1$, $x = 0$, and $x = -1$, that is, through the ordinates, y_1 , y_0 , and y_{-1} . Equation (5) assumes that a still better approximation to the curve can be obtained by a fourth-degree parabola drawn through the ordinates of the curve at $x = 2$, $x = 1$, $x = 0$, $x = -1$, and $x = -2$, that is, through y_2 , y_1 , y_0 , y_{-1} , and y_{-2} . Elderton also gives equations based upon these same degrees of approximation, that express the area over an interval in terms

¹ The area computed later from a table of probabilities for a Gram-Charlier curve is 16.98, which shows how good this approximation is.

of the ordinates of the curve at the ends instead of the mid-points of that interval and of neighboring class intervals.¹

FITTING FREQUENCY CURVES

To fit a frequency curve to a given set of sample data means to determine numerical values for the constants of the curve equation. This is usually done by relating the curve constants to the various statistics of the data. The problem will now be considered with reference to the more important types of curves that are fitted to sample data.

The Normal Curve. When the variable is expressed as an absolute deviation from the zero origin, the equation for the normal curve is

$$y = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(X - \bar{X})^2}{2\sigma^2} \right] \quad (6)$$

where $\bar{X}$ and σ are the mean and standard deviation of the curve. It would appear, therefore, that the normal curve could be fitted to a set of sample data by merely putting the mean of the data and the standard deviation of the data in the curve equation. This is essentially the method that is used. A particular adjustment must be made, however, whenever the sample standard deviation is computed from grouped data. This adjustment will now be discussed.

When data are grouped for the purpose of calculating a mean, standard deviation, or other statistic, a certain arbitrary distortion may result from considering all the cases of a given class interval as being all concentrated at the center of the interval or, what amounts to the same thing, as being uniformly distributed throughout the interval. If there were enough cases in an interval, it would probably be found that the frequencies of the cases in the interval would actually form a smooth curve that tended to rise toward the center of the distribution (see Fig. 40a) instead of forming a horizontal line as assumed by the grouping of the data (see Fig. 40b).

The error that results from grouping is generally negligible in the case of odd-powered moments such as the mean, since errors in measuring positive deviations tend to be offset by errors in measuring negative deviations. For even powered

¹ See ELDETON, W. P., *op. cit.*, pp. 25-26, 48.

moments, however, such as the variance, the errors of measurement are cumulative and result in a net error due to grouping. W. F. Sheppard¹ has put the error in the variance at $\frac{1}{12}i^2$ where i is the size of the class interval.

Before the normal curve is fitted to a sample histogram, therefore, the variance of the sample, when calculated from grouped data, must be corrected for grouping by subtracting $\frac{1}{12}i^2$. The square root of the corrected variance will give the corrected

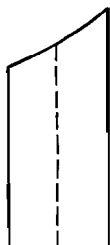


FIG. 40a.—Actual distribution of frequencies in a class interval.

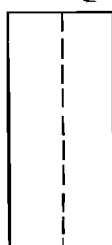


FIG. 40b.—Assumed distribution of frequencies in a class interval.

standard deviation. When the mean of the data and this corrected value for the standard deviation are put in the general formula for the normal curve, the result will be a normal curve that has the same mean and same standard deviation as the given data.²

Whether the curve actually fits the data depends on how truly normal the data are. The goodness of fit can usually be made evident by plotting the normal curve on the same chart as the histogram of the data.³ If the fit or lack of fit is somewhat doubtful, various methods may be employed to test the goodness of fit. The χ^2 test of goodness of fit is discussed below.⁴

Nonnormal Curves. The reason for fitting a normal curve to a given set of data is to determine whether the data are or are not normally distributed. If the distribution is normal, then the

¹ See *Proceedings of the London Mathematical Society*, Vol. 29 (1898) p. 353.

² In fitting a frequency curve the data are generally very numerous, so the multiplication of the sample variance by $\frac{N}{N-1}$ to improve the estimate of the population variance affects the result but little and can therefore be neglected.

³ See pp. 137-139.

⁴ See pp. 139-150.

normal table can be used to calculate probabilities and the results of sampling can be predicted by the carefully worked out sampling theory applicable to normal populations.

When data are distinguished as nonnormal, there may be further advantages in fitting a nonnormal curve to the data. Such a curve, for example, may serve to "smooth" the histogram and may thus permit a more accurate determination of the relative frequencies of the population from which the sample was taken. The identification of the distribution of given data with a particular frequency curve may also serve to distinguish them from other data the distribution of which is identified with a different frequency curve. The "fitting" of frequency curves may thus help to classify data as to type. These are the two principal reasons for fitting nonnormal frequency curves.

A Gram-Charlier Curve. The usual type of nonnormal curve fitted to data of everyday life is either a Pearsonian curve or a Gram-Charlier curve. Since the latter is the easier to fit, it will be discussed first.

In the case of the normal curve, the curve constants were themselves commonly computed statistics, *viz.*, the mean and standard deviation of the data (the latter corrected for errors of grouping). In the case of a Gram-Charlier curve, the general frequency equation is¹

$$y = \frac{1}{\sigma \sqrt{2\pi}} \left(1 + \frac{A}{\sigma^3} \left[3 \frac{(X - \bar{X})}{\sigma} - \frac{(X - \bar{X})^3}{\sigma^3} \right] + \frac{B}{\sigma^4} \left[3 - 6 \frac{(X - \bar{X})^2}{\sigma^2} + \frac{(X - \bar{X})^4}{\sigma^4} \right] \right) \exp \left[-\frac{(X - \bar{X})^2}{2\sigma^2} \right] \quad (7)$$

where

$$A = -\frac{\mu_3}{3!} \quad \text{and} \quad B = \frac{\mu_4 - 3\mu_2^2}{4!}$$

Hence the constants of the curve are again functions of commonly computed statistics, *viz.*, the mean $\bar{X}$, the standard deviation σ , the third moment μ_3 , and the fourth moment μ_4 . Therefore, a Gram-Charlier curve can be fitted to a set of sample data by substituting the sample values for the population values $\bar{X}$, σ , μ_3 , and μ_4 of the curve equation.

If the sample values of $\bar{X}$, σ , etc., have been computed from grouped data, as is usually the case, then a correction for group-

¹ See pp. 92-99.

ing must be made in the case of the even-power moments $\sigma^2 = \mu_2$, and μ_4 . Sheppard's corrections in these instances are

$$\left. \begin{aligned} \mu_2 &= \mu_2 \text{ (uncorrected)} - \frac{1}{12}(i)^2 \\ \mu_4 &= \mu_4 \text{ (uncorrected)} - \frac{(i)^2}{2} \mu_2 \text{ (uncorrected)} + \frac{7}{240}(i)^4 \end{aligned} \right\} \quad (8)$$

where the μ 's stand for the corrected moments and for the uncorrected moments.

The goodness of fit of the curve to the data may be examined by simply making a graphic comparison of the curve and sample histogram to which it is fitted, or a χ^2 test may be undertaken. These tests of goodness of fit are described and illustrated below.

A Pearsonian Curve. Karl Pearson, it will be recalled,¹ found that frequency curves in general might be represented by an equation of the type

$$\text{Relative slope} = \frac{x + a}{b_0 + b_1x + b_2x^2} \quad (9)$$

The logical basis upon which this system of curves rests was discussed in Chap. IV. It is the purpose here to describe the fitting of such a curve.

The first step is to relate the constants a , b_0 , b_1 , and b_2 of Eq. (9) to various statistics computed from the sample data. The method by which this is accomplished is explained by W. P. Elderton in his *Frequency Curves and Correlation* (pages 39 to 40). The essence of the procedure is to determine a , b_0 , b_1 , and b_2 so that the first four moments of the fitted curve will be the first four moments of the equation. The algebra is elaborate and will not be repeated here. When the results obtained by Elderton's analysis are substituted in Eq. (9), it becomes²

$$\text{Relative slope} = \frac{x + \frac{\sqrt{\mu_2} \sqrt{\beta_1} (\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}}{\frac{\mu_2(4\beta_2 - 3\beta_1) + \sqrt{\mu_2} \sqrt{\beta_1} (\beta_2 + 3)x + (2\beta_2 - 3\beta_1 - 6)x^2}{2(5\beta_2 - 6\beta_1 - 9)}} \quad (10)$$

¹ See Chap. IV (p. 57).

² Elderton's method involves the assumption that $x^n(b_0 + b_1x + b_2x^2)y$ vanishes at the ends of the range of the distribution, i.e., that there is close contact with the x -axis at both ends.

It will be noted that the mode of the curve is that point at which the relative slope is zero. Equation (10) thus shows that the mode of a Pearsonian curve comes at

where $x = X - \bar{X}$, $\beta_1 = u_3^2/u_2^3$ and $\beta_2 = u_4/u_2^2$, and $\sqrt{\beta_1}$ is to have the same sign as u_3 .

The Pearsonian equation is now expressed in terms of the moments and the β coefficients of the curve and the equation can be fitted by substituting the moments of the sample data for the curve moments. The values of the second and fourth moments must again be adjusted for Sheppard's corrections if they have originally been computed from grouped data [see Eqs. (8)].

The Pearsonian equation (9), it will be recalled,¹ yields different types of curves, depending on the value of the criterion $\kappa = b^2/4b_0b_2$. When the b 's are given their values from Eq. (10), this criterion becomes²

$$\kappa = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \quad (11)$$

The types of curves distinguished on the basis of this equation are as follows:³

Value of criterion	Curve type
$\kappa = 0, \beta_1 = 0, \beta_2 > 3$	VII
$\kappa = 0, \beta_1 = 0, \beta_2 = 3$	Normal curve
$\kappa = 0, \beta_1 = 0, \beta_2 < 3$	IIa
$\kappa = 0, \beta_1 = 0, \beta_2 < 1.8$	IIb
$0 < \kappa < 1$	IV
$\kappa = 1$	V
$1 < \kappa < \infty$	VI
$\kappa = \infty$, that is, $2\beta_2 - 3\beta_1 - 6 = 0$	III
$\kappa < 0$	I

$$x = \frac{-\sqrt{u_2}\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

Since $x = X - \bar{X}$ and $\sqrt{u_2} = \sigma$,

$$M_0 = \bar{X} - \frac{\sigma\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

where $\sqrt{\beta_1}$ is to be given the same sign as u_3 . Since skewness can be

measured by $\frac{\bar{X} - M_0}{\sigma}$,

$$sk = \frac{\pm\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

¹ See Chap. IV (p. 57).

² See ELDBERTON, *op. cit.*, p. 42.

³ Taken from *Tables for Statisticians and Biometricians*, Part I, p. lxiii.

coefficients in the criterion formula or by the location of the β values on Fig. 41. The equation for the curve (in differential form) will be obtained by substituting the corrected moments and β coefficients in Eq. (10), as was indicated above. Since this is a differential equation that gives the relative slope of a curve at any point and not the actual ordinate, it is necessary to derive an equation for the ordinates of a curve before the curve can be graphed. This problem will be discussed more fully in the next section on testing goodness of fit.¹

Fitting Sampling Curves. Sometimes laboratory experiments are undertaken to test certain sampling theories. These yield empirical sampling distributions. Sampling theory may suggest that these empirical distributions should conform to certain theoretical sampling distributions. To test this theory or possibly just to see how well the empirical distribution is approximated by some well-known theoretical distribution, a particular sampling curve is fitted to the empirical distribution. In such instances the process of fitting is relatively simple. To fit the t curve, χ^2 curve, or F curve it is necessary merely to determine the proper values for n or n_1 and n_2 and then plot the curve as described in the first section. The goodness of fit can be determined by graphic comparison or by a χ^2 test such as is described in the next section.

TESTING GOODNESS OF FIT

Testing a Normal Curve. Two methods of testing whether a normal curve fits a given set of sample data will be described here. These will be simple graphic comparison and the χ^2 test. Other methods of testing for normality are described in Chap. XVI.

Graphic Comparison. In the first section of this chapter the graphing of a standard normal curve was explained. This consisted merely in plotting the ordinates of the standard curve at selected values of x/σ . The problem now will be to adjust the abscissa scale and the curve ordinates so that the curve will fit a histogram with a given mean and a given standard deviation. The process is as follows:

First find what the mid-points of the histogram class intervals are in terms of standard deviation units measured from the mean

¹ See pp. 137-152.

as an origin. This can be done by subtracting the mean from each of the mid-values and dividing these deviations by the (corrected) standard deviation. For the values of x/σ so obtained, the ordinates of the standard normal curve may be found from Table VI in the Appendix. If the histogram with which the normal curve is to be compared is such that relative frequencies are measured by the areas of the rectangles erected on each interval, then the ordinates of the standard curve should be divided by σ to allow for the fact that the abscissa scale for the histogram is in absolute units while the abscissa scale for the standard curve is in standard deviation units.¹ If the histogram is such that absolute frequencies are measured by the heights of the rectangles erected on each interval, then the ordinates of the standard curve must be multiplied by Ni/σ , where i is the size of the class interval and N the number of cases.² When the

TABLE 17a.—CALCULATION OF THE ORDINATES OF THE NORMAL CURVE THAT FITS THE DISTRIBUTION OF HEIGHTS OF 300 PRINCETON FRESHMEN

(1)	(2)	(3)	(4)	(5)
X	$X - \bar{X} = x$	$\frac{X - \bar{X}}{\sigma}$	Ordinate of standard curve	Col. (4) $\times \frac{Ni}{\sigma}$
62.5	-7.97	-3.22	0.00224	0.27
63.5	-6.97	-2.82	0.00748	0.91
64.5	-5.97	-2.42	0.02134	2.59
65.5	-4.97	-2.01	0.05292	6.43
66.5	-3.97	-1.59	0.11270	13.69
67.5	-2.97	-1.19	0.19652	23.87
68.5	-1.97	-0.80	0.28969	35.19
69.5	-0.97	-0.39	0.36973	44.91
70.5	0.03	-0.01	0.39892	48.45
71.5	1.03	0.42	0.36526	44.36
72.5	2.03	0.82	0.28504	34.62
73.5	3.03	1.22	0.18954	23.02
74.5	4.03	1.63	0.10567	12.83
75.5	5.03	2.04	0.04980	6.05
76.5	6.03	2.44	0.02033	2.47
77.5	7.03	2.84	0.00707	0.86

$$\bar{X} = 70.47 \quad \sigma \text{ (corrected)} = 2.47$$

¹ The area of the histogram in this case is one regular unit while the area of the standard curve is one standard deviation unit. To make the two equal, the ordinates of the standard normal curve must all be divided by σ .

² The area of this histogram is $\Sigma Fi = Ni$. Hence, after the ordinates

standard ordinates have been properly adjusted, they can be plotted on the same graph as the given histogram and the goodness of fit may be examined visually. Fits that are obviously good and bad may be determined in this way.

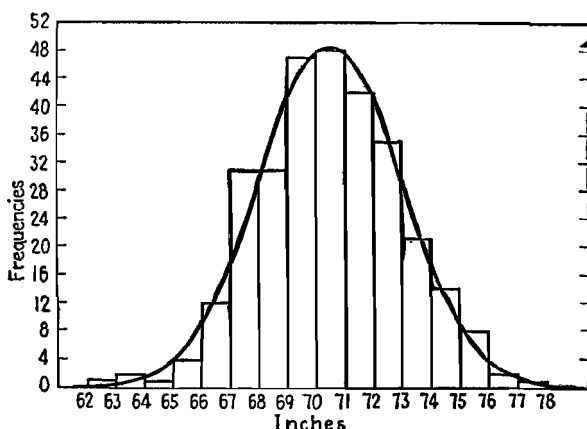


FIG. 42a.—Normal curve fitted to heights of 300 Princeton freshmen.

The whole procedure of graphically comparing a histogram with a fitted normal curve is illustrated by Table 17a, and the result obtained is shown in Fig. 42a. For this example, the fit is clearly a good one. It must be concluded, therefore, that the heights of young men of approximately the same age are normally distributed.

The χ^2 Test. When it is difficult to tell by graphic comparison whether a normal curve is a good fit to a given set of data, some more refined method must be undertaken. One such method is the χ^2 test, so called because it makes use of the χ^2 distribution in determining the goodness of fit. The essence of the test is a numerical comparison of the frequencies of the curve and histogram, interval by interval. The procedure is as follows:

To calculate the curve probabilities for each interval it is first necessary to express the limits of these intervals in terms of standard deviation units measured from the mean as an origin. Thus x/δ for each class limit can be found by subtracting the

have been adjusted for the difference in abscissa scales by division by σ , as explained above, they must be multiplied by Ni to take account of the fact that the area of this second form the histogram is Ni absolute units and not one unit.

mean from the class limit and dividing by the corrected standard deviation. The process is illustrated in columns (1) and (2) of Table 17a.

The next step is to find from the area table for the normal curve the probabilities, or relative frequencies, of cases lying between the mean and the various class limits. Then the relative frequencies, or probabilities, for each interval may be found by a process of subtraction, and these may be converted to absolute frequencies by multiplying by the number of cases N . This step is illustrated in columns (3), (4), and (5) of Table 17a.

If it should turn out that the curve frequency for any interval is less than 5, it should be combined with a neighboring interval or intervals so that the frequency of every interval is at least 5. Such adjustments are almost always necessary at the ends of the curve. In this connection it should be pointed out that the end intervals should always be taken as running to $\pm \infty$, so that the total frequency for the curve will be the same as that for the histogram.

The final step is to compute the quantity $\frac{(F - f)^2}{f}$ for each interval and then sum for all intervals. Here f stands for the curve frequency and F for the histogram frequency. The value of $\sum \frac{(F - f)^2}{f}$ is the final criterion of goodness of fit. As explained more fully below,¹ if normal curves are fitted to many sample histograms from a truly normal population, the various sample values of $\sum \frac{(F - f)^2}{f}$ will tend to form a sampling distribution that is of the form of a χ^2 distribution with n equal to the number of class intervals for which comparisons are made (a combined interval is treated as a single interval), minus 3. Hence, a selected point of the χ^2 table for the proper value of n , say the .05 point, will serve as a critical value for $\sum \frac{(F - f)^2}{f}$. For if the population from which the sample is taken is truly normal, there are only 5 chances out of a 100 that the value of $\sum \frac{(F - f)^2}{f}$ will equal or exceed the .05 χ^2 value. Hence, values

¹ See pp. 331-333.

TABLE 17b.—CALCULATION OF $\sum \frac{(F-f)^2}{f}$ FOR THE HEIGHTS OF 300 PRINCETON FRESHMEN

(1)	(2)	(3)	(4)*	(5)	(1')	(5')	(6)	(7)	(8)	(9)	(10)
X	$X - \bar{X}$	$\frac{X - \bar{X}}{\sigma}$	Area from $-\infty$	Area for each interval			Aggregate curve frequencies $(5') \times N = f$	Histogram frequencies $\frac{f}{F}$	$F - f$	$(F - f)^2$	$\frac{(F - f)^2}{f}$
-63	-7.47	-3.02	0.00130	0.00130							
-64	-6.47	-2.62	0.00440	0.00310							
-65	-5.47	-2.22	0.01321	0.00881							
-66	-4.47	-1.81	0.03515	0.02194	$-\infty$ to 66	0.03515	10.55	8	-2.55	6.5025	0.616
-67	-3.47	-1.41	0.07927	0.04412	66 to 67	0.04412	13.24	12	-1.24	1.5376	0.116
-68	-2.47	-1.00	0.15866	0.07939	67 to 68	0.07939	23.82	31	7.18	51.5524	2.164
-69	-1.47	-0.60	0.27425	0.11559	68 to 69	0.11559	34.68	31	-3.68	13.5424	0.391
-70	-0.47	-0.19	0.42465	0.15040	69 to 70	0.15040	45.12	47	1.88	3.5344	0.078
-71	0.53	0.22	0.58706	0.16241	70 to 71	0.16241	48.72	48	-0.72	0.5184	0.011
-72	1.53	0.62	0.73237	0.14531	71 to 72	0.14531	43.59	42	-1.59	2.5281	0.058
-73	2.53	1.02	0.84614	0.11377	72 to 73	0.11377	34.13	35	0.87	0.7569	0.022
-74	3.53	1.43	0.92364	0.07750	73 to 74	0.07750	23.25	21	-2.25	5.0625	0.218
-75	4.53	1.83	0.96638	0.04274	74 to 75	0.04274	12.82	14	1.18	1.3924	0.109
-76	5.53	2.24	0.98745	0.02107	75 to ∞	0.03362	10.08	11	0.92	0.8464	0.084
-77	6.53	2.64	0.99585	0.00840							
			1.00000	0.00415							
										$\sum \frac{(F-f)^2}{f} = 3.867$	

* The items in this column are obtained by subtracting from 0.50000 the figures found for each $-(X - \bar{X})/\sigma$ and by adding to 0.50000 the figures found for each $+(X - \bar{X})/\sigma$ in Table VI of the Appendix, p. 469.

of $\sum \frac{(F - f)^2}{f}$ greater than this .05 value suggest that the population is not truly normal; they are an index of bad fit.

An illustration of the procedure for calculating $\sum \frac{(F - f)^2}{f}$ is given in Table 17b. The value of $\sum \frac{(F - f)^2}{f}$ there obtained supports the graphic analysis in showing that the normal curve is a good fit to the heights of young college men. In fact, the fit is almost too good to be true.¹

A Gram-Charlier Curve. The goodness of fit of a Gram-Charlier curve may also be tested by graphic comparison and by a χ^2 test. These will now be described.

Graphic Comparison. The graphing of a Gram-Charlier curve is almost as easy as the graphing of a normal curve. Equation (7) shows that the ordinates of a Gram-Charlier curve are equal to the normal ordinates plus two different multiples of these ordinates, viz., $-\frac{\mu_3}{3!d^3} \left(\frac{3x}{d} - \frac{x^3}{d^3} \right)$ and $\frac{\mu_4 - 3\mu_2^2}{4!d^4} \left(3 - \frac{6x^2}{d^2} + \frac{x^4}{d^4} \right)$. Graphing of the Gram-Charlier ordinates is facilitated by use of tables of the normal ordinate [called $\varphi_0(x/d)$], the normal ordinate times $\left(\frac{3x}{d} - \frac{x^3}{d^3} \right)$ [called $\varphi_3(x/d)$], and the normal ordinate times $\left(3 - \frac{6x^2}{d^2} + \frac{x^4}{d^4} \right)$ [called $\varphi_4(x/d)$]. These values will be found in the Appendix, Table VI.

To get the ordinates of a particular Gram-Charlier curve it is necessary merely to multiply the values given for $\varphi_3(x/d)$ by $-\frac{\mu_3}{6\sigma^3}$ and the values given for $\varphi_4(x/d)$ by $\frac{\mu_4 - 3\mu_2^2}{24\sigma^4}$ and to add the algebraic sum of these (or to subtract if the sum is negative) to the values given for $\varphi_0(x/d)$.* When the ordinates so computed are multiplied by Ni/σ , they will become immediately comparable with the sample histogram from which the moments were computed and the two may be plotted on the same graph.

¹ Values of χ^2 less than the .95 point suggest that the randomness of the data be further investigated.

* If μ_3 and μ_4 have been calculated from grouped data, they must be adjusted for Sheppard's corrections before substituting in these equations (see p. 134).

To illustrate the fitting of a Gram-Charlier curve and the graphing of its ordinates, consider the distribution of weights of the 300 Princeton freshmen, shown in Fig. 42b. The mean of this distribution is 151.8 pounds, its standard deviation is 17.8 pounds, and the values of μ_2 , μ_3 , and μ_4 for this distribution are $\mu_2 = 318.094$, $\mu_3 = 3,566.48$, and $\mu_4 = 472,727$, Sheppard's cor-

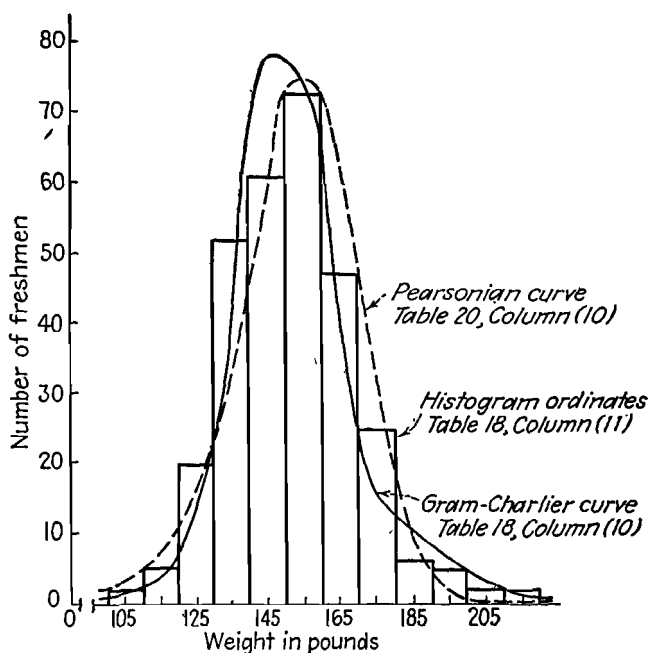


FIG. 42b.—A Gram-Charlier curve and a Pearsonian curve fitted to the distribution of weights of 300 Princeton freshmen.

rections for grouping being applied in all cases. The equation for the Gram-Charlier curve that fits this distribution is thus

$$y = \frac{1}{\sigma \sqrt{2\pi}} \left[1 - .1048 \left(\frac{3x}{\sigma} - \frac{x^3}{\sigma^3} \right) + .0697 \left(3 - \frac{6x^2}{\sigma^2} - \frac{x^4}{\sigma^4} \right) \right] \exp \left[-\frac{x^2}{2\sigma^2} \right]$$

or,

$$y = \left[\varphi_0 \left(\frac{x}{\sigma} \right) - .1048 \varphi_3 \left(\frac{x}{\sigma} \right) + .0697 \varphi_4 \left(\frac{x}{\sigma} \right) \right]$$

The steps taken to obtain the ordinates of this curve at the mid-points of various class intervals are indicated in Table 18. Thus in column (1) are written the values of the mid-points of the intervals, in column (2) their deviation from the mean of the distribution, and in column (3) the measure of these deviations in standard deviation units. In columns (4), (5), and (6) are entered the values of $\varphi_0(x/\sigma)$, $\varphi_2(x/\sigma)$, and $\varphi_4(x/\sigma)$ given in Table VI of the Appendix¹ for the values of x/σ listed in column (3). In column (7) is entered the product of $-\mu_3/6\sigma^3 (= -.1048)$ times column (5) and in column (8) the product of $\frac{\mu_4 - 3\mu_2^2}{24\sigma^4}$ ($= .0697$) times column (6). Column (9) contains the algebraic sums of the items of columns (4), (7), and (8), and in column (10) these sums are multiplied by $Ni/\sigma = 168.2$. These final figures are the ones in plotted Fig. 42b. Column (11) contains the ordinates of the histogram.

TABLE 18.—COMPUTATION OF THE ORDINATES OF THE GRAM-CHARLIER CURVE THAT FITS THE DISTRIBUTION OF WEIGHTS OF 300 PRINCETON FRESHMEN

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
X	$X - \bar{X} = z$	$\frac{z}{\sigma}$	$\varphi_0\left(\frac{z}{\sigma}\right)$	$\varphi_2\left(\frac{z}{\sigma}\right)$	$\varphi_4\left(\frac{z}{\sigma}\right)$	$\frac{-\mu_3}{6\sigma^3} \varphi_2$	$\frac{\mu_4 - 3\mu_2^2}{24\sigma^4} \varphi_4$	$\frac{1}{(4) + (7) + (8)}$	$(9) \times \frac{Ni}{\sigma}$ Curve ordinates	Histogram ordinates
105	-46.8	-2.62	.0129	+.1305	+.1152	-.0137	+.0080	+.0072	1.21	2
115	-36.8	-2.06	.0478	+.1225	-.2129	-.0128	-.0148	.0202	3.40	5
125	-26.8	-1.50	.1295	-.1457	-.7043	+.0153	-.0491	.0957	16.10	20
135	-16.8	-.94	.2565	-.5102	-.3901	+.0535	-.0272	.2828	47.57	52
145	-6.8	-.38	.3712	-.4028	+.7996	+.0422	+.0557	.4691	78.90	61
155	3.2	.18	.3924	+.2097	+1.1017	-.0220	+.0768	.4472	75.22	73
165	13.2	.74	.3034	+.5506	+.0043	-.0577	+.0003	.2460	41.38	47
175	23.2	1.30	.1714	+.2918	-.7341	-.0306	-.0512	.0896	15.09	25
185	33.2	1.86	.0707	-.0605	-.4095	+.0063	-.0285	.0485	8.16	6
195	43.2	2.42	.0214	-.1475	+.0461	+.0154	+.0032	.0400	6.73	5
205	53.2	2.99	.0046	-.0811	+.1337	+.0085	+.0093	.0224	3.77	2
215	63.2	3.54	.0008	-.0256	+.0643	+.0027	+.0045	.0080	1.35	2

$$-\frac{\mu_3}{6\sigma^3} = -.1048 \quad \frac{\mu_4 - 3\mu_2^2}{24\sigma^4} = .0697 \quad \frac{Ni}{\sigma} = 168.2$$

The χ^2 Test. The χ^2 test of goodness of fit of a Gram-Charlier curve is the same as the χ^2 test for a normal curve. Frequencies

¹ It is to be noted that for negative values of x/σ the signs of $\varphi_3(x/\sigma)$ entries in the table must be reversed.

of curve and histogram are compared interval by interval, and the value $\sum \frac{(F - f)^2}{f}$ is computed and compared with the .05 point of a χ^2 table. The procedure will be illustrated with reference to the distribution of weights discussed in the previous section.

Before the procedure itself is described, a brief explanation of how to compute relative frequencies or probabilities (areas) for a Gram-Charlier curve is in order. To calculate the area under a Gram-Charlier curve for a given range of x/σ values it is necessary first to compute the normal area and then adjust for the effects of the $\varphi_3(x/\sigma)$ and $\varphi_4(x/\sigma)$ terms. The adjustment required by the $\varphi_3(x/\sigma)$ term is given by the product of three factors, $\frac{x^2}{\sigma^2} - 1$, the ordinate of the normal curve for x/σ , and $-\varphi_3/6\sigma^3$. The adjustment required by the $\varphi_4(x/\sigma)$ term is simply $\varphi_3\left(\frac{x}{\sigma}\right)$ times $\frac{\varphi_4 - 3\varphi_2^2}{24\sigma^4}$. Thus to get the area under a Gram-Charlier curve from $-\infty$ to x/σ write down the area given by Table VI of the Appendix [this table gives the area between the mean 0 and x/σ ; to find the area from $-\infty$ to x/σ , subtract the table area from .5 if x/σ is negative, or add it to .5 if x/σ is positive], and add to it algebraically the value of $\left(\frac{x^2}{\sigma^2} - 1\right) \left[\varphi_0\left(\frac{x}{\sigma}\right) \right] \left(\frac{-\varphi_3}{6\sigma^3}\right)$ plus the value of $\left[\varphi_3\left(\frac{x}{\sigma}\right) \right] \left(\frac{\varphi_4 - 3\varphi_2^2}{24\sigma^4}\right)$. To calculate the area between any two values of x/σ , calculate the areas from $-\infty$ to each of these values, and take the difference.

Table 19 illustrates the procedure just outlined by showing how the areas under the Gram-Charlier curve fitted to the weights of the 300 Princeton freshmen are determined for various class intervals. Thus columns (1) to (3) convert the original class limits into x/σ deviations from the mean. Column (4) gives the areas under the normal curve from $-\infty$ to the upper limit of each class interval, now measured in x/σ units. Columns (5), (6), (7), and (8) show the computation of the adjustment required in each case by the $\varphi_3(x/\sigma)$ term [the final adjustment here is entered in column (8)], and columns (9) and (10) show the computation of the adjustment required by the $\varphi_4(x/\sigma)$ term [final adjustment given in column (10)]. Column (11) is the sum

of the items of columns (4), (8), and (10) and represents the final areas from $-\infty$ to the given values of x/σ . In column (12) the areas for the individual class intervals are computed by taking the successive differences between the items in column (11). These figures give the proportion of the total area contained in each interval. As indicated in the headings of the table, the sample values of μ_2 , μ_3 , and μ_4 are taken to be the parameter values, so the areas computed refer to the fitted curve.

The remaining part of Table 19 is concerned with testing the goodness of fit of the curve to the given histogram. Thus in column (13) the relative frequencies given in column (12) (areas under a frequency curve, it will be recalled, measure relative frequencies) are multiplied by $N = 300$ to give absolute frequencies that are comparable with the frequencies of the sample histogram. Then columns (14) to (17) carry out the calculations

necessary to compute the value of $\sum \frac{(F - f)^2}{f}$. This is found to be 11.6675. A χ^2 table shows that for $n = 10 - 5$ (i.e., the number of class intervals¹ minus 5) the probability of as great a value as 11.6675 is between .02 and .05, which does not indicate a very good fit. Apparently the Gram-Charlier type curve is not flexible enough in this instance to adjust itself to the sharp contours of the histogram. It is possible that a Pearsonian type curve might give better results; this will be examined in the next section.

A Pearsonian Curve. Graphic comparison and the χ^2 test can be used to test the goodness of fit of a Pearsonian curve as well as a normal or Gram-Charlier curve. The following section will thus be devoted to the problems peculiar to a Pearsonian curve and avoid as far as possible duplication of previous discussion.

Finding the Curve Equation. The Pearsonian equation (9), as noted above, is not an equation for the frequency curve itself, but one for its relative slope.² The former can be readily obtained from the latter, however, by the use of the integral calculus. Although it is possible in any practical case to substitute the computed values of the moments in Eq. (9) and then find the frequency curve to which it gives rise, it is generally

¹ See p. 333 for a fuller discussion of the basis for selecting n .

² See p. 57.

TABLE 19.—TEST OF GOODNESS OF FIT OF A GRAM-CHARLIER CURVE TO THE DISTRIBUTION OF WEIGHTS OF 300 PRINCETON FRESHMEN¹

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
	$X - \bar{X}$	$\frac{X - \bar{X}}{\sigma}$	Area under normal curve — ∞ to $\frac{x}{\sigma}$	$\frac{x^2}{\sigma^2} - 1$	$\exp\left(\frac{x^2}{\sigma^2}\right) \exp\left(\frac{x^2}{\sigma^2} - 1\right)$	$\frac{x^2}{\sigma^2} - 1$	$\frac{x^2}{\sigma^2} - 1$	$\exp\left(\frac{x^2}{\sigma^2}\right)$	$\frac{x^2}{\sigma^2} - 1$	Sum of (4), (8), and (10)	First differences of (11)	$f = \frac{N \times}{(12)}$	F	$F - f$	$(F - f)^2$	$\frac{(F - f)^2}{f}$
120	-31.8	-1.78	.0376	2.1634	.0819	.1776	-.0185	+.0242	+.0017	.0207	.0207	6.21	7	+.79	.6241	.1005
130	-21.8	-1.22	.1113	.4834	.1866	.0926	-.0097	-.3494	-.0243	.0773	.0566	16.98	20	+ 3.02	9.1204	.5371
140	-11.8	-.66	.2546	-.5644	.3209	-.1811	+.0190	-.5426	-.0378	.2358	.1585	47.55	52	+ 4.45	19.8025	.4165
150	-1.8	-.10	.4502	-.9900	.3970	-.3930	+.0412	-.1187	-.0083	.4931	.2573	77.19	61	-16.19	262.1161	3.3657
160	8.2	.46	.6772	-.7884	.3588	-.3829	+.0206	+.4599	+.0320	.7388	.2547	73.71	73	-.71	.5041	.0068
170	18.2	1.02	.8460	-.0404	.2372	.0096	-.0010	+.4735	+.0330	.8780	.1392	41.76	47	+ 5.24	27.4576	.6575
180	28.2	1.58	.9429	1.4964	.1145	.1713	-.0179	+.0812	+.0064	.9314	.0534	16.02	25	+ 8.98	80.6404	5.0337
190	38.2	2.14	.9838	3.5796	.0405	.1450	-.0152	-.1364	-.0095	.9591	.0277	8.31	6	- 2.31	5.3361	.6421
200	48.2	2.70	.9965	6.2900	.0104	.0654	-.0069	-.1207	-.0084	.9812	.0221	6.63	5	- 1.63	2.6569	.4007
			1.0000			0	0	0	0	1.0000	.0188	5.64	4	- 1.64	2.6896	.4769
																$\sum \frac{(F - f)^2}{f} = 11.6575$

¹ Several class intervals have been combined at the ends so that the curve frequencies in these intervals will not be too small (see p. 140 and fuller discussion on pp. 309, 326).

* Upper limits of class intervals.

easier to carry out the transformation (integration) first and then substitute the values of the moments in the resulting equation. The details of the process of integration and the development of curve equations for the various types of curves is explained in detail in W. P. Elderton, *Frequency Curves and Correlation*. Their use will again be illustrated by reference to the data on the weights of 300 Princeton freshmen. The μ 's and β 's for these data are $\mu_2 = 318.094$, $\mu_3 = 3,566.48$, $\mu_4 = 472,727$, $\beta_1 = .4063$, and $\beta_2 = 4.6720$. On substitution in Eq. (12), it is found that the criterion of curve type κ equals .161, which indicates a type IV curve. This is also indicated by Fig. 41.

If the integration is carried out, it will be found that the formula for a type IV curve is¹

$$y = y_0 \left(1 + \frac{x'^2}{a^2}\right)^{-m} e^{-v \tan^{-1} \frac{x'}{a}} \quad (12)$$

where

$$\begin{aligned} r &= \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6} \\ m &= \frac{1}{2}(r + 2) \\ v &= \frac{r(r-2) \sqrt{\beta_1}}{\sqrt{16(r-1) - \beta_1(r-2)^2}} \\ a &= \sqrt{\frac{\mu_2}{16}} \sqrt{16(r-1) - \beta_1(r-2)^2} \\ y_0 &= \frac{N}{aF(r,v)} \end{aligned}$$

$F(r,v)$ is a special function of r and v

and the origin is at the mean plus va/r . An alternative equation is²

$$y = y_0 \cos^{r+2} \theta e^{-v\theta} \quad (13)$$

where $\theta = \tan^{-1} \frac{x'}{a}$ and θ in $e^{-v\theta}$ is measured in radians. The function $F(r,v)$ may be evaluated for various values of r and v from Karl Pearson's *Tables for Statisticians and Biometricians*, Table LIV, explained on pages lxxxi to lxxxiii.

Equation (13) is the easier form to employ in plotting ordinates although it does not give ordinates at equal x intervals. If the

¹ See ELDERTON, *op. cit.*, p. 64.

² See *ibid.*, p. 65.

latter is essential, as it might be if the ordinates were to be used to compute areas, then Eq. (12) must be used. Equation (12) will be employed here. The calculations required for determining the curve equations are as follows:

Since $\beta_1 = .4063$ and $\beta_2 = 4.6720$, then, on substituting these sample values for the population parameters which they serve to estimate, it follows that

$$\begin{aligned} r &= \frac{6(4.6720 - .4063 - 1)}{2(4.6720) - 3(.4063) - 6} = 9.2204 \\ m &= \frac{1}{2}(9.2204 + 2) = 5.6102 \\ v &= \frac{9.2204(9.2204 - 2) \sqrt{.4063}}{\sqrt{16(9.2204 - 1) - .4063(9.2204 - 2)^2}} \\ &= 4.0398 \\ a &= \sqrt{\frac{318.094}{16}} \sqrt{16(9.2204 - 1) - .4063(9.2204 - 2)^2} \\ &= 46.8374 \end{aligned}$$

To find the value of $F(r, v)$ to be used in computing y_0 , resort must be had to Karl Pearson's *Tables for Statisticians*, Table LIV. The first step is to find φ , defined by the relationship $\tan \varphi = v/r$. For the given data,

$$\tan \varphi = \frac{4.0398}{9.2204} = .43814 \quad \text{and} \quad \varphi = 23.6605^\circ$$

For $\varphi = 23^\circ$ and $r = 9$, Pearson's *Tables* give $\log F(r, v) = 9.2186422 - 10$; for $\varphi = 24^\circ$, $r = 9$, $\log F(r, v) = 9.2488237 - 10$. Hence, for $\varphi = 23.6605^\circ$ and $r = 9$, the value of $\log F(r, v)$ is, by straight-line interpolation, $9.2385650 - 10$. Again, for $\varphi = 23^\circ$ and $r = 10$, $\log F(r, v) = 9.2347504 - 10$; for $\varphi = 24^\circ$ and $r = 10$, $\log F(r, v) = 9.2686087 - 10$. Interpolation gives, for $\varphi = 23.6605^\circ$ and $r = 10$, $\log F(r, v) = 9.2571003 - 10$. Finally, interpolation between $9.2385650 - 10$ and $9.2571003 - 10$ gives, for $\varphi = 23.6605^\circ$ and $r = 9.2209$, $\log F(r, v) = 9.2426502 - 10 = -.7573498$. Consequently,

$$\log y_0 = \log 300 - \log 46.8374 - (-.7573498) = 1.5638783,$$

and $y_0 = 36.63$. The formula for the curve is thus

$$y = 36.63 \left[1 + \frac{x'^2}{(46.84)^2} \right]^{-5.610} e^{-4.040 \tan^{-1} \frac{x'}{46.84}}$$

TABLE 20.—COMPUTATIONS FOR GRAPHING A PEARSONIAN TYPE IV CURVE

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
X	$x' = \frac{X - 172.3}{172.3}$	$\frac{x'}{46.8374}$	$\frac{1.0000 + (3)^2}{(3)^2}$	$\log (4)$	$\tan^{-1} (3),$ radians	$(6) \times 4.0398$ $\times .43294$	$(5) \times 5.6102$	$\frac{1.56388}{-(7) - (8)}$	antilog (9)	Histogram ordinates
105	-67.3	-1.4369	3.0647	.48639	-.96281	-1.68921	2.72875	.52434	3.3	2
115	-57.3	-1.2234	2.4967	.39737	-.88554	-1.55365	2.22933	.88820	7.7	5
125	-47.3	-1.0099	2.0199	.30533	-.79034	-1.38662	1.71296	1.23754	17.3	20
135	-37.3	-.7964	1.6343	.21333	-.67253	-1.17993	1.19682	1.54699	35.2	52
145	-27.3	-.5829	1.3398	.12704	-.52775	-.92592	.71272	1.77708	59.9	61
155	-17.3	-.3694	1.1365	.05557	-.35384	-.62080	.31176	1.87292	74.6	73
165	-7.3	-.1559	1.0243	.01043	-.15466	-.27134	.05851	1.77671	59.8	47
175	2.7	+.0576	1.0033	.00143	+.05754	+.10095	.00802	1.45491	28.5	25
185	12.7	+.2712	1.0735	.03080	+.26483	+.46463	.17279	.92646	8.4	6
195	22.7	+.4847	1.2349	.09163	+.45133	+.79184	.51406	.25798	1.8	5
205	32.7	+.6982	1.4875	.17246	+.60952	+.1.06938	.96754	1.52696	.3	2
215	42.7	+.9117	1.8312	.26274	+.73924	+.1.29697	1.47402	2.79289	.1	2

the origin being at $\frac{4.0398}{9.2204}$ (46.8374) plus the mean or at^{*}
 $20.5 + 151.8 = 172.3$.

TABLE 21.—COMPUTATION OF AREAS FROM ORDINATES OF TABLE 20

<i>X</i>	<i>y</i>	Δy	Area or frequency
105*	3.3	7.7 - 3.3 = 4.4	3.3
115	7.7	17.3 - 7.3 = 10.0	7.7 - $\frac{1}{2} \times (4.4 - 10.0) = 7.7 + .2 = 7.9$
125	17.3	35.2 - 17.3 = 17.9	17.3 - $\frac{1}{2} \times (10.0 - 17.9) = 17.3 + .3 = 17.6$
135	35.2	59.9 - 35.2 = 24.7	35.2 - $\frac{1}{2} \times (17.9 - 24.7) = 35.2 + .3 = 35.5$
145	59.9	74.6 - 59.9 = 14.7	59.9 - $\frac{1}{2} \times (24.7 - 14.7) = 59.9 - .4 = 59.5$
155	74.6	59.8 - 74.6 = -14.8	74.6 - $\frac{1}{2} \times (14.7 + 14.8) = 74.6 - 1.2 = 73.4$
165	59.8	28.5 - 59.8 = -31.3	59.8 - $\frac{1}{2} \times (-14.8 + 31.3) = 59.8 - .7 = 59.1$
175	28.5	8.4 - 28.5 = -20.1	28.5 - $\frac{1}{2} \times (-31.3 + 20.1) = 28.5 + .5 = 29.0$
185	8.4	1.8 - 8.4 = -6.6	8.4 - $\frac{1}{2} \times (-20.1 + 6.6) = 8.4 + .6 = 9.0$
195	1.8	.3 - 1.8 = -1.5	1.8 - $\frac{1}{2} \times (-6.6 + 1.5) = 1.8 + .2 = 2.0$
205	.3	.1 - .3 = -0.2	.3 - $\frac{1}{2} \times (-1.5 + .2) = .3 + .1 = 0.4$
215*	.1		0.1

* Areas of these intervals are taken equal to ordinates.

 TABLE 22.—THE χ^2 TEST OF GOODNESS OF FIT OF THE CURVE

<i>X</i>	<i>F</i>	<i>f</i>	<i>F</i> - <i>f</i>	(<i>F</i> - <i>f</i>) ²	$\frac{(F - f)^2}{f}$
Below 110.....	2	6.5*	-4.5	20.25	3.12
110-.....	5	7.9	-2.9	8.41	1.06
120-.....	20	17.6	2.4	5.76	.33
130-.....	52	35.5	16.5	272.25	7.67
140-.....	61	59.5	1.5	2.25	.38
150-.....	73	73.4	-.4	.16	
160-.....	47	59.1	-12.1	146.41	2.48
170-.....	25	29.0	-4.0	16.00	.55
180 and above†.....	15	11.5	3.5	12.25	1.06
Sum	300	300			16.65

* Taken to make total area equal to 300, it being assumed that the area above 220 is zero.

† Consolidated so that total area for interval of comparison is at least equal to 5.

Graphic Comparison. From this equation the ordinates of the frequency curve *y* may be computed as indicated in Table 20. It will be noted that the *x'* values are taken as the mid-points of the various class intervals. The final results are to be found in column (10); the ordinates of the histogram to which the curve is fitted are given in column (11) for the sake of comparison. The curve and histogram are graphed in Figure 42*b*.

The χ^2 Test of Goodness of Fit. To test the goodness of fit the areas under the curve for the various class intervals may be computed from the quadrature equation (4) or (5).^{*} This has been done in Table 21 by using Eq. (4). The final results are compared with the histogram frequencies in Table 22 and a χ^2 test of goodness of fit carried out. This table yields

$$\sum \frac{(F - f)^2}{f} = 16.14.$$

For $n = 9 - 5 = 4$,[†] the .01 point of a χ^2 distribution is 13.277, which indicates that for the case in question the probability of a value of $\sum \frac{(F - f)^2}{f}$ equal to or greater than 16.65 is less than .01. Hence the fit is not an especially good one. Apparently, owing to the sharp peak in the center, the given data cannot be well fitted by a smooth frequency curve.

^{*} See p. 128.

[†] For explanation, see p. 333.

PART II

Elementary Theory of Random Sampling

CHAPTER VIII

A PREVIEW OF SAMPLING THEORY

An important part of statistical theory is concerned with the logical basis of inferences about a population, *i.e.*, about a large set of cases, from which a sample has been taken. This is called the "theory of sampling."

In many instances a study of the whole population is impracticable, if not impossible. The set of children born to members of the white race goes back to the dim past and, so far as we know, will continue into the unknown future. For all practical purposes it is an infinite population of children. Any study of this population, for example, a study of the percentage of male and female births, must be made from a sample. Again, the registered voters in the United States form a population of many millions. If an institute of public opinion wishes to discover the trend of political sentiment in the country, it might conceivably approach every one of these millions of voters and ask him what party or candidate he favors. Such a "straw-vote," however, would necessitate considerable preliminary preparation and would be very costly, and the tabulation of returns would be an elaborate clerical and statistical problem; only the United States Bureau of the Census could venture to undertake such a job. A small private institute must necessarily determine public opinion from a carefully selected sample.

For similar reasons, sampling is employed in many other fields. It is used to study qualities of manufactured products, yields of various agricultural techniques, results of different medical treatments, effects of suggested educational methods, and the like. Sampling analysis is not confined to human populations but may be applied to the study of populations of inanimate

objects, plants, and animals; it has general applicability. Statisticians often speak of a "universe" instead of a "population"; the two words are interchangeable in sampling analysis.

RANDOM SAMPLING

Random Sampling and Probability. If a sample of data is obtained in a manner that may be characterized as "random," it is possible to make certain inferences about the population from which the sample has been drawn. With respect to a random sample the calculus of probability and the theory of frequency curves developed in previous chapters may be applied with a reasonable degree of accuracy. In random sampling, for example, it is possible to compute the probability of wrongfully rejecting a given hypothesis regarding the population. It is also possible to calculate ranges of values that may be stated to cover the actual population values with a given probability.¹ The randomness of a sample accordingly affords an essential basis upon which inferences regarding the population may be logically based.

Sampling that is purposive and not random will be discussed briefly at the end of this chapter. Samples obtained by this method may be "thought" to be good representations of the population, but just how good is indeterminate; nor can it be determined by the use of this method how often incorrect inferences regarding the population may occur. Random sampling is the only method so far devised that permits logical inferences about a given population.²

Types of Populations. Before discussing the technique of random sampling it is desirable to distinguish several different types of populations. A population may be either "existent" or "hypothetical." The registered voters in the United States, the stock of machine parts in a given storeroom, the wool sheep on a given ranch are all existent populations. The set of heads

¹ Testing of hypotheses and determination of "confidence intervals" are discussed briefly in a subsequent section of this chapter (see pp. 163-174) and developed more fully in the chapters that follow.

² It must be admitted, however, that confidence in an inference based on a random sample is dependent on the "thought" or firm belief that it is a truly random one. Whether thought with respect to randomness is any sounder, as a basis for inferences, than thought with respect to representativeness of a sample obtained by some other method is a debatable question.

and tails to be obtained by the indefinite tossing of a given coin, the children that have been and will be born to members of the white race, the results of the repeated performance of a given physical, biological, or other type of scientific experiment are all hypothetical populations. In some cases an existent population is part of a larger hypothetical population; for example, the machine parts in a storeroom, which of themselves form an existent population, may be viewed as a portion of the hypothetical population of parts that would be produced by the indefinite continuation of the given manufacturing process. Again, hypothetical populations may sometimes be the products of random selection from given existent populations. The balls in a bag, for example, form an existent population, but the balls that might be drawn from the bag with replacements would form a hypothetical population.

Populations may also be classed as "finite" or "infinite." The body of registered voters in the United States is a finite population, although it is so large that for some purposes it may be considered infinite. The set of heads and tails obtained by the endless tossing of a coin, the continuous births of white children, the results of an indefinite repetition of any physical or biological process—all constitute infinite populations. It is possible, of course, for a hypothetical population to be finite or infinite. The first 100 heads and tails to be obtained from the tossing of a coin is a finite hypothetical population. Existent populations, however, are almost universally finite.¹

Technique of Random Sampling. As indicated in the chapter on probability, randomness is largely a matter of intuition. Probability theory considers the set of all possible different samples and derives their distribution according to some criterion. If probability theory is to be used in predicting the results of any method of sampling, the method should be such that, if repeated a large number of times, it will tend to yield all possible different samples with equal frequency. Such a method is called a "random" method.

¹ The foregoing classification of populations follows closely that of G. Udny Yule and M. G. Kendall, *Introduction to the Theory of Statistics*, pp. 332-334. Also, see M. G. Kendall and B. Babington Smith, "Randomness and Random Sampling Numbers," *Journal of the Royal Statistical Society*, Vol. 101 (1938), pp. 147-166.

General notions regarding the concept of probability suggest that, if every member of the population is given an equal chance of being selected, or, to put it another way, if the selection of any member of the population is independent of the attribute it assumes, the results may be those desired. There is no definite assurance, however, that any special technique will conform to these criteria. All that can be done, and all that has been done, is to apply to a known population some apparently random method and to compare the results obtained with those expected upon the basis of probability theory. If the actual and theoretical results agree reasonably well when applied to the known population, it is concluded that the given method will also yield random samples when applied to an unknown population. The danger exists, nevertheless, that a method of selection yielding random samples in one instance may not give random samples in another instance. In the final analysis, belief that a particular method will produce random results rests on intuition guided by past experience. Some of the methods that have been devised to obtain random samples are ordinal selection, mechanical randomizing devices, tables of numbers, random sampling numbers, and natural selection.

Ordinal Selection. The methods of selection described in this and the next three sections are methods devised to obtain random samples from finite populations. They are frequently used in the social sciences, since the sampled populations in these fields are often *finite*.

If a given population consists of a list of objects in which the attribute of an object is independent of its position on the list, members of the population selected by some ordinal position (every tenth, every twentieth, or every tenth position per page) will be a random sample for the purpose of studying the given attribute. For example, a list of students alphabetically arranged in a college directory would presumably be an arrangement independent of student heights. Accordingly, some ordinal selection, say the tenth and twentieth name on each page, would give a random sample for the purpose of studying student heights.

The method of ordinal selection must be used with care. If a reporter of a college newspaper visits every tenth room in a given

dormitory and if the dormitory is so arranged that each entry has only 10 rooms, it might happen that he would visit only a first-floor room in each entry. If he were seeking data of an economic or social character, the preferred location of the rooms visited might give a definite bias to the sample obtained. This would be a case in which the method of selection failed to be independent of the attributes of the members of the population in which the reporter was interested.

In the use of ordinal selection, care must also be taken to note whether the list from which selection is made comprises the whole population being studied or is merely assumed to be representative of that population. The list of telephone subscribers in the city of New York might be taken, for example, to be representative of the whole adult population of the city. For some purposes, however, this would be a risky assumption. For very poor families cannot afford telephones, and any study of the social characteristics of the population would be biased unless it included representatives of this poorer section. The *Literary Digest* poll in 1936 failed accurately to predict the outcome of the presidential election because it relied heavily upon lists of names that were not representative of the whole population with respect to attributes affecting their votes.

Mechanical Randomizing Devices. Random samples are often sought by the use of various "randomizing" devices. Suppose, for example, that each member of a finite population is identified by a number. If this number is written on a small piece of paper and inserted in a small metal cylinder and if these cylinders are put in a revolving drum and thoroughly mixed, the selection of a sample of cylinders from the drum might possibly be taken to be a random sample from the given population. This type of randomizing device is used in many lotteries.

The shuffling of cards is another randomizing device. If the population is small enough, the numbers representing its various members may be written on cards and these may then be thoroughly mixed by shuffling. The withdrawal of a given number of cards from the shuffled pack may be taken as a random sample from the given population. Such a sample may be taken without replacing any of the cards; or the number of each card may be recorded, the card replaced, and the pack reshuffled before the

next card is drawn. Both these would constitute random samples, but the probability of a given type of sample would be different under the two procedures.¹

Similar devices make use of dice, roulette wheels, and other gambling instruments. Although a random sample may possibly be procured by such means, hidden bias in the mechanical device or in the procedure must be carefully avoided. If lottery cylinders are withdrawn by hand, the person making the withdrawals may have an unconscious predilection toward the cylinders of a certain shape, smoothness, or texture. If numbers are written on cards with ink, the difference in the size of the number might affect the stickiness of the cards and cause some to be withdrawn less often than others. The use of dice notoriously results in bias; Karl Pearson's study of the gaming results at Monte Carlo indicates that the odds against the absence of bias are extremely large.² If the possibility of deliberate falsification is precluded, the bias in the Monte Carlo results "would appear to arise from small imperfections in the roulette wheel which direct the ball into some compartments in preference to others."³ For these reasons, mechanical randomizing devices of this kind are no longer in high favor.

Tables of Numbers. When it is possible to associate a number with each member of the population, tables of numbers such as tables of squares and other powers, tables of logarithms, various statistical tables, and even tables of telephone numbers are sometimes employed in the attempt to secure a random sample. For example, if a population consists of 100 members and a number from 0 to 99 is associated with each member of the population, a statistical table involving seven-place figures, say, might be opened to any arbitrarily selected page and the last two digits of each of the first 10 lines of the table read off. The members of the population whose numbers correspond to the 10 pairs of two digits obtained in this way (09, 01, 07, etc., being read 9, 1, 7, etc.) might be considered to be a random sample of 10 from the given population.

¹ For further discussion of these two cases, see Chap. IX, pp. 186-190, 209-211.

² *Chances of Death*, Vol. I, Edward Arnold, London and New York (1897), pp. 42-62.

³ KENDALL and BABINGTON SMITH, *op. cit.*, p. 156.

The success of this method of obtaining a random sample depends on the digits in the table occurring in a random order. Unfortunately in many tables of numbers, such as tables of logarithms, a relationship exists between the figures in successive rows. Such tables, of course, cannot be used to draw a random sample. Even tables that are apparently random in character must be used with caution. Kendall and Babington Smith give the following account of an experiment with telephone numbers:¹ "We have attempted to construct a random series by selecting digits from the London Telephone Directory. In order to exclude bias as far as possible, pages were taken by opening the book haphazardly; numbers of less than four digits were ignored; numbers associated with names printed in heavy type were also ignored; and only the two right-hand digits were taken.

"It was found that a series of this kind was significantly biased. There appeared a deficiency of fives and nines. . . .

"The reasons for this effect are complicated and are not confined to the obvious one that telephone engineers would avoid fives and nines because of their assonance. It thus appears that the London Directory is useless as a source of random digits."

Random Sampling Numbers. Because ordinary tables of numbers may fail to yield random sets of digits, attempts have been made to construct special tables of random sampling numbers that may be used with some confidence. One such table is Tippet's *Random Sampling Numbers*. This consists of digits picked at random from British census reports. The digits are arranged in groups of four so as to provide 26 pages of 400 four-figured numbers, or a total of 10,400 four-figured numbers. These figures have been subjected to a number of different tests of randomness. Generally the results were deemed satisfactory, although one investigator found that the figures were somewhat "patchy" in meeting the tests.² The table and a description of the various methods of using it have been published by the Department of Applied Statistics, University of London, University College, as Tract XV of its *Tracts for Computers*.

Another set of 5,000 random sampling numbers has been published in the *Journal of the Royal Statistical Society* by

¹ *Ibid.*, pp. 156-157.

² See YULE, G. UDNY, "A Test of Tippet's Random Sampling Numbers," *Journal of the Royal Statistical Society*, Vol. 101 (1938), pp. 167-172.

M. G. Kendall and B. Babington Smith.¹ These were obtained from a special randomizing machine, which its inventors describe as follows:² "Essentially the machine consists of a disk divided into 10 equal sections, on which the digits 0 to 9 are inscribed. The disk rotates rapidly at a speed which can, if necessary, be made constant to a high degree of approximation by means of a tuning fork. The experiment is conducted in a dark room, and the disk is illuminated from time to time by an electric spark or by a flash of a neon lamp, which is of such short duration that the disk appears to be at rest. At each flash a number is chosen from the apparently stationary disk by means of a pointer fixed in space.

"In the actual experiment, the disk was rotated by an electric motor at about 250 revolutions per minute. It was illuminated by a neon lamp in parallel with a condenser in an independent electric circuit which was broken by means of a key. Owing to experimental conditions, the time between the making of the circuit and the passing of the flash varied, but to add an extra element of randomness the key was tapped irregularly by the experimenter. Flashes occurred, on the average, about once in 3 or 4 seconds."

These random sampling numbers of Kendall and Babington Smith have also been subjected to various tests of randomness with satisfactory results.

To draw a random sample, tables of random sampling numbers are used according to the procedure already explained in connection with other tables of numbers. After the members of the population are assigned numbers, a sample of size N is picked by selecting any set of N numbers from the table of random sampling numbers. Suppose, for example, that a sample of 100 from a population of 10,000 commercial banks doing business in a given area is desired. The banks are, it happens, listed in a directory. The first bank in the directory can thus conveniently be numbered 0 and the last 9999; and then Tippet's tables can be opened to any page and the numbers in the first column read off. The banks whose numbers correspond to the numbers so selected will then constitute a random sample of 100 from the whole population of 10,000 banks.

¹ *Op. cit.*, pp. 164-166.

² *Ibid.*, pp. 157.

Instead of taking the numbers from the columns of the table, they could have been selected by rows or even by diagonals. In each case, a random sample will be obtained. On the assumption that Tippett's numbers are a random set, the selection of any subset of these numbers in a way that is independent of the numbers themselves will yield a random sample. Kendall's and Babington Smith's numbers can be used in the same way.

Natural Selection. The methods of sampling described above are of use primarily in sampling from finite existent populations. For the most part they cannot be used to obtain random samples from an unknown hypothetical infinite population.¹ A population that would be hypothetically generated by the continuous operation of some physical, biological, or social process is an unknown hypothetical infinite population. When these processes occur in everyday life, any existent results of that process may be tentatively taken as a random sample of the infinite population of results. The selection here is a "natural" one in that it is effected by the ordinary forces of everyday life.

The role of the investigator in the case of hypothetical infinite populations is to study the circumstances under which the natural selection of the sample was effected in order to see whether some unusual influences might have given a special bias to this particular sample and consequently might have destroyed its naturally random character. Thus any set of white birth statistics might tentatively be taken as a random sample of the population of white births. If, however, the sex ratio were being investigated and it were known that geographical location had a significant effect on the ratio of male to female births, then data on white births in a given area could be looked upon as a random sample only of white births in that area and not of births in all areas.

Naturally occurring data thus provide their own method of selection. This is true whether the process giving rise to them is one of everyday life or a process carried out in some experimental laboratory or research station. In testing the effects of a certain serum, care may be taken to see that the guinea pigs

¹ When the population is known, however, and sampling is undertaken primarily for experimental purposes, random sampling numbers may be used to obtain random samples from infinite populations. See YULE, G. UDNY, and M. G. KENDALL, *op. cit.*, pp. 343-344.

used for the experiment are not noticeably abnormal, but otherwise the peculiar characteristics of each pig are those that chance happens to provide. Again, in agricultural experiments with fertilizer, the peculiar influences that affect each plot are those that are provided by the chance variations in soil, wind, rain, sunshine, and other natural forces.

In experimental research, however, care must be taken to design an experiment so that the balance of natural forces is not all to the one side or to the other for certain parts of the data. Thus, if the difference in rate of growth of self-fertilized and cross-fertilized plants is being studied, and if the cross-fertilized plants are all placed on the sunny side of the experimental plot, the effect of the sun and any possible effect of the difference in fertilization will be so confounded that it will be difficult, if not impossible to determine whether any significant difference in the rate of growth is due to the one or to the other. If an equal number of plants of both types are assigned at random to the two sides of the plot, however, then the constant effects of the sun will be eliminated. With this arrangement or design of the experiment, the possibility that the recorded difference in rate of growth may be due to chance instead of to a difference in method of fertilization may be determined by probability theory.¹ It is the role of the experimenter in these cases so to design his experiment that natural forces do provide him with a random sample.

In the case of very large finite populations, natural selection is also sometimes used to get a random sample. An organization investigating public opinion, for example, may send an agent to a given locality where the first dozen people he meets, of the sort whose opinion is desired, may be taken as a random sample of the given class of people. The sample is thus picked for him by chance forces. For the 1940 United States census, schedules were printed so that those persons whose names happened to fall on certain lines of each schedule were asked supplementary questions. Two out of the forty lines on each side of a schedule were marked off for this purpose so that a 5 per cent sample was obtained on the supplementary questions.²

¹ Cf. FISHER, R. A., *The Design of Experiments*, Chap. III.

² See STEPHAN, F. F., W. E. DEMING, and M. H. HANSEN, "The Sampling Procedure of the 1940 Population Census," *Journal of the American Statistical Association*, Vol. 35 (1940), pp. 615-630.

SAMPLING AND THE THEORY OF STATISTICAL INFERENCE

If a sample is drawn from an unknown population in a random manner, certain inferences about the population may be made on the basis of probability theory. The following sections will discuss in a general way the testing of particular hypotheses regarding a population and the estimation of population parameters. The application of this general theory to special problems will be discussed in detail in subsequent chapters.

Testing Hypotheses Regarding the Population. Often a problem presents itself in such a way that a definite hypothesis regarding the population is offered for testing. A fruit merchant about to buy a carload of oranges wishes them to be of good quality; he does not want to buy the carload, let us say, if more than 10 per cent of the oranges are substandard. Accordingly, he will want to test the hypothesis that the carload is 10 per cent substandard. If this hypothesis is rejected, because a random sample suggests that the number of substandard oranges is more than 10 per cent, he will not buy the carload. But if the hypothesis is not rejected, because the random sample suggests that the number of substandard oranges is not greater than 10 per cent, he will buy the oranges.

A similar problem is illustrated by the manufacturer of electric-light bulbs who does not wish to adopt a proposed new process of production unless the mean length of life of the bulbs producible by it is 1,000 kilowatt-hours or more. He knows from past experience that electric-light bulbs vary in length of life in accordance with the normal frequency curve. He will thus want to test the hypothesis that the population of bulbs producible by the new process is a normal population with a mean of 1,000 kilowatt-hours. If this hypothesis is rejected, because a random sample suggests that the mean length of life is less than 1,000 kilowatt-hours, the manufacturer will not adopt the new process. But if the hypothesis is not rejected, because the random sample suggests that the mean length of life is not less than 1,000 kilowatt-hours, he will adopt the new process. These are examples in which the problem itself suggests a particular hypothesis to be tested and alternatives to it.

Errors in Testing Hypotheses. In testing a particular hypothesis regarding a population, two kinds of errors may be made. The first of these errors, hereafter referred to as "error I," is the

rejection of the hypothesis when it is actually true. That is, the hypothesis will be deemed unreasonable on the basis of the sample obtained, although in fact the population is exactly as the hypothesis assumes. The second kind of error, which will be referred to as "error II," is the failure to reject the hypothesis, *i.e.*, the failure to consider it unreasonable on the basis of the given sample, when in fact the population is not the same as the hypothesis assumes. The aim of the statistical testing of hypotheses is twofold: it seeks on the one hand to limit the risk of the first kind of error to a preassigned amount and, on the other hand, to minimize the risk of the second kind of error.

Procedure for Limiting Risk of Error I. To limit the risk of falsely rejecting a given hypothesis to a preassigned amount, the following procedure is adopted:

First, the hypothesis to be tested is assumed to be true.

Second, a certain statistic is selected such as a mean or standard deviation, and probability theory is employed to show how this statistic might be expected to vary from sample to sample on the assumption that the hypothesis is true. The process consists in supposing that a large number of samples of given size have been drawn at random from the assumed population and then finding the frequency distribution of the sample values of the selected statistic. This distribution of sample values is called the "sampling distribution" of the statistic, and the standard deviation of this distribution is called the "standard error" of the statistic. The argument here is purely theoretical and is based upon the probability calculus.

The third step is to specify the degree of risk that the investigator is willing to take in rejecting the given hypothesis when it is true. This degree of risk is called the "coefficient of risk."

The fourth step is to study the sampling distribution of the selected statistic and to mark off a range of values for which the probability of the statistic falling within it is just equal to the coefficient of risk. This range is called the "region of rejection;" and the remaining range of values is called the "region of acceptance." To illustrate, suppose that the adopted coefficient of risk is .05. Furthermore, suppose that the statistic used to test the given hypothesis is the mean of the sample and that the probability calculus indicates that means of random samples from the assumed population will be distributed in the form of a

normal frequency distribution, the mean of which is 1,000 and the standard deviation of which is 20.¹ From the properties of a normal distribution, it is known that .05 of the frequencies of a normal distribution lie below the mean minus 1.645σ . Hence the range of values below $1,000 - 1.645(20) = 967.1$ could be taken as a .05 region of rejection. The range of values from 967.1 upward would constitute the corresponding region of acceptance (see Fig. 43).

The final step is to note whether the value of the selected statistic for the given sample falls in the region of rejection or the

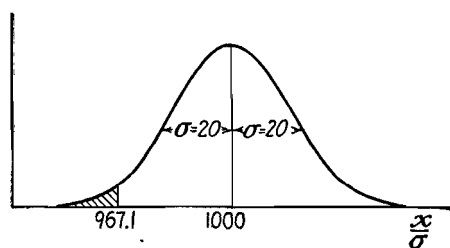


Fig. 43.—An .05 limit in the lower tail of a normal sampling distribution.

region of acceptance. For example, if the sampling distribution and the regions of rejection and acceptance were as indicated above, and if the mean of the given sample were 965.2, the given hypothesis would be rejected because this sample value would fall in the range below 967.1. If, on the other hand, the sample mean had been 972.3, then the hypothesis would have been accepted, for the sample value would then fall in the range above 967.1.

This procedure ensures that the risk of rejecting a given hypothesis when it is true will be just .05. For if it is always followed in testing hypotheses, false rejections will be made only 5 per cent of the time, since sample values will fall in regions of rejection only that often when hypotheses are true.

It will be noted that there are three arbitrary elements in the procedure. One is the selection of the statistic to be used, the

¹ As indicated in Chap. VI, the standard deviation of the sampling distribution of the mean, *i.e.*, the standard error of the mean, is equal to the standard deviation of the population divided by the square root of the size of the sample. For samples of 25, say, a standard error of 20 implies a population standard deviation of 100.

second is the selection of the coefficient of risk, and the third is the selection of the regions of rejection and acceptance. The choice of the statistic will depend in part on the ease of calculation and in part on the size and position of the regions of rejection and acceptance to which it gives rise. The latter consideration is of prime importance since, as pointed out below, the size and position of the regions of rejection are a major factor in reducing the error of accepting a hypothesis when it is not true. This will be more fully discussed later.

The selection of the coefficient of risk will depend in most instances upon the nature of the problem. If action based upon nonacceptance of a hypothesis is not of great significance, the investigator may be willing to run considerable risk of falsely rejecting a hypothesis. In testing a certain manufacturing process, for example, it may be that any tendency for the process to deteriorate may be corrected with relatively little expense. The manufacturer may in such instances be willing to undergo an occasional overhauling of his process when in fact there is no real need for such overhauling. In other cases, however, the expense of overhauling may be very great, and the manufacturer may be willing to undergo an unnecessary overhauling only very infrequently and may gladly undertake considerable expense for statistical testing to avoid any such unnecessary overhauling.

In the former case, the manufacturer might be willing to adopt a coefficient of risk of .10 (a risk that 10 per cent of the overhaulings will be unnecessary); in the latter, even a coefficient of risk of .01 (a risk that 1 per cent of the overhauling will be unnecessary) might be too great. In cases where serious or dangerous medical treatment is involved, a still smaller coefficient of risk might be adopted. A coefficient of .05 is one that has come to be frequently used in cases where there is no strong reason for adopting a very high or a very low figure.

The third arbitrary element, the choice of the regions of rejection and acceptance, is illustrated in Fig. 44. For any given statistic and coefficient of risk it is possible to pick innumerable regions of rejection and acceptance. In the previous illustration, the range of values below 967.1 was taken as the region of rejection, and the range of values from 967.1 upward was taken as the region of acceptance. The range of values above $1,000 + 1.645\sigma$, that is, above 1,032.9, could equally well have been used as a

.05 region of rejection and the range of values from 1,032.9 downward as the corresponding region of acceptance, for the probability of a sample mean falling above the mean, plus 1.645σ , is likewise .05. This probability of .05 could also have been

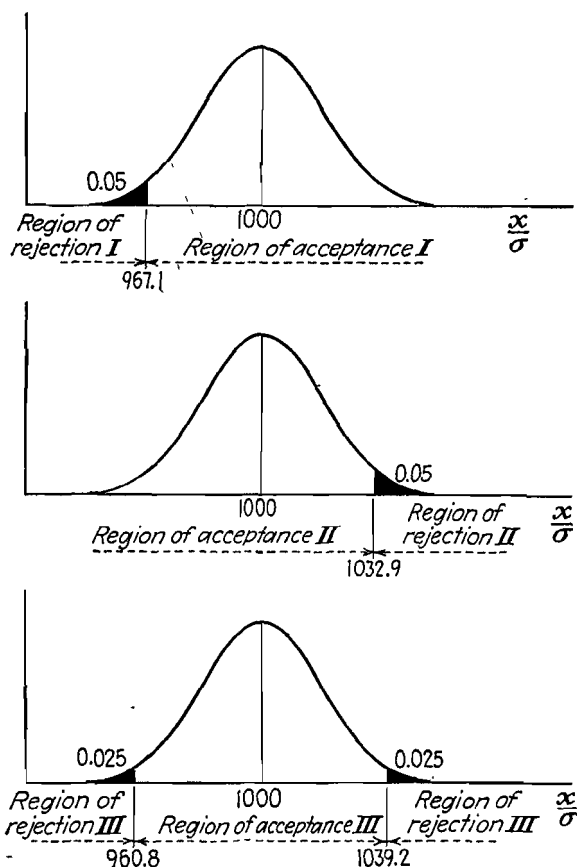


FIG. 44.—Three combinations of regions of rejection and acceptance that will make the risk of falsely rejecting the hypothesis $\bar{X} = 1,000$ just equal to .05.

obtained by splitting the region of rejection so that it included the range of values below $1,000 - 1.96\sigma$, that is, below 960.8, and the range of values above $1,000 + 1.96\sigma$, that is, above 1,039.2. The corresponding region of acceptance in this case would have been the range of values from 960.8 to 1,039.2, as shown in the

third section of Fig. 44.¹ Still other regions of rejection could have been found that would have had a probability of .05.

Any of these regions would have limited the risk of falsely rejecting the given hypothesis to the preassigned coefficient of risk, *viz.*, .05. They would not all have been equally good with respect to the risk of the second kind of error, *viz.*, that of accepting the given hypothesis when it is not true. It is with the problem of selecting regions of rejection and acceptance that will minimize this second kind of error that the ensuing discussion is concerned.

Procedure for Minimizing Risk of Error II. If the population is not as specified by a given hypothesis, the probability of a sample falling in any region of acceptance selected for limiting the risk of error I will depend on the position of the region in relation to the actual character of the population. If there were one particular region of acceptance for which the probability of a sample falling within it was less than that of any other region giving the same risk of error I, no matter how the population differed from the given hypothesis, this would obviously be the best region that could be selected. For, in this case, the probability of accepting the false hypothesis would be less than it would be for any other region giving the same risk of error I.

Such a best critical region is usually impossible to determine. Ordinarily, one region will be the best that can be chosen if the population differs from the given hypothesis in one direction, while another region will be the best if the population differs from it in another direction. If the statistician is concerned with the possibility of the population differing from the given hypothesis in whatever direction possible, some compromise region is generally employed.

The selection of good regions of rejection and acceptance will be illustrated by reference again to the three diagrams of Fig. 44. Suppose that the hypothesis to be tested is that the mean of the population from which a sample has been drawn equals 1,000, and suppose that probability theory, together with certain facts known about the population, indicates that the sampling distribution of means of samples from this assumed population will be a normal distribution with a mean of 1,000 and a standard

¹ This is the only type of region that was discussed in J. G. Smith and A. J. Duncan, *Elementary Statistics and Applications*, Chap. XII.

deviation of 20. Finally, suppose that the three regions of rejection and acceptance shown in Fig. 44 are considered for adoption. The question is which of these regions will be the best to employ. The answer is that it depends on the nature of the problem; this can be illustrated by concrete cases.

Suppose that the data of the given example pertain to the length of life of electric-light bulbs, measured in kilowatt-hours, producible by some new process. The manufacturer of these bulbs, it can be argued, will desire especially to avoid acceptance of the hypothesis that the average length of life of the new bulbs is 1,000 kilowatt-hours when in fact it is less than that. For if he accepts the given hypothesis when actually the mean length of life is greater than 1,000 kilowatt-hours, he stands to lose nothing; he in fact gets a better process than he expected. But if he accepts the hypothesis of a mean length of life of 1,000 kilowatt-hours when actually the mean is less than that, he gets a poorer process than he expected. The quality of his product will not reach the desired standard, and the result may be a considerable loss of money. In this instance, therefore, the manufacturer will want to adopt regions of rejection and acceptance that will minimize the risk of accepting the hypothesis when actually the mean length of life of the bulbs is less than 1,000. This set of regions of acceptance and rejection is shown in the first diagram of Fig. 44.

By the selection of this set of regions of rejection and acceptance, the risk of error II is minimized; this is illustrated by three diagrams in Fig. 45. The three diagrams in this figure show the effects of applying the three sets of regions of rejection and acceptance shown in Fig. 44. The three curves in Fig. 45 show the true probabilities of obtaining sample means of varying sizes, forming a normal distribution about the true population mean of 980, superimposed over the scale showing the regions of rejection and acceptance according to the hypothesis of a mean of 1,000, shown in Fig. 44. The diagrams in Fig. 44, from which the regions of acceptance were obtained, show the probabilities of obtaining sample means of varying sizes, forming a normal distribution with a mean of 1,000 and a standard deviation of 20.

Figure 45 suggests that the risk of accepting the hypothesis of 1,000 when the actual mean is less than 1,000 is less for the

region of acceptance running from 967.1 upward than for any other region that involves the same risk of rejecting the given hypothesis when it is true (*i.e.*, for any other .05 region). In each of the three diagrams the proportion of the area under the curve crosshatched indicates the probability of accepting the

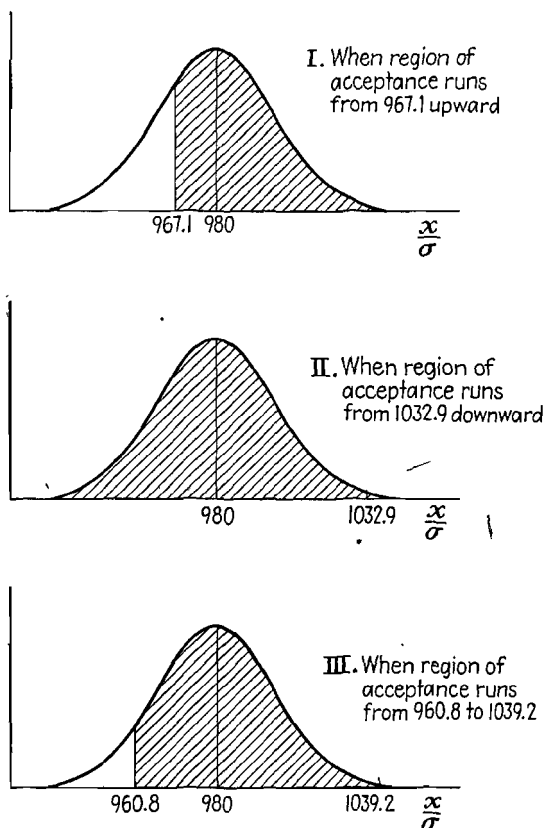


FIG. 45.—Risk of accepting hypothesis that population mean is 1,000 when actually it equals 980.

hypothesis that the mean is 1,000, when in fact it is 980, that is, the probability of accepting a hypothesis when it is not true. Part I of the figure shows the probability of a sample mean falling above 967.1, part II shows the probability of its falling below 1,032.9, and part III the probability of its falling between 960.8 and 1,039.2. These probabilities are the risks of accepting the hypothesis of a mean of 1,000 when the true mean is 980

and when the region of acceptance for testing the given hypothesis runs from (I) 967.1 upward, (II) 1,032.9 downward, and (III) from 960.8 to 1,039.2.

It is clear that this risk is least for the first of these three regions, as shown by the fact that the crosshatched portion of the area under the curve is less in the first diagram of Fig. 45. Further study of this kind shows that, whenever the mean of the population is less than 1,000, region I will always give a lower risk of accepting the hypothesis of 1,000 than any other region

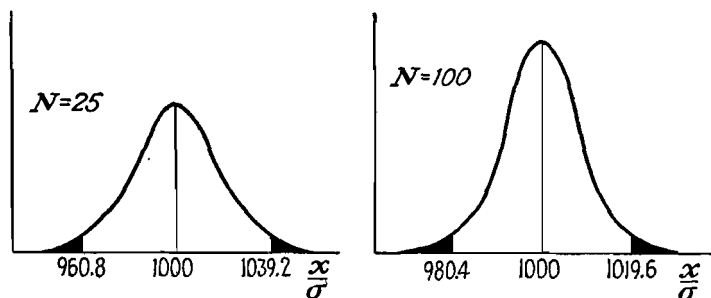


FIG. 46.—Sampling distribution of the mean of a sample when the mean of the population is 1,000 and the standard deviation of the population is 100. Region of rejection, .025 points at each end, when $N = 5$ compared to when $N = 100$.

of acceptance for which the risk of falsely rejecting the hypothesis is .05. For the problem illustrated, therefore, it is the best region of acceptance that may be employed; since the manufacturer wants to minimize the risk of accepting the hypothesis that the mean is 1,000 when in fact it is less.¹ In other instances, one of the other regions might be preferred.

Effect of the Size of the Sample. In testing hypotheses the size of the sample is of prime importance; for, the larger the sample, the narrower the sampling distribution, *i.e.*, the smaller the standard error, of the statistic used to test a hypothesis. The standard error of the mean, for example, is equal to the standard deviation of the population, divided by the square root of the size of the sample. Hence the standard error of the mean varies inversely with the square root of the size of the sample.²

¹ Of course, if the risk of rejecting the hypothesis were increased, the risk of falsely accepting the hypothesis could be reduced.

² See Chaps. XIII, XIV.

From this inverse relationship between the spread of the sampling distribution and the size of the sample it follows that, the larger the sample, the closer the finite limits of the region of acceptance to the central tendency of the sampling distribution of that statistic, assuming, of course, a constant coefficient of risk. This increases the likelihood of rejecting a hypothesis that is not true. In other words, the larger the sample, the closer the value of a sample statistic must be to its expected average value if a given hypothesis is to be accepted. This is illustrated by Figs. 46 and 47, which contrast the regions of rejection and acceptance with samples of 25 and 100, respectively.

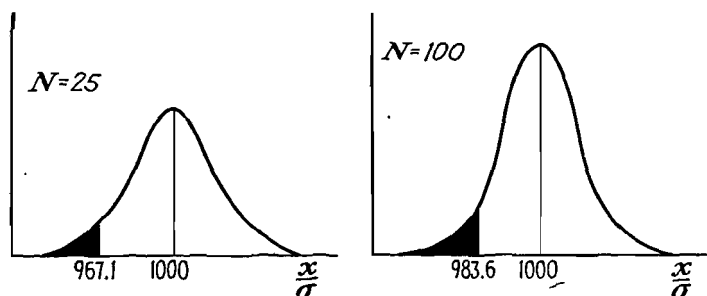


FIG. 47.—Sampling distribution of the mean of a sample when the mean of the population is 1,000 and the standard deviation of the population is 100. Region of acceptance, .05 point at the lower end, when $N = 5$ compared to when $N = 100$.

Still more important is the effect of the relationship between the size of the sample and the standard error of the sample statistic upon the risk of accepting a hypothesis that is not true. Since the sampling distribution becomes narrower as the size of the sample increases, the larger the sample, the less the probability that a given hypothesis will be accepted when it is not true. This is illustrated in Figs. 48a and 48b. It is principally for this reason that greater confidence is put in a test making use of a large sample than in a test employing a small sample. In Figs. 48a and 48b, the probabilities of samples occurring in the regions of acceptance are represented by the crosshatched portions of the sampling distributions about the true mean in each case.

Effect of the Statistic Selected. As pointed out above, the procedure for testing a given hypothesis may be varied by taking different sample statistics. A hypothesis regarding the mean

of a normal population, for example, may be tested by using either the mean or the median of the sample. The former has the advantage, however, of having a sampling distribution that

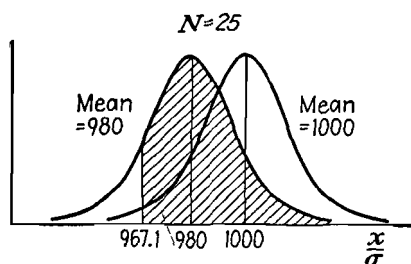


FIG. 48a.—Standard deviation of the population known to be 100; if the hypothesis is that the mean of the population is 1,000, the range of values downward from 967.1 will be a .05 region of rejection for testing this hypothesis with reference to a sample of 25. The range of values upward from 967.1 will be the corresponding region of acceptance. This figure shows the probability of a sample falling in this region of acceptance when the true mean is 980 and not 1,000.

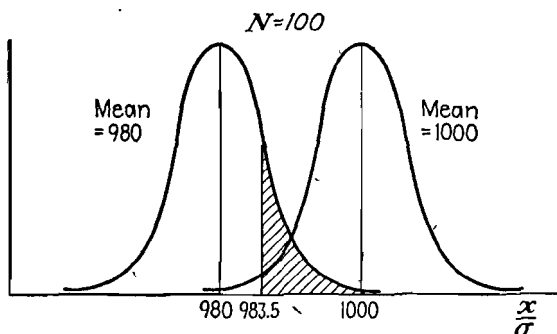


FIG. 48b.—Standard deviation of the population known to be 100; if the hypothesis is that the mean of the population is 1,000, the range of values downward from 983.5 will be a .05 region of rejection for testing this hypothesis with reference to a sample of 100. The range of values upward from 983.5 will be the corresponding region of acceptance. This figure shows the probability of a sample falling in this region of acceptance when the true mean is 980 and not 1,000.

is narrower than that of the median. In other words, the standard error of the mean is less than the standard error of the median.¹ The use of the mean in preference to the median, therefore, is the equivalent of employing a larger sample; or, to put it another way, as good a test can be made with the mean

¹ This result may be reversed for other than normal universes.

of a smaller sample as with the median of a larger sample. The former is accordingly the better statistic to employ for the test.

Estimation of Population Parameters. Often the statistician is interested in more than the testing of a single hypothesis. In some cases, no particular hypothesis is suggested by the problem. Even if it is, the statistician may wish to know, not only whether a particular hypothesis is acceptable on the basis of the sample return, but also what hypotheses in general are acceptable and what are not. More exactly, he may wish to draw a boundary line for which it can be said that the chances are, say, 95 out of 100, that this boundary includes the true population. In addition, he may want to select a single hypothesis as the best hypothesis to be adopted. The theory of estimation seeks an answer to these questions.

Specification of a Population. A population is precisely specified when its functional form is given together with the values of its parameters. A normal population, for example, is precisely specified when the fact of its normality is given, together with the values of its mean and standard deviation. In many cases, however, a population will be reasonably well determined if its more important parameters are given, such as its mean, standard deviation, β_1 , and β_2 . For in this case a Pearsonian curve or a Gram-Charlier curve can be derived from which the probabilities of the population can be approximately computed. Thus, if the form of a population is known, estimates of its parameters will exactly specify it; if the form is not known, estimates of its parameters will at least approximately determine the population. Hence, estimates of a population become largely estimates of its parameters.¹

In the ensuing discussion it is assumed that the form of the population is known a priori. This will make for greater simplicity and precision in the analysis and will facilitate the presentation of the argument. It will not affect any of the general conclusions of this chapter.

Confidence Intervals. A principal consideration in the theory of estimation is to determine confidence intervals for a population parameter. A confidence interval for a given parameter is

¹ In many instances, knowledge of a particular parameter or set of parameters is all that is desired, so that full knowledge of the population is not required.

a range of values that, on the basis of a given sample, has a specified probability of including the true value. The probability associated with any confidence interval is called the "confidence coefficient" for the interval.

The procedure by which such a confidence interval for a given parameter may be obtained will now be traced step by step.

1. Assume at the start that the form of the population is known a priori. For simplicity, also assume that the population

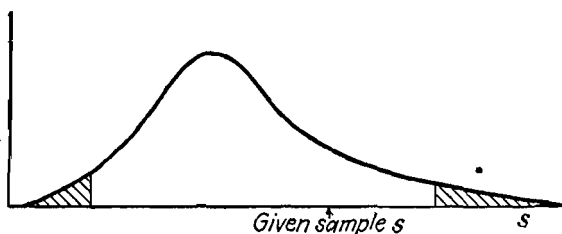


FIG. 49.—A given sample in a region of acceptance.

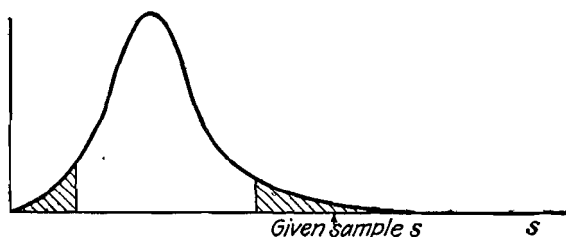


FIG. 50.—A given sample in a region of rejection.

is characterized by the value of a single parameter, which may be designated as θ .

2. Let some sample statistic s be chosen for the purpose of estimating the value of θ . What statistic is chosen is immaterial to the present argument.

3. From knowledge of the population, derive by use of the probability calculus the sampling distribution of s for samples of the given size. This will give the relative frequencies with which sample s 's may be expected to have different values, since the sampling distribution of s is derived from the original population, the precise nature of the distribution will depend on the value of the parameter θ . The way in which a sampling distribution might vary with variations in a parameter θ is suggested in Figs. 49 to 52.

4. Let the confidence coefficient for the given interval be set at .95. This means that the interval to be determined should have a probability of .95 of covering the true value of θ .

5. Calculate the value of s for the sample actually obtained. Then determine the value of the parameter θ that will cause the upper .025 point of the sampling distribution of s (as noted in item 3, this distribution varies with the value assigned to θ) to coincide with the computed sample value of s . Such a value of

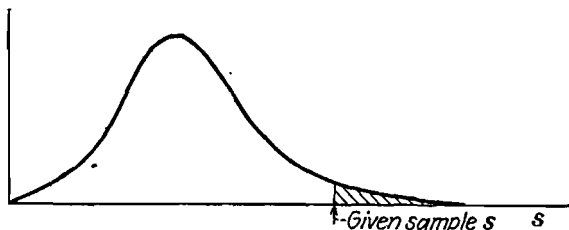


FIG. 51.—Lower confidence limit.

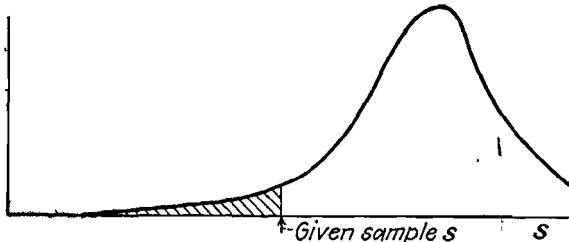


FIG. 52.—Upper confidence limit.

θ is pictured in Fig. 51. Call this value of θ " θ lower," and designate it by θ_l . Next select a value of θ that will cause the lower .025 point of the sampling distribution of θ 's to coincide with the computed sample value of s . A value of θ that will give this result is pictured in Fig. 52. Call this " θ upper," and designate it θ_u . The interval $\theta_l - \theta_u$ is the desired confidence interval. For if the foregoing procedure is always followed in setting up confidence intervals, the true value of θ will fail to be covered only in those instances in which the sample value of s falls in the upper or lower .025 tail of the true sampling distribution of s ; and this will occur on the average only 5 times out of 100. Hence, on the average, the confidence interval $\theta_l - \theta_u$ will cover the true value of θ 95 times out of 100. The values θ_l and θ_u are called the " θ lower" and " θ upper confidence limits" of θ .

In conclusion, it may be noted that there are three arbitrary elements in the procedure for the determination of confidence intervals, just as there were three arbitrary elements in the procedure for testing a hypothesis. One is the choice of the sample statistic s to be used; the second is the selection of the confidence coefficient; and the third is the selection of the points of the sampling distribution of s that are to coincide with the sample value.

Since the sampling distributions of some statistics are narrower than others, for the same confidence coefficient a smaller confidence interval can be obtained in some cases than in others.¹ Some statistics thus provide more accurate estimates of population values than others.

Confidence intervals can be made smaller,² if the confidence coefficient is smaller. Thus, in a given instance one may be able to say that the chances are 80 out of 100 that the interval 50-70 includes the true value; or he might be able to say that the chances are 95 out of 100 that the interval 30-90 includes the true value. If the investigator is willing to run greater chances of being wrong, he may thus reduce the size of the interval that is said to include the true value. In matters involving life and death a very high confidence coefficient should be adopted. In testing to discover the effect of an advertising campaign, on the other hand, a much smaller coefficient might be used.

The third arbitrary element in the determination of confidence intervals is the selection of the points of the sampling distribution of s to serve as the points of coincidence with the sample value of s . In the analysis above, the upper and lower .025 points were chosen. For a confidence coefficient of .95 any other set of points that included 95 per cent of the sample values of s could have been used, such, for example, as the lower .01 point and the upper .04 point, or the lower .03 point and the upper .02 point. If an upper or lower limit only is desired for a confidence interval, the lower or upper .05 point of the distribution of s can be used. The particular points that are chosen in any case will depend on the problem in hand. Normally in estimating θ

¹ Or if, as noted on pp. 177-178, a confidence interval has only one finite bound, this finite bound will be closer to the sample value.

² Or the finite bound of an interval that has one infinite bound can be brought closer to the sample value.

both an upper and lower bound will be desired. In some instances, however, the statistician may not care if he overestimates a certain figure, but he may wish especially not to underestimate it. In other instances, he may wish particularly to avoid overestimation.

The whole process of determining confidence intervals may be illustrated by reference to the previously used electric-light bulb problem. Suppose that a sample batch of 25 bulbs has a mean length of life of 965 kilowatt-hours, and suppose that the manufacturer desires to determine limits for the true mean length of life that have a probability of .95 of covering this true value. To simplify the analysis, suppose that the standard deviation in length of life of the bulbs is known to be 100 kilowatt-hours and that only the mean is unknown.¹

To determine confidence limits for the mean of the population given a sample mean of 965 kilowatt-hours, the procedure is as follows: First identify the mean of the population as the unknown parameter θ , and let the statistic s , which will be used to estimate θ , be the mean of the sample. Next note that the sampling distribution of the means of random samples from a normal population can be shown by the probability calculus to be a normal distribution with a mean equal to the mean of the population and a standard deviation equal to the standard deviation of the population divided by the square root of N , the size of the sample. In the problem in hand, therefore, the standard deviation of the sampling distribution of s will have a standard deviation of $100/\sqrt{25} = 20$ and a mean of which the value will be unknown but which will equal θ .

As the third step in the analysis, note that the value of the mean of the population that will cause the upper .025 point of the sampling distribution of s to coincide with the given sample value of s is 925.8. For the upper .025 point of the sampling distribution of the mean lies at just 1.96 times σ above the mean of the population, *i.e.*, at 1.96 times $20 = 39.2$ above the mean of the population. If this is to coincide with the given sample mean of 965, then the population mean must have the value $965 - 39.2 = 925.8$. The value 925.8 will thus be taken as the lower confidence limit for the mean of the population.

¹ Knowledge of the standard deviation might be obtained from technical considerations.

Similarly, note that the value of the mean of the population that will cause the lower .025 point of the sampling distribution of sample means to coincide with 965 is $965 + 39.2 = 1,004.2$. The value 1,004.2 may thus be taken as the upper confidence limit for the mean of the population. The whole confidence interval will thus be $925.8 - 1,004.2$, and it may be said that the chances are 95 out of 100 that this interval includes the true population mean.

If the statistician wished to make an especially conservative report to the manufacturer about the new process, he might choose a confidence interval that has only an upper bound (but still a confidence coefficient of .95). This upper bound would be determined by finding the value of the mean of the population that caused the given sample value to fall at the lower .05 point of the sampling distribution of the mean. For the problem illustrated, this upper bound would be given by $965 + 1.645$ times $20 = 997.9$ (for the .05 point of a normal distribution comes at 1.645 times σ from the mean). The statistician would in this instance report to the manufacturer that there is a chance of 95 out of 100 that the range of values up to 997.9 includes the true value, or, to put it another way, that there is only a chance of 5 out of 100 that the range of values above 997.9 includes the true value.

Effect of Size of Sample. For confidence intervals making use of the same statistic, the same confidence coefficient, and the same type of interval, the larger the sample the smaller the confidence interval, or, in the case of an interval having only an upper or lower bound, the closer the bound to the sample value. That is, a larger sample is always a better means of estimating the character of a population than a smaller sample. Of course, large samples may be more costly to procure than small ones, and their greater accuracy may not be worth the additional expense. In all cases, however, improvement in estimates to be obtained from larger samples should be given consideration, for the gain from increased size may far overbalance the higher cost.

Maximum-likelihood Estimates of Population Parameters. Sometimes a problem requires something further than setting up a range of values that probably includes the true value of a population parameter. It may be desirable to have a single figure that can be considered the "best," or "optimum," estimate

that can be made of the population parameter from a given sample.

It should be noted at the start that various methods may be employed to estimate population parameters, and it is doubtful if any single method is the best for all circumstances. There is one method; however, the method of maximum likelihood, that has been received with considerable favor, and it is this method that will be described here.

The method of maximum likelihood reasons entirely from the sample to the population. It consists in estimating the values of the population parameters so that the probability of the given sample among all possible samples of the same size is a maximum. It will be noted that, if the values of the population parameters are given, a particular set of sample values will have a definite probability. As the parameter values are changed, the probability of the given sample values also changes. According to the method of maximum likelihood, the "best estimates" that can be made of unknown population parameters, given a particular sample, are those values of the parameters which, if they were the true values, would make the probability of the given sample a maximum.

Mathematically the method of maximum likelihood may be described as follows: Suppose that for given values of the population parameters $\theta, \theta', \theta'', \dots$, the probability of a sample consisting of $X_1, X_2, \dots, X_n$ is given by the probability distribution $F(X_1, X_2, \dots, X_n, \theta, \theta', \theta'', \dots)$. Then for given values of $X_1, X_2, \dots, X_n$, that is, for a given sample, the best estimates that can be made of $\theta, \theta', \theta'', \dots$ are those values that maximize the logarithms of the function F . The logarithm of F , instead of F itself, is maximized because the mathematical analysis is easier. The results are the same, however, for F is a maximum when $\log F$ is a maximum. If the form of F is known, these maximizing values can be found by the use of the differential calculus. Thus the optimum estimates of θ, θ' , and $\theta'', \dots$ must satisfy the conditions

$$\frac{\partial \log F}{\partial \theta} = 0 \qquad \frac{\partial \log F}{\partial \theta'} = 0 \qquad \frac{\partial \log F}{\partial \theta''} = 0$$

These give as many equations as there are parameters, and their common solution affords the desired estimates.

It will be noted that the values of the population parameters that are found by the method of maximum likelihood are values that depend on the given sample values. In other words, the estimate of each population parameter is expressed as a certain "function" of the sample values. These functions of the sample values constitute "statistics," for that is what a statistic is, a function of the observed sample values. Usually it is found that the maximum-likelihood estimate of a given population parameter is the sample counterpart of the parameter itself. If the population is normal, for example, the mean and standard deviation of the sample are the maximum-likelihood estimates, respectively, of the mean and standard deviation of the population.

Sometimes it is desired to estimate a population parameter in a way that will be independent of the estimates of other parameters. Such an independent estimate of a population parameter may be made under either of two conditions. If it turns out that in making the maximum-likelihood estimates one of the partial derivatives gives an equation containing only a single parameter, then this parameter may be estimated immediately without troubling with the estimates of the other parameters.

The second condition permitting an independent estimate of a population parameter may be described as follows: Sometimes it happens that the function F giving the probability of a given sample can be broken up into the product of two factors, one of which depends on the sample values and on a single population parameter. If this is true, the value of that parameter can be estimated independently of the other parameters by choosing the value that maximizes this part of the total probability. For example, it so happens that the probability of a sample from a normal population can be factored into a part that depends only on the sample values and on the standard deviation of the population. The other part depends on the sample values and on both the mean and the standard deviation of the population. By taking as an estimate of σ the value that maximizes the first factor, there will be obtained a maximum-likelihood estimate of the population standard deviation that is independent of the population mean. When the mathematics of this is actually carried out,¹ it is found that this independent estimate of the

¹ See pp. 290-294.

population standard deviation is

$$\sigma = \sqrt{\frac{(X_i - \bar{X})^2}{N - 1}}$$

In other words, the maximum-likelihood estimate of the standard deviation of a normal population that is independent of the population mean is given by the standard deviation of the sample multiplied by $\sqrt{\frac{N}{N-1}}$. When the mean and standard deviation are estimated jointly, the maximum-likelihood estimate of the standard deviation of the population is the standard deviation of the sample. The factor $\sqrt{\frac{N}{N-1}}$ that occurs in the independent estimate may thus be viewed as a correction factor that makes allowance for neglect of the estimate of the population mean.

Maximum-likelihood Statistics as "Optimum" Statistics. Statistics derived by the method of maximum likelihood are deemed to be "optimum statistics" in certain senses.¹ First these statistics are said to be "consistent." By this it is meant that as the size of the sample is increased the sample statistic approaches closer and closer to the population parameter it is used to estimate. A more precise statement is that the sampling distribution of the given statistic becomes more and more concentrated around the value of the population parameter and the probability of any given finite deviation from the population value becomes less and less. Such an approach of the sample value to the population value as the size of the sample is increased is spoken of as "stochastic convergence." A maximum likelihood estimate is thus a consistent statistic in that it approaches the population value stochastically.

Maximum-likelihood estimates are also optimum statistics in that they are "efficient." By this it is meant that in the limit as the size of the sample is indefinitely increased there is no other statistic that has a smaller sampling variance. For a finite sample some other statistic used to estimate the same

¹ See FISHER, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London, Series A*, Vol. 222 (1922), pp. 309-368.

population parameter may have a somewhat smaller sampling variance, but as the size of the sample is increased the sampling variance of the maximum-likelihood statistic will tend to become at least as small as that of the other statistic. Maximum-likelihood statistics are thus said to have sampling variances that are a minimum asymptotically.

A third optimum characteristic of maximum-likelihood estimates is their "sufficiency." This characteristic ensures that, when a maximum-likelihood estimate has been made of a population parameter, no further information about that parameter can be obtained from any statistic that is independent of (*i.e.*, not functionally related to) the maximum-likelihood statistic. Again this characteristic is obtained only in the limit as the size of the sample is indefinitely increased.

Maximum-likelihood statistics are thus optimum statistics in three ways. They are consistent, efficient, and sufficient.

STRATIFIED, OR REPRESENTATIVE, RANDOM SAMPLING

Often sampling fluctuations can be greatly reduced by use of partial or supplementary knowledge about a population instead of relying entirely upon the sampling process itself. Suppose, for example, that it is known that with respect to the division of public opinion on a given question religious preference was an importance influencing factor. And suppose further that census data are available giving the number of people in the given community belonging to each religious category. If reliance was placed upon random sampling only to give a representative sample of public opinion, then it would be hoped that the sample was representative regarding the proportion of various religious adherents in each category as well as of the division of public opinion within each. When knowledge of the relative number of people in each religious category is available, however, then there is no need to rely upon sampling for it. In this case, the sampling can be so devised that the number of people sampled in each group is proportional to the number of people in each group in the whole population. The randomness of the sampling will in this instance be restricted to the selection of the individuals within each religious category. This is known as "stratified" or "representative" random sampling. It is especially impor-

tant in public-opinion analysis, index-number construction,¹ and the like.

The significance of stratified, or representative, random sampling is that it reduces sampling errors. It makes use of knowledge of correlation between the variable which is being studied and one or more other variables which are correlated with this variable and about which information is available. By using this correlation it diminishes the extent of the chance fluctuations.²

PURPOSIVE SAMPLING

Some sampling methods do not employ the random technique but select their samples to conform to chosen criteria. This is spoken of as "purposive sampling." For example, suppose a given research bureau is in search of some particular knowledge regarding the urban population of the United States. To get this they may go over carefully all the cities in the country and select a city that is "typical," say, as to size, as to proportion of heavy and light industries in the city, as to percentage of foreign born, etc. This city then will be taken as a typical sample of American cities, and the data it reveals on the problem in question will be considered as typical of American urban population.

Purposive sampling is also used in combining census data in various ways, in order to save the time and money that would be required by use of the complete data. In the Danish census of 1923, for example, it was decided to pick a representative sample that was to be about 20 per cent of the whole country and also 20 per cent of the total in each of the country's 22 counties.³ The procedure for securing this sample was as follows:

For each of the 1,300 parishes in the country, the number of cows per 100 hectares of farm area was computed and the parishes in each county were grouped according to the magnitude of this ratio, five groups being distinguished in each county. From each of these groups the parish was selected whose agricultural

¹ See SMITH and DUNCAN, *op. cit.*, Chap. XIX.

² It will be recalled that the variation around a line of regression, *i.e.*, the scatter, or second-order variance, is always less than the total variation in the dependent variable.

³ See JENSEN, ADOLPH, "The Representative Method in Practice," *Bulletin de l'Institut international de statistique*, tome 22 (1926), 1^{ère} livraison, pp. 420-421.

area most nearly amounted to one-fifth the total agricultural area of the group. In some cases, two or more parishes were combined or a single parish divided in order to get this 20 per cent sample. Next the parishes so selected were drawn on a map to see if the various parts of the country were about equally well represented. When this was not the case, certain parishes with approximately the same number of cows per hectare as parishes included in the sample were substituted for the latter so as to improve the geographical distribution. Finally this preliminary sample was tested with regard to special factors such as the number of farms, total agricultural area, grain area, number of milch cows, number of pigs, etc., to see whether the sample represented with respect to these factors approximately one-fifth the whole country. Where there were marked divergencies in this respect, adjustments were made to make the sample approximately 20 per cent of the whole for these auxiliary items without at the same time disturbing the area relationship or the geographical representation.

Purposive sampling of this kind is hard to appraise. The argument is that if a sample is representative in certain respects it will be representative in other respects, but this need not be a necessary consequence. Samples of this kind depend largely on the judgment of those making them and are worth just about as much.¹

¹ Cf. remarks on p. 154.

CHAPTER IX

SAMPLING FROM A DISCRETE TWOFOLD POPULATION

Sampling from a discrete population in which all the elements fall into one or the other of two categories is one of the simplest sampling problems. This type of problem is illustrated when public opinion is being investigated, when a manufacturer is testing the quality of a given production process, or when a sociologist is determining the ratio of male to female births.

Sampling from a discrete twofold population affords an opportunity to study in a more detailed yet relatively simple way most of the aspects of sampling theory sketched in the previous chapter. The following discussion consequently offers an easy approach to the general principles of sampling, and the reader is urged to master it completely. The arguments, for the most part, will follow the essential steps that were outlined in the preceding chapter. Chapter XIII generalizes the argument for a manifold population.

THE PROBLEM AND INITIAL ASSUMPTION

To avoid excessive abstraction the argument will be expounded with reference to a concrete example. Suppose that a manufacturer has produced an order of machine parts numbering several hundred thousand pieces. According to the standards of the trade an order of this kind is not acceptable if more than 10 per cent of the parts are defective. Routine inspection of each part is costly, however, and the only feasible method of testing the lot is to take a sample. The problem then arises as to what inferences can be made about the percentage of defectives in the whole lot from the determination of the percentage of defectives in the sample.

If the sample yields 11 per cent defectives, is it a reasonable hypothesis that the percentage of defectives in the whole lot is only 10 per cent? What are the limits within which the percentage of defectives in the whole lot may be reasonably considered to lie? What is the best estimate that can be made of

the percentage of defectives in the whole lot? It is with these questions that the following analysis is concerned.

Assumptions. In selecting the size of the sample to be tested the manufacturer is pulled in opposing directions. The larger the sample tested, the greater the expense to the manufacturer. He will seek, therefore, to test as small a sample as is practicable. On the other hand, in choosing a small sample he increases the risk of having the whole lot rejected when in fact it may be above standard.¹

In the ensuing argument it will be assumed that the sample selected is small relative to the size of the population, although it may be large absolutely. More precisely, it will be assumed that the withdrawal of the sample parts does not materially change the percentage of defective and nondefective parts in the whole lot. Suppose, for example, that the initial lot consists of 500,000 parts and that a sample of 2,500 is taken. If 10 per cent of the whole lot were defective and 40 per cent of the sample were defective (an extremely unlikely result), the percentage of defectives in the whole lot after the entire sample had been taken would still be 9.85, which differs but little from 10 per cent. In what follows it will be assumed that, as the sample is drawn, the percentage of defectives in the whole lot remains unchanged. This will greatly simplify the analysis.²

It will further be assumed that the process of sampling is a random one, as described in the previous chapter. The full implication of this will be understood only as the analysis proceeds. For the moment it may be noted that with randomness the results of repeated sampling should be predictable by mathematical probabilities calculated from some appropriate mathematical model.

DISTRIBUTION OF SAMPLE PERCENTAGES

In the problem in hand the population of parts is divided into groups, defective parts and nondefective parts; it is therefore a twofold population. The problem is to make inferences regarding the percentage of defective parts in the population from a sample set of data. The only sample statistic that is appropriate

¹ See pp. 171-172.

² For a discussion of the complications that arise when this assumption cannot be made, see pp. 209-211.

for making inferences about the population percentage is the percentage of defective parts in the sample. For subsequent analysis, therefore, it will be necessary to have the distribution of sample percentages from a twofold population, *i.e.*, a formula giving the probabilities of sample percentages taking on various values in repeated random sampling from a given twofold population. The derivation of this sampling distribution is the next task.

Derivation of the Sampling Distribution. To derive a sampling distribution it is necessary to find a mathematical model that symbolizes the conditions of sampling and to use this model for the calculation of the desired probabilities. In drawing up the model, therefore, the first step is to note carefully the conditions of sampling. In the present problem these may be described as follows:

Conditions of Sampling. Prior to the withdrawal of any sample let the percentage of defective parts in the whole population be p_1 , and let the percentage of nondefective parts be p_2 . After the withdrawal of the first member of a sample, the percentages of defective and nondefective parts remaining in the population will not, if the population is finite, be exactly p_1 and p_2 . It is one of the fundamental assumptions of the problem, however, that the changes in p_1 and p_2 will be so slight that they may be ignored. Hence the percentages of defective and nondefective parts in the population when the second member of the population is to be drawn will still be practically p_1 and p_2 . The same reasoning may be applied to subsequent withdrawals, so that on each draw the percentages of defective and nondefective parts in the population will be for all practical purposes p_1 and p_2 . The withdrawal of a sample of N cases from the given population without replacement will therefore be considered to give practically the same results as withdrawals with replacement. The following analysis will assume, then, that a sample item is replaced before the next item is drawn.

The immediate problem is a study of what will happen under repeated sampling. If the sampling process is random, each member of the population will have an equal chance of being drawn on every occasion. This means that as a large number of samples are drawn, with replacements, every member of a population will presumably be drawn sometime or other with every

member of every other population and in the long run no particular combination of members will appear any more frequently than any other combination.¹

The Mathematical Model. These conditions of sampling suggest that the sampling distribution of percentages may be derived from the following model: Suppose there are ten packs of cards in each of which p_1 per cent are black cards and p_2 per cent are white cards. All possible combinations of N cards each are formed by combining without restriction each card of each pack with every card of every other pack. Since the probability (relative frequency) of a black card in each of the 10 packs from which the various cards are selected is p_1 and the probability of a white card is p_2 , and since the combination of cards is without restriction so that the selection of a black or white card from any pack does not affect the probabilities in other packs, then, according to the multiplication theorem, the probability of a combination in which, for example, the first 4 cards are black and the last 6 are white is $p_1^4 p_2^6$. But the same would be true for any combination having 4 black cards and 6 white cards.

Since there are $C_{10}^4 = \frac{10 \cdot 9 \cdot 8 \cdot 7}{1 \cdot 2 \cdot 3 \cdot 4} = 210$ ways of picking 10 cards so that 4 of them will be black, there will be 210 such combinations. The probability of a combination having 40 per cent black cards and 60 per cent white will therefore be $210 p_1^4 p_2^6$. In general, for N packs, the probability of a combination having $\frac{N_1}{N}$ per cent black cards and $\frac{N - N_1}{N}$ per cent white cards will be

$$\text{The model: } C_{N_1}^N p_1^{N_1} p_2^{N-N_1} = \frac{N!}{N_1!(N - N_1)!} p_1^{N_1} p_2^{N-N_1}$$

If this mathematical model is correct, the equation just given will yield the probabilities of samples having $\frac{N_1}{N}$ per cent defective parts and $\frac{N - N_1}{N}$ nondefective parts in the whole set of

¹ It should be carefully noted here that the members are identified as individuals so that a particular combination means a particular combination of members. Thus a combination of three members in which the first two were defective and the last nondefective would not be the same combination in which the first was defective and the last two nondefective.

samples of N each that might be drawn at random (with replacements) from a population containing p_1 per cent defective and p_2 per cent nondefective parts. For, as noted above, if the process of sampling is random and each sample item is replaced before the next is drawn, it is reasonable to suppose that every particular combination of parts will appear as frequently as every other particular combination of parts so that the probability of samples having $\frac{N_1}{N}$ per cent defective and $\frac{N - N_1}{N}$ nondefective parts will tend to be the same as the probability of combinations of cards having $\frac{N_1}{N}$ per cent black and $\frac{N - N_1}{N}$ per cent white cards among the set of all possible combinations of cards that might be made of one card from each of N different packs, each pack containing p_1 per cent black and p_2 per cent white cards. Although this model is derived for sampling with replacements, it also gives approximate results, as noted above, for sampling without replacements.¹

If the percentage of defective parts in the sample $\frac{N_1}{N}$ is chosen as the sample statistic, the sampling distribution of this statistic will therefore be

$$P \frac{N_1}{N} = \frac{N!}{N_1!(N - N_1)!} p_1^{N_1} p_2^{N - N_1} \quad (1)$$

This will be recognized as the general equation for a binomial distribution.² Hence it follows that the sampling distribution of a percentage has³ a mean equal to p_1 , a standard deviation equal to $\sqrt{\frac{p_1 p_2}{N}}$, a $\beta_1 = \frac{(p_2 - p_1)^2}{N p_1 p_2}$, and a $\beta_2 = 3 + \frac{1 - 6 p_1 p_2}{N p_1 p_2}$.

Factors Affecting the Distribution. *Effect of Size of the Sample.* The foregoing equations show that the character of the sampling

¹ See p. 188.

² See p. 43.

³ Cf. p. 45. Since the attribute is taken as $\frac{N_1}{N}$ instead of N_1 , as in the previous discussion of the binomial distribution, the mean of $\frac{N_1}{N}$ is $\frac{N p_1}{N} = p_1$ and the standard deviation of $\frac{N_1}{N}$ is $\sqrt{\frac{N p_1 p_2}{N^2}} = \sqrt{\frac{p_1 p_2}{N}}$. Since β_1 and β_2 are coefficients their values are unaffected by this change in units.

distribution of a sample percentage from a twofold population is dependent on the size of the sample N . With a small value of N , the distribution of probabilities may be very skewed, especially if p_1 is markedly different from p_2 , and the standard deviation will be large. With a large value of N , the distribution will be more symmetrical (β_1 approaches 0 as N increases), and the standard deviation will be small. If large samples are taken, therefore, the probabilities of samples in which the percentage of a selected attribute deviates slightly from the population percentages will be more or less symmetrically distributed about the population percentage as a central value and the probabilities of samples in which the percentage deviates by a large amount from the population percentage will be very small. That is, the larger a sample, the smaller the probability of its differing greatly from the population.

A further consequence of increasing the size of the sample N is that the distribution of probabilities tends to approach the normal curve whose mean and standard deviation is the same as that of the binomial distribution.¹ Hence, in large samples, the probabilities of various types of samples can be computed from a normal probability table.² This is an especially important conclusion for the practical application of the mathematical model to the problem.

Effect of the Population Parameter p_1 . The sampling distribution described by Eq. (1) gives the probabilities of various types of samples on the assumption that the percentage of defective parts in the whole population is p_1 . As the value of p_1 is varied, the character of the sampling distribution will be changed. For the subsequent discussion it is important to note the nature of these changes.

For samples of size 10 (this small size is taken temporarily for purposes of exposition), Fig. 53 and Table 23 show how the dis-

¹ Cf. pp. 46-47.

² Actually the ordinates of the binomial distribution are approached by those of the normal curve (see Appendix to Chap. IV, pp. 68-74). If, however, the probability of a given value of N_1/N is represented, not by the height of a line or bar, but by the area of a rectangle whose base is δ and whose height is $P\left(\frac{N_1}{N}\right)/\delta$, then as N is increased (and hence δ , which equals $\sqrt{p_1 p_2 / N}$, is decreased) the rectangles become thinner and thinner and their area tends to be approximated by that of the standard normal curve.

tribution of a sample percentage changes as the value of p_1 is varied from 0 to 1 by intervals of .05. It will be noted that from $p_1 = 0$ to $p_1 = .5$ the sampling distribution changes from a positively skewed distribution to a symmetrical one, and then from $p_1 = .5$ to $p_1 = 1$ it changes to a negatively skewed distribution. The reader should here center his attention on the character of the rows of Table 23 or on the left-right variation

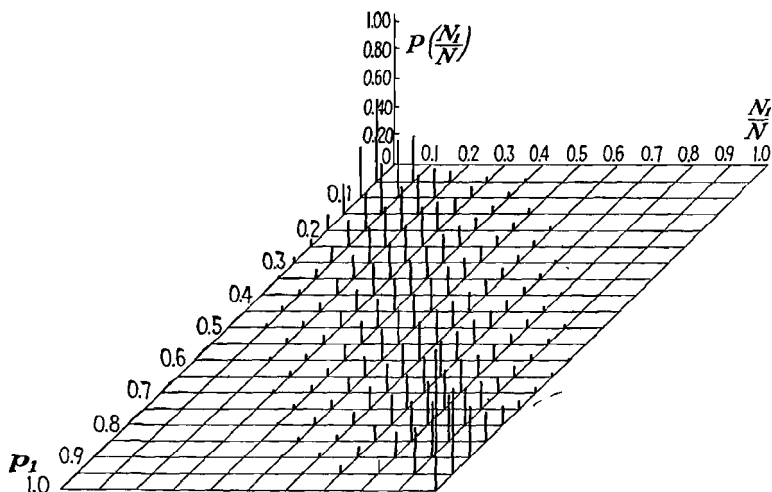


FIG. 53.—The variation in the character of the distribution of a sample percentage with changes in the population parameter p_1 .

of Fig. 53 as he proceeds from back to front. It will also be noted that for a given sample percentage N_1/N the probability of the sample rises as the population percentage p_1 approaches the point $p_1 = N_1/N$. That is, the probability of a given sample result is a maximum when $p_1 = N_1/N$. Here the reader should center his attention on the variation from row to row for a given column of Table 23 or from back to front for a given value of N_1/N in Fig. 53.

If the population is very large, variation in p_1 can be assumed to be practically continuous, that is, p_1 can be given any hypothetical value between 0 and 1. On the other hand, for samples of size 10, only 11 values of N_1/N are possible; variation in this quantity is therefore discrete. Accordingly, Fig. 53 consists essentially of a series of curves running from back to front and separated from left to right by intervals of .1. If sampling is

made with a much larger sample, say a sample of 2,500, the variation in N_1/N may also be considered to be practically continuous. In that instance, Fig. 53 can be replaced by the smooth surface of Fig. 54. As pointed out in the previous section, except for the extreme ends where p_1 is very small, the left to right sections of this surface are practically normal curves.

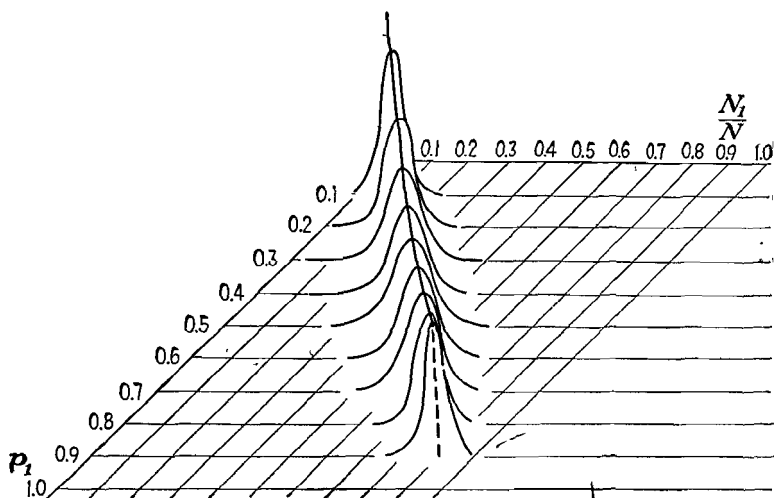


FIG. 54.—The surface that approximates Fig. 53, when N is large.

USES OF THE DISTRIBUTION OF THE SAMPLE PERCENTAGE FOR TESTING HYPOTHESES

Up to this point the argument has been purely deductive. Given a certain population, it has been asked, how frequently will random sampling produce certain types of samples? The crucial part of the argument is the reverse of this: Given a certain sample, what inferences can be made about the population from which it was obtained? It is this part of the argument that will now be examined.

The Hypothesis. With reference to the original illustration, suppose that the manufacturer of machine parts is particularly interested in the hypothesis that the whole set or population of parts has just the standard 10 per cent of defectives. He takes a sample of 2,500 parts and finds that 11 per cent of the sample are defective. How does his hypothesis fare in the light of this sample result?

Coefficient of Risk. The first step is to decide upon the coefficient of risk. The manufacturer recognizes, of course, that even if the number of defective parts in the whole population is just 10 per cent he might by chance get a sample in which the percentage of defective parts is much greater or less than 10 per cent. There is always some risk that he will reject the 10 per cent hypothesis when it is actually true. The initial step, therefore, is to decide upon how great a risk of this kind he is willing to run. Suppose he decides that on the average he does

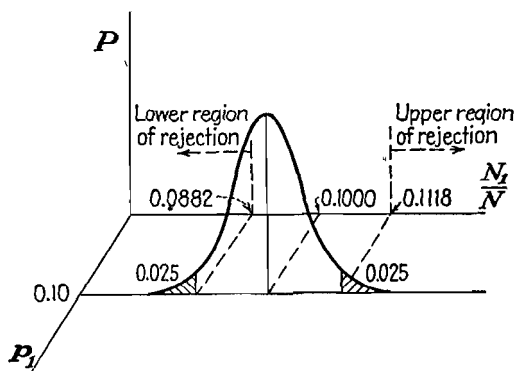


FIG. 55.—A region of rejection for testing the hypothesis $p_1 = .10$. Coefficient of risk = .05.

not want to reject the hypothesis when it is true more than 5 times out of 100. He will then adopt a coefficient of risk of .05.

A .05 Region of Rejection. From the analysis of the previous section, the manufacturer knows that if he takes a sample of 2,500 from a large population the probabilities of the various possible types of samples will be given approximately by a normal probability curve whose mean is at the population percentage p_1 and whose standard deviation is $\sqrt{p_1 p_2 / N}$. From this he knows that if the number of defectives in the whole population is actually 10 per cent ($p_1 = .10$), then the probability of obtaining a sample in which the percentage of defectives is equal to or less than $.10 - 1.96\sigma$ is .025 (see normal table in Appendix, Table VI):

He also knows that the probability of obtaining a sample in which the percentage of defectives is equal to or greater than $.10 + 1.96\sigma$ is likewise .025. Hence, according to the addition theorem, the probability of a sample in which the percentage of defectives lies outside of the limits $.10 + 1.96\sigma$ and $.10 - 1.96\sigma$

is $.025 + .025 = .05$. Since $\sigma = \sqrt{\frac{(.1)(.9)}{2,500}} = .006$, the actual values of these limits are 8.82 per cent and 11.18 per cent (see Fig. 55). It follows, therefore, that the risk of rejecting the hypothesis when it is true will be just .05 if the manufacturer rejects it whenever a sample contains defective parts numbering less than 8.82 per cent or more than 11.18 per cent. For if the number of defectives in the whole population is exactly 10 per cent, samples will fall in these extreme regions (below 8.82 per

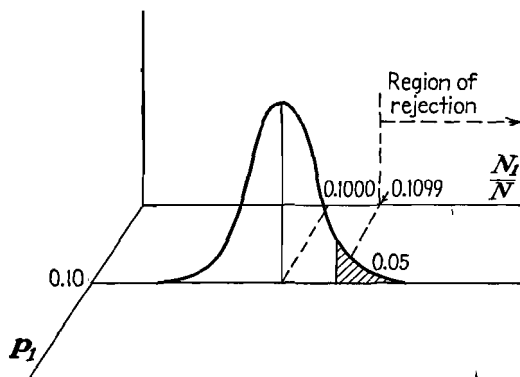


FIG. 56.—An alternative region of rejection for testing the hypothesis $p_1 = .10$
Coefficient of risk = .05.

cent and above 11.18 per cent) only 5 per cent of the time. Hence, if the manufacturer follows this rule he will, on the average, reject the hypothesis when it is true only 5 times out of 100, which is the degree of risk he is willing to undergo.

Alternative Regions of Rejection. It will be noted, however, that the suggested region of rejection is not the only one that the manufacturer might adopt. He can follow the rule of rejecting the hypothesis whenever the sample percentage falls above $.10 + 1.645\sigma$, that is, above 10.99 per cent, and his risk of rejecting the hypothesis when it is true will still be only .05. For, as will be noted from a table of the normal curve, the probability of a deviation from the mean of 1.645σ or more is just .05. The region beyond 10.99 per cent is therefore an alternative region of rejection with a coefficient of risk of .05. This is illustrated in Fig. 56. Still a third region with a coefficient of risk of .05 is the region lying below $.10 - 1.645\sigma$, that is,

below 9.11 per cent. For, owing to the symmetry of the normal curve, the probability of a negative deviation from the mean of 1.645 σ or more is also .05. This third alternative region of rejection is illustrated in Fig. 57. Yet a fourth region with a coefficient of risk of .05 is the region lying below $.10 - 2.327\sigma$

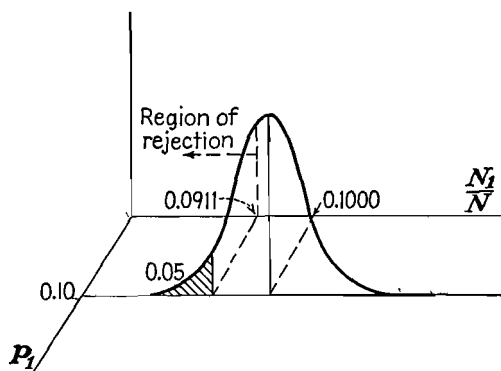


FIG. 57.—A third alternative region of rejection for testing the hypothesis $p_1 = .10$. Coefficient of risk = .05.

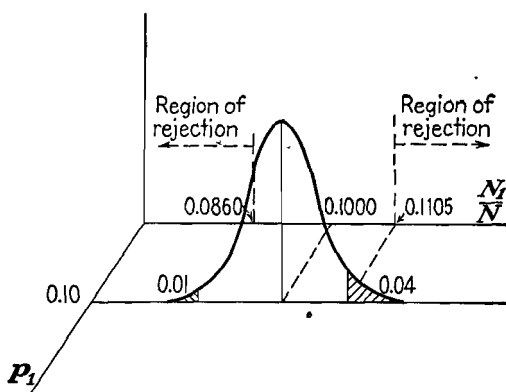


FIG. 58.—A fourth alternative region of rejection for testing the hypothesis $p_1 = .10$. Coefficient of risk = .05.

and above $.10 + 1.751\sigma$, that is, below 8.60 per cent and above 11.05 per cent. This is illustrated in Fig. 58.

In fact, there are an infinite number of regions of rejection that the manufacturer might adopt, all of which have associated with them a coefficient of risk of .05. What is the criterion that he should follow in choosing from among these various possible regions?

The Best Region of Rejection. Each region of rejection has a corresponding region of acceptance. As suggested in the previous chapter, the "best" region of acceptance would be the one for which the risk of accepting a given hypothesis when it is not true is a minimum. Since the probability of a sample falling in the region of acceptance is equal to 1 minus the probability of its falling in the region of rejection, the former will be a mini-

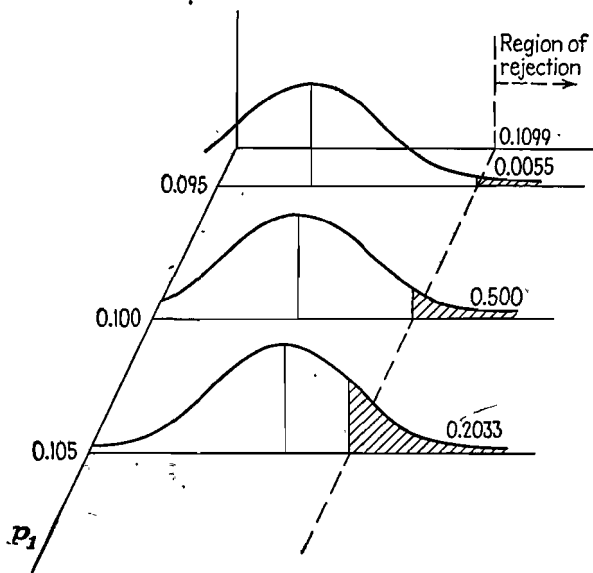


FIG. 59.—The probability of a sample falling in the region of Fig. 56 varies with the population percentage p_1 , illustrating the procedure for finding the data of Table 24 from which Fig. 60 is plotted.

imum when the latter is a maximum. The best region of rejection, therefore, will be a region for which the probability of rejection when the given hypothesis is not true is a maximum.¹

How the various regions suggested in the previous paragraph fare in respect to this criterion is indicated by the probabilities shown in Table 24. Table 24 gives data for each of the four alternative regions, showing how the probability of a sample falling in the region varies with the population percentage p_1 .

¹ In technical language this probability of rejection when the hypothesis is not true is called the "power" of the test associated with the selected region. The criterion is to select the region with the highest power. The region so selected will then give the "most powerful" test.

The method of computing the data in Table 24 is illustrated in Fig. 59, and the results are presented in Fig. 60, which is based upon Table 24.

It is obvious from this figure that there is no one region for which the probability of a sample is the greatest for all possible values of the population percentage different from the one set up by hypothesis. The region above .1099 (the region of Fig. 56) has the highest probability for population percentages

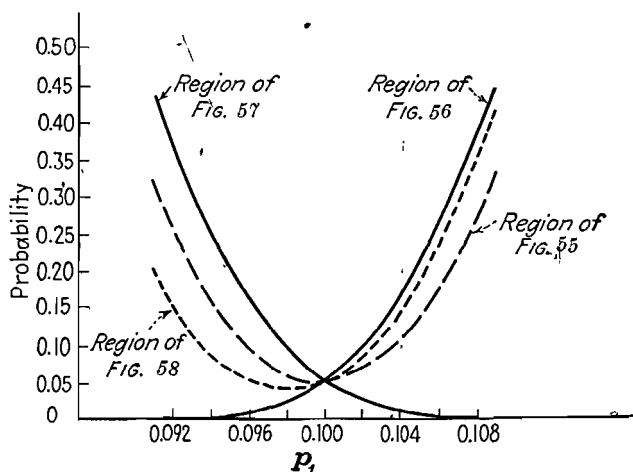


FIG. 60.—Effect that variation in the population percentage p_1 has on the probability of a sample falling in given regions of rejection.

greater than the hypothetical percentage being tested (that is, $p_1 = .10$ in this case), and the region below .0911 (the region of Fig. 57) has the highest probability for population percentages less than the hypothetical percentage. The region below .0882 and above .1118 (the region of Fig. 55) does well for values of p_1 both greater and less than the hypothesis being tested, but it does not have the maximum probability for any values. There is not in this instance any single region of rejection that is best for all problems.

In order to determine the region of rejection that is best for the present problem the manufacturer must draw upon other considerations. Suppose he wishes particularly to avoid unnecessary and expensive refinements in production methods and accordingly aims to keep the percentage of defectives as close as possible to the accepted standard of 10 per cent defective

parts.¹ At the same time, suppose he does not care if orders are sent out that are below standard; it may be that the purchaser will carry out his own test to avoid buying a substandard set. Accordingly, the manufacturer will select the region of Fig. 57,

TABLE 24.—VARIATIONS IN THE PROBABILITY OF A SAMPLE FALLING IN ANY GIVEN REGION FOR DIFFERENT VALUES OF THE POPULATION p_1 *

Values of p_1	Probability of a sample in			
	Region of Fig. 55	Region of Fig. 56	Region of Fig. 57	Region of Fig. 58
.091	.3157	.0005	.4404	.1952
.092	.2581	.0010	.3745	.1522
.093	.2067	.0018	.3121	.1174
.094	.1623	.0033	.2546	.0876
.095	.1272	.0055	.2033	.0671
.096	.0971	.0091	.1587	.0524
.097	.0758	.0154	.1270	.0435
.098	.0619	.0227	.0934	.0401
.099	.0521	.0344	.0694	.0464
.100	.0500	.0500	.0500	.0501
.101	.0597	.0708	.0359	.0636
.102	.0653	.0968	.0250	.0834
.103	.0824	.1292	.0174	.1102
.104	.1069	.1685	.0116	.1439
.105	.1389	.2033	.0078	.1851
.106	.1782	.2643	.0049	.2333
.107	.2218	.3228	.0032	.2846
.108	.2718	.3821	.0020	.3448
.109	.3304	.4443	.0013	.4053

* These figures were calculated as follows: The mean value $Np_1 = 2,500p_1$ was calculated for each value of p_1 and also the standard deviation, $\sqrt{p_1p_2/N} = \sqrt{p_1p_2/2,500}$. The difference between the mean and the boundary or boundaries of a region was then divided by the standard deviation, and the probability of a deviation greater than this value, or values, of z/σ was found from a normal probability table. These are the figures given in the table above.

i.e., below .0911, as the best region for his purpose. He will thus accept the hypothesis that the number of defectives in the population is 10 per cent least often when that percentage is actually below 10 per cent.

¹ The assumption is that if a sample shows a percentage of defectives that is considered too low, the manufacturer would discontinue some expensive refinement in the method of production that might lower the quality somewhat but would yield him more profit.

On the other hand, if the manufacturer desires especially to protect his reputation and wishes particularly to avoid sending out an order that is actually below standard, he will select the region of Fig. 56, *i.e.*, above .1099, as the best region to use. In this instance, he will accept the hypothesis of 10 per cent least often when the population percentage is actually above 10 per cent. He runs the least risk in this case of damaging his reputation.¹

If the manufacturer wishes especially to avoid the danger of damaging his reputation by sending out substandard lots but nevertheless desires to give some consideration to the additional expense of attaining too high a standard, he might adopt a region such as that lying above .1105 and below .0860, that is, the region of Fig. 58. If he is equally indifferent or equally concerned about his reputation and the additional expense of attaining a high standard, he would do best to choose the region lying above .1118 and below .0882, the region of Fig. 55. For this latter region favors values neither above nor below the hypothesis being tested; it is an "unbiased" region.²

In the problem illustrated, suppose that the manufacturer adopts the unbiased region lying below .0882 and above .1118, the region of Fig. 55. He knows that if this region is always adopted in his sampling tests he will in the long run reject the 10 per cent hypothesis when it is actually true only 5 times out of a 100. He also knows that the probability of accepting the 10 per cent hypothesis when it is not true is evenly distributed with respect to actual percentages above and below 10 per cent. He will in this instance not cause undue damage to his reputation nor induce himself to incur an unnecessary expense of production.

¹ That is, least risk for the size region adopted. If the manufacturer had adopted a coefficient of risk of .10 of rejecting the 10 per cent hypothesis when it was actually true, his regions of rejection would all have been larger and his regions of acceptance correspondingly smaller. There would be less risk of accepting the hypothesis in this case when it was not true. When it is said in the text that a certain region gives the least risk of accepting the hypothesis being tested when some other values are true, it is meant that this is the least risk for any region of the specified size, *i.e.*, for any region for which the associated risk of rejection is the specified coefficient .05.

² Generally a region is said to be "unbiased" if the probability of a sample falling within it is less when the given hypothesis is true than when any alternative hypothesis is true.

The Test of the Hypothesis. The actual sample result is 11 per cent. Since this does not fall in the adopted region of rejection, the hypothesis of $p_1 = 10$ per cent is not rejected and the given lot of parts is deemed to be of standard quality.

USE OF THE DISTRIBUTION OF A SAMPLE PERCENTAGE TO ESTIMATE THE VALUE OF THE POPULATION PERCENTAGE

The foregoing pages were primarily concerned with the rejection or acceptance of a particular hypothesis regarding the percentage of defective parts in the population. It did not indicate the limits within which this percentage might reasonably lie, nor did it indicate what might be a good estimate of the actual percentage of defectives in the whole population. Often these considerations are of more importance than the testing of a particular hypothesis.

Determining Confidence Intervals. The procedure for determining a reasonable range for a population parameter, given a particular sample, is much the same as that followed in testing hypotheses regarding that parameter. First it is necessary to decide upon the degree of risk that is to be run in failing to include the true value of the parameter within the estimated range. To put it positively, it is first necessary to determine the degree of confidence that may be had in the estimated range covering the true value. Furthermore, if no one type of range is found to be the best, it is necessary to decide whether failure to include the true value because the range is placed too low is of equal importance as failure to include the true value because the range is placed too high. These matters will now be considered in some detail.

Determining an Unbiased Interval. In the example of the preceding section the manufacturer of machine parts took a sample of 2,500 parts from a large lot and found that, of these 2,500 parts, 275, or 11 per cent, were defective. On the basis of this sample, the manufacturer may determine an unbiased¹ con-

¹ The interval is called "unbiased" because it is so determined that the probability of the interval failing to cover the true value because it is too high is equal to the probability of the interval failing to cover the true value because it is too low. It happens in the present instance that this unbiased interval is also symmetrical about the sample value. In other cases, however, it will be found that an interval that is unbiased in the probability sense is not symmetrical around the sample value (see pp. 287-289).

fidence interval as follows: The confidence coefficient associated with this interval will be put at .95. First let the manufacturer find the value of p_1 that will make the difference between it and the sample percentage, that is $\frac{N_1}{N} - p_1$, just equal to 1.96σ .

Since $\sigma = \sqrt{\frac{p_1 p_2}{N}}$, this value of p_1 will be given by the solution of the equation $\frac{N_1}{N} - p_1 = 1.96 \sqrt{\frac{p_1 p_2}{N}}$ for p_1 , which leads to the approximate equation¹

$$p_1 = \frac{N_1}{N} - 1.96 \sqrt{\frac{N N_1 - N_1^2}{N^3}} \quad (2)$$

In the present instance this formula yields the result $p_1 = .098$. It will be noted that this lower limit for p_1 is designated $\underline{p}_1$ (called " p_1 lower"). The procedure is illustrated in Fig. 61.

Next let the manufacturer find the value of p_1 that makes the difference between it and the sample percentage just equal to -1.96σ . This second value will be given by the equation

$\frac{N_1}{N} - p_1 = -1.96 \sqrt{\frac{p_1 p_2}{N}}$ or by the approximate equation²

$$\bar{p}_1 = \frac{N_1}{N} + 1.96 \sqrt{\frac{N N_1 - N_1^2}{N^3}} \quad (3)$$

For the given problem this yields $\bar{p}_1 = .122$. Again it will be noted that this upper limit for p_1 is designated by $\bar{p}_1$ (called " p_1 upper"). For illustration the reader is again referred to Fig. 61.

¹ The approximate formula is obtained by substituting $\frac{N_1}{N}$ for p_1 and $\frac{N - N_1}{N}$ for p_2 in the radical on the right, that is σ is estimated from the sample percentages. The formula given by the solution of the equation is

$$p_1 = \frac{N_1 + 1.92 - 1.96 \sqrt{.96 + N_1 - \frac{N_1^2}{N}}}{N + 3.84}$$

but when N is large, as it should be when the normal curve is taken as an approximation to the binomial distribution, this more exact formula differs but little from the approximate one given in the text.

² The exact formula is the same as that given in the preceding footnote except that the square root is now preceded by a plus sign instead of a negative sign.

These two values of p_1 , viz., $\underline{p}_1$ and $\bar{p}_1$, mark off a confidence interval that the manufacturer may claim has a probability of .95 of covering the true value. This means that, if he continuously follows the foregoing procedure in sampling from a twofold population, the interval he sets up will on the average

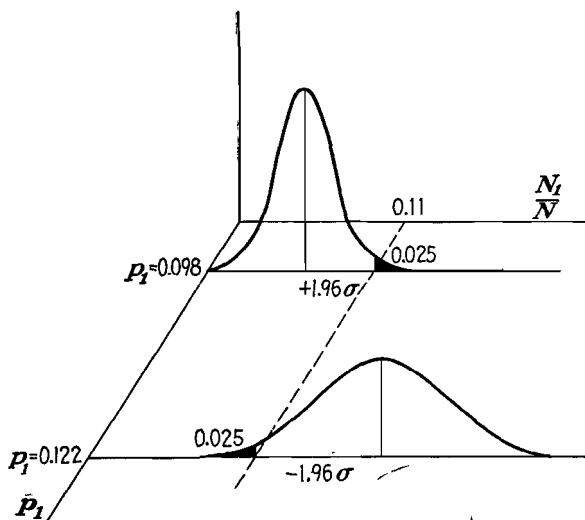


FIG. 61.—Determination of an unbiased confidence interval for p_1 . Confidence coefficient = .95. Note that $\delta = Np_1p_2$ and hence varies with the values of p_1 and p_2 .

cover the population value 95 times out of a 100. For, on the assumption that the normal curve is a good approximation to the binomial distribution, sample percentages will fall within the range $\pm 1.96\delta$ of the population percentage 95 per cent of the time. Hence, ranges of $\pm 1.96\delta$ about the sample percentage will include the population percentage 95 per cent of the time. It is only when the sample percentages deviate more than 1.96δ from the population value that a range of $\pm 1.96\delta$ about the sample percentage will fail to include the population value, and according to the normal probability curve the probability of such an event is only .05. It will be noted that the manufacturer speaks, not of the probability of the population value falling in the confidence interval, but of the probability of the interval covering the population value. For it is the interval that varies from sample to sample, not the population value; the latter is an unknown constant.

Other Confidence Intervals. The confidence interval just determined is not the only interval with a confidence coefficient of .95. Other intervals with the same confidence coefficient may also be derived. For example, it is possible for the manufacturer to select an interval that has only a variable lower limit, the upper limit being the maximum attainable value of $p_1 = 1.00$. This lower limit may be computed by finding the value of p_1 that will make the probability of getting the given sample percentage $\frac{N_1}{N}$ or a greater value just .05. Such a value of the lower limit will be given by the equation $\frac{N_1}{N} - p_1 = 1.645\sigma$ or the equation

$$p_1 = \frac{N_1}{N} - 1.645 \sqrt{\frac{NN_1 - N_1^2}{N^3}}.$$
 For the given case in which $N = 2,500$ and $N_1 = 275$, this lower limit will be .100. The manufacturer can therefore say with equal validity that the interval .100-1.000 has a chance of .95 out of 100 of including the population value.

Again, with the same confidence coefficient of .95, the manufacturer may choose an interval that has only a variable upper bound, the lower bound being the minimum possible value of 0. This upper limit can be computed by finding the value of p_1 that will make the probability of the sample percentage of a lower percentage just equal to .05. The value will be given by the equation $\frac{N_1}{N} - p_1 = -1.645\sigma$ or by the formula

$$p_1 = \frac{N_1}{N} + 1.645 \sqrt{\frac{NN_1 - N_1^2}{N^3}}.$$

For a sample of 2,500 and a sample percentage of 11, this upper limit will be .120. Hence, with as much truth as in the other cases, the manufacturer can say that the range 0-.120 has a chance of .95 of including the population value.

Finally, the manufacturer may adopt an interval determined in the following manner: He may determine a lower limit such that, if it were the population value, the probability of getting the sample percentage of 11 per cent or a greater value will be just .04. And he may determine an upper limit such that if it were the population value the probability of getting the sample percentage or a lower value will be just .01. These two limits will

be given by the equations

$$\frac{N_1}{N} - p_1 = 1.751\sigma \quad \text{and} \quad \frac{N_1}{N} - p_1 = 2.327\sigma$$

or by the equations, $p_1 = \frac{N_1}{N} - 1.751\sqrt{\frac{NN_1 - N_1^2}{N^3}}$ and

$\bar{p}_1 = \frac{N_1}{N} + 2.327\sqrt{\frac{NN_1 - N_1^2}{N^3}}.$ * For a sample of 2,500 and a sample percentage of 11, these limits will be .099 and .125. Thus the manufacturer might say, as in the other cases, that the interval .099-.125 has a chance of .95 of including the population figure.

The Best Interval. There are accordingly an infinite number of different ranges, all associated with a confidence coefficient of .95, that might be adopted in setting up confidence limits for the population parameter p_1 . Is there any one range that might be designated the "best"? Before answering this question, consider the various values covered by the different intervals.

If the confidence interval given by the limits

$$\frac{N_1}{N} \pm 1.96\sqrt{\frac{NN_1 - N_1^2}{N^3}}$$

is compared with the confidence interval given by the limits $\frac{N_1}{N} - 1.645\sqrt{\frac{NN_1 - N_1^2}{N^3}}$ and 1, it is found that for an actual population value of, say, 10 per cent the latter interval would include the value above 10 per cent, especially the higher values such as 13 per cent, 14 per cent, etc., much more frequently than the former would. For the latter always runs up to 100 per cent no matter what the sample percentage, while the former runs up as far as 13 per cent, say, only if the sample value is as high as 11.68 per cent,¹ which, on the assumption that the population value is 10 per cent, is likely to be a rare phenomenon indeed.²

* For .04 of the normal curve lies about 1.751 σ from the mean and .01 lies below -2.327σ from the mean.

¹ This is calculated from the exact formula $\frac{N_1}{N} - p_1 = 1.96\sqrt{\frac{p_1 p_2}{N}}$.

² If $p_1 = 10$ per cent, then $\sigma = .006$, $\frac{11.68 \text{ per cent} - 10 \text{ per cent}}{\sigma} = 2.80$; and the probability of a normally distributed variate exceeding its mean by more than 2.80 times σ is only .0026.

On the other hand, the second type of interval always has a larger value for its lower limit than the first, and it will therefore include values below 10 per cent (the actual population value) less frequently than the first type of interval. If the same type of comparison were made between the interval given by the limits $\frac{N_1}{N} \pm 1.96 \sqrt{\frac{NN_1 - N_1^2}{N^3}}$ and the interval given by the

limits 0 and $\frac{N_1}{N} + 1.645 \sqrt{\frac{NN_1 - N_1^2}{N^3}}$, it would be found that the former would include values of p_1 below the actual value of 10 per cent much less frequently than the latter, while the latter would include values above 10 per cent less frequently.

In the problem illustrated it would seem that the manufacturer would prefer the interval running from 0 to

$$\bar{p}_1 = \frac{N_1}{N} + 1.645 \sqrt{\frac{NN_1 - N_1^2}{N^3}},$$

for this would put the best face on his product. It would include the true value 95 per cent of the time, just as the other intervals would, and it would put the upper limit no higher than necessary. The buyer would be at no disadvantage if this method of stating the limits were employed. For if he understood their derivation he would know that 5 times out of 100 he might get a lot in which the percentage of defective parts exceeded the stated upper limit, and he would not buy unless he were willing to run this risk.

Influence of the Size of the Sample on the Confidence Interval. It will be noted that in the foregoing analysis the size of an interval depended upon the value of σ and this in turn depended on the value of N , the size of the sample. Since $\sigma = \sqrt{p_2 p_2 / N}$, it is seen that the larger the sample, the smaller the value of σ , and hence the narrower the confidence interval. That is, for a given confidence coefficient, say a probability of .95, the larger the sample, the closer the population value may be estimated.

Influence of the Size of the Confidence Coefficient on the Confidence Interval. The size of the confidence interval is also dependent on the size of the confidence coefficient. For example, if in setting up an unbiased confidence interval the manufacturer had been willing to have the interval cover the population value only 90 times out of 100, he could have set up limits given by

$\frac{N_1}{N} \pm 1.645\sigma$ instead of $\frac{N_1}{N} \pm 1.96\sigma$.¹ Accordingly, by choosing a lower confidence coefficient, the manufacturer could have narrowed his confidence interval. The same is true for other types of confidence intervals.

A Single Estimate of p_1 . The foregoing has had to do with the determination of a range of reasonable values for p_1 . In some instances, however, the manufacturer may desire a single estimate of the population percentage. He may wish, for example, to report to a buyer some measure of the quality of the lot being sold, and an estimate of the percentage of defective parts in the lot may be a suitable figure for this purpose. Estimation of population parameters by the method of maximum likelihood was outlined in the previous chapter. Consider how it may be applied to the present instance.

In estimating p_1 from the sample result, the manufacturer might proceed as follows: He might turn to Fig. 53 or to its more general form, Fig. 54, and pick out the point on the N_1/N axis representing his sample percentage. The chart would have to be one for which the size of the sample, N , was the same as that taken by the manufacturer. Then he could proceed in the direction of increasing p_1 values until he had reached the highest point on the surface, *i.e.*, until he found the value of p_1 that made the probability of the given sample a maximum. If he had taken a sample of 10 items, for example, and 2 had been found defective, he would start at the point .2 on the N_1/N axis of Fig. 53 and proceed perpendicularly to this axis until he had reached the value $p_1 = .2$. There he would find that the probability of his sample result would be a maximum. The maximum-likelihood estimate of p_1 would be .2, the same as the sample percentage. This, and all estimates made in this way, are called "maximum-likelihood estimates," since they choose the value of the population parameter, in this case p_1 , so that the probability, or likelihood, of the given sample result, here N_1/N , is a maximum.

In general, the maximum-likelihood estimate of a population percentage is the sample percentage. This may be demonstrated mathematically as follows: Equation (1) shows that if the percentage of attribute A in the population is p_1 , the probability

¹ For the probability of a normally distributed variate deviating from its mean by $\pm 1.645\sigma$ or more is just .10.

of a sample of N having N_1 A's is¹

$$P\left(\frac{N_1}{N}\right) = \frac{N!}{N_1!(N - N_1)!} p_1^{N_1}(1 - p_1)^{N - N_1}$$

According to the differential calculus, the value of p_1 that will maximize this probability for a given value of N_1 and N must satisfy the condition²

$$\frac{d\left[P\left(\frac{N_1}{N}\right)\right]}{dp_1} = N_1 p_1^{N_1-1} (1 - p_1)^{N - N_1} - p_1^{N_1} (N - N_1) (1 - p_1)^{N - N_1 - 1} = 0$$

which reduces to

$$N_1(1 - p_1) - p_1(N - N_1) = 0$$

or

$$p_1 = \frac{N_1}{N}$$

If the maximum-likelihood estimate of p_1 is written $\hat{p}_1$ to distinguish it from the population value itself, then $\hat{p}_1 = N_1/N$ is the maximum-likelihood estimate of p_1 , that is, the sample percentage N_1/N is the maximum-likelihood estimate of the population percentage p_1 .

SAMPLES FROM RELATIVELY SMALL POPULATIONS

One of the fundamental assumptions of the foregoing analysis was that the population was very large relative to the sample taken. For it was on this basis that the probability of a given attribute in the population was assumed not to be materially affected by the withdrawal of various members of the sample. It was argued, for example, that, if a sample of 2,500 parts was taken from a lot of 500,000 parts in which the actual number of defective parts was 10 per cent, the percentage of defectives in the 499,000 parts left after 1,000 parts had been drawn would still be close to 10 per cent (actually, 9.82 per cent) even if all the 1,000 parts drawn turned out to be defective. Hence the analysis proceeded on the assumption that the percentage of defective parts in the population remained unaffected as the sample items were taken out. That is, the sampling discussed was sampling with replacements. It is the purpose of this section

¹ Here p_2 is replaced by its equal $1 - p_1$.

² Generally the logarithm of the probability is maximized, but here it is just as easy to differentiate the function itself.

to consider how the analysis must be modified when nonreplacement of the sample items is taken into account.

When the population is of such a size relative to the sample that consideration must be given to the effects of the withdrawal of the sample items on the percentage of a given attribute in the population, then the distribution of a sample percentage can no longer be adequately described by the binomial distribution. Instead, as may be shown by an analysis similar to that of Chap. IV, the mathematical law describing the distribution of a sample percentage when replacements are not made is the hypergeometrical distribution. Thus, if a random sample of N cases is taken from a population consisting of S cases and if the percentage of cases in the population having attribute A is p_1 and the percentage not having the attribute A is $1 - p_1 = p_2$, then the probability of a sample (drawn without replacements) containing N_1/N A 's, is given by the formula

$$P\left(\frac{N_1}{N}\right) = \frac{(p_1 S)!(p_2 S)!(S - N)!N!}{(p_1 S - N_1)!(p_2 S - N + N_1)!S!N_1!(N - N_1)!} \quad (4)$$

This substitution of the hypergeometrical distribution for the binomial distribution is the fundamental modification that must be made in the previous analysis when sampling is made without replacements.

If both the population and the sample are relatively small, then either the hypergeometrical distribution itself may be employed to calculate the probabilities of various sample results or one of the Pearsonian frequency curves that approximates this distribution can be used for this purpose. Calculations of this kind are quite complicated, however, and will not be further considered here. The interested reader is referred to the Appendix to Chap. IV (page 81) for a discussion of the appropriate Pearsonian curve to fit and to pages 127 to 131 for a method of measuring probabilities from such a curve.

When the population and sample are both moderately large, the hypergeometrical distribution can be approximated by a normal distribution, which again greatly simplifies the analysis. Specifically, if the sample N is less than half the population S , if $1/\sqrt{N}$ and hence $1/\sqrt{S - N}$ are so small as to be negligible, and if Np_1 and Np_2 are both moderately large, then the hypergeometrical distribution can be approximated by a normal distribu-

tion; whose mean is p_1 and whose standard deviation is given by the equation¹

$$\sigma = \sqrt{\frac{p_1 p_2}{N} \left(1 - \frac{N}{S}\right)} \quad (5)$$

When it is possible to use the normal distribution in place of the hypergeometrical distribution, the analysis proceeds in almost exactly the same manner as described in the preceding sections. The only difference is that, instead of using the value $\sqrt{\frac{p_1 p_2}{N}}$, Eq. (5) is used in its place. It will be noticed that the effect of this modification is to reduce the value of σ employed, since $1 - \frac{N}{S}$ is necessarily less than 1. This means that confidence intervals, regions of acceptance, and the like, will be smaller when account is taken of the size of the population. As S becomes large relative to N , however, the factor $1 - \frac{N}{S}$ becomes practically 1 and the size of the population may be disregarded. The analysis then becomes that of the previous sections of this chapter.

SAMPLES FROM POPULATIONS IN WHICH p_1 IS VERY SMALL

There are some instances in which the chance of a "favorable" occurrence is very small. If a large enough sample is taken, however, a few cases will be found. The example most commonly referred to is that of Borthewitch, who found that for 10 Prussian army corps the average number of deaths occurring from horse kick during 10 years was .61 per corps per year. Another example is the counting of cells in certain biological tests and experiments where the chance of occurrence of cells of a certain type on a given plate or a given section of a plate is very small.²

Whenever the chance of a favorable case is very small, the binomial distribution is not well approximated by the normal distribution unless N is very large.³ In these instances the

¹ Cf. BOWLEY, A. L., *Elements of Statistics*, 5th ed., pp. 282-284.

² Cf. FISHER, R. A., *Statistical Methods for Research Workers*, par. 15.

³ See p. 73.

binomial distribution is better approximated by the Poisson distribution.¹ This has the form²

$$P(N_1) \doteq \frac{\mathbf{m}^{N_1}}{N_1!} e^{-\mathbf{m}} \quad (6)$$

which may also be written

$$P(N_1) \doteq \frac{(N\mathbf{p}_1)^{N_1}}{N_1!} \exp [-N\mathbf{p}_1]$$

where $P(N_1)$ represents the probability of N_1 cases out of N , $\mathbf{p}_1$ is the probability of a favorable case $\mathbf{m} = N\mathbf{p}_1$, and $e = 2.718+$ is the base of Napierian logarithms. For example, if $\mathbf{p}_1 = .003$ and $N = 1,000$, then $\mathbf{m} = 3$ and the probabilities of 0, 1, 2, 3, 4, . . . favorable occurrences are³

N_1	$P(N_1)$	
0	$e^{-3} \frac{3^0}{1}$	= .0498
1	$e^{-3} \frac{3^1}{1}$	= .1494
2	$e^{-3} \frac{3^2}{2}$	= .2240
3	$e^{-3} \frac{3^3}{6}$	= .2240
4	$e^{-3} \frac{3^4}{24}$	= .1680
5	$e^{-3} \frac{3^5}{120}$	= .1008
6	$e^{-3} \frac{3^6}{720}$	= .0504
7	$e^{-3} \frac{3^7}{5,040}$	= .0216
8	$e^{-3} \frac{3^8}{40,320}$	= .0081
9	$e^{-3} \frac{3^9}{362,880}$	= .0027
10	$e^{-3} \frac{3^{10}}{3,628,800}$	= .0008
.	.	.
.	.	.
.	.	.

¹ Also known as the "law of small numbers" or "law of small chances."

² The derivation of this approximation is given in the Appendix to this chapter (p. 215).

³ These probabilities may be computed directly by noting that

These are graphed in Fig. 62.

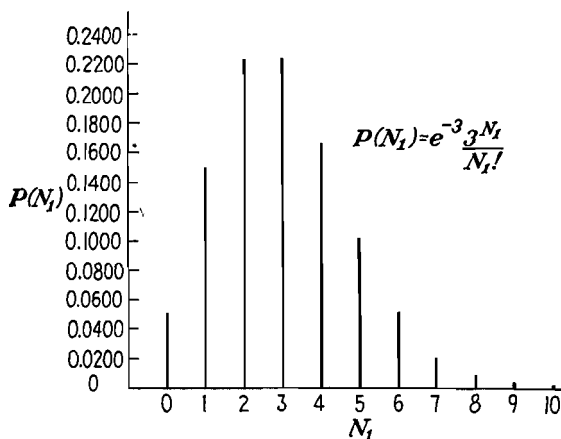


FIG. 62.—The Poisson distribution, $m = 3$, $p_1 = .003$, $N = 1000$ [see Eq. (6)].

If it is desired to find the probability of N_1 exceeding or falling short of a given value, it is necessary merely to sum the individual probabilities in question. For example, if $p_1 = .003$ and $N = 1,000$ as in the table above, the probability that two or fewer favorable cases will occur is equal to

$$.0498 + .1494 + .2240 = .4232.$$

The probability that six or more favorable cases will occur equals $.0504 + .0216 + .0081 + .0027 + .0008 + \dots$, which equals approximately .0836. Since the table is not completed in this direction owing to the small probabilities, the probability in question is more accurately computed by subtracting from 1 the probability of five or less favorable occurrences. The latter probability equals

$$.0498 + .1494 + .2240 + .2240 + .1680 + .1008 = .9160.$$

Hence a better approximation to the probability of six or more favorable cases is $1 - .9160 = .0840$.

The use of the Poisson distribution in testing a hypothesis about a population may be illustrated as follows: Suppose it is

$$\log_{10} e^{-3} = -3 \log_{10} e = -3(.4343) = -1.3029$$

and therefore $e^{-3} = .0498$, or they may be obtained from Karl Pearson's *Tables for Statisticians and Biometricians*, Table LI.

wished to determine the chance of dying from a new type of inoculation, say against smallpox. The existing method, it will be assumed, results in the death of not more than .5 per cent of those inoculated, and the new method will not be adopted widely if its death rate appears to be significantly more than that. A thousand test cases are made, and of these seven die. Is it reasonable on the basis of these tests to reject the hypothesis that the true death rate from the inoculation is more than .5 per cent?

To answer this question, set the coefficient of risk at approximately .05, and place the region of rejection all at the upper end of the distribution. (For the risk of accepting the hypothesis when the death rate is actually higher than .5 per cent is to be made a minimum.) It will be noted that the Poisson distribution, like the binomial distribution, is discrete so that it may not be possible to obtain a region of rejection for which the probability is exactly .05. One coming as close as possible to this standard, however, will be adopted.

If .5 per cent is the true death rate and samples of 1,000 are taken, the distribution of sample death rates will be approximated by a Poisson distribution for which $m = (1,000)(.005) = 5$. The individual probabilities for the more likely sample results will thus be¹

N_1	$P(N_1) = \frac{e^{-5} 5^{N_1}}{N_1!}$
0	.0067
1	.0337
2	.0842
3	.1404
4	.1755
5	.1755
6	.1462
7	.1044
8	.0653
9	.0363
10	.0181
11	.0082
12	.0034
13	.0013
14	.0005
15	.0002

¹ These are taken from Pearson's *Tables for Statisticians and Biometricians*, Table LI.

The total probability of a sample containing 8 or less deaths is .9319; thus the probability of a sample containing 9 or more deaths is .0681. This is close enough to .05 to take values of N_1 equal to 9 or more as the desired region of rejection. The actual sample has 7 deaths, which is not in the region of rejection; therefore the hypothesis of a true death rate of 5 per 1,000 cannot be rejected. From the figures above it is seen that a sample could have as high as 8 deaths without causing the hypothesis of a general rate of 5 deaths per 1,000 to be rejected.

An upper bound for the true death rate that will have a probability of .95 of including the true rate may be found as follows: Note first that the sample return is 7 out of 1,000. Then examine Pearson's tables to see for what distribution the sum of the probabilities of 0 to 6 deaths per 1,000 is just .05. This will be found to be the distribution for which $m = 11.8$. Since the sample contains 1,000 cases and $m = Np_1$, it follows that the .95 upper bound for p_1 , given a sample rate of .007, is .0118. It may be said, therefore, that the chances are 95 out of 100 that the range 0-1.18% covers the true percentage.

The mean of a Poisson distribution is m , its variance is also m , its $\beta_1 = \frac{1}{m}$, and its $\beta_2 = \frac{1}{m} + 3$. When N is so large that $Np_1 = m$ is also large, the Poisson distribution is fairly well approximated by the normal curve. This is merely another way of saying that even when p_1 is very small and hence $p_2 - p_1$ is relatively large, if N is big enough, the binomial distribution approaches the normal distribution.¹ In this instance it makes little difference whether the variance of the normal distribution is taken as $m = Np_1$ or as Np_1p_2 since p_2 is very close to 1.

APPENDIX

Derivation of the Poisson Distribution and Its Properties. The Poisson distribution is derived as follows: According to the binomial equation the probability of N_1 cases out of N is²

$$P(N_1) = \frac{N!}{N_1!N_2!} p_1^{N_1} p_2^{N_2} \quad (1)$$

where $p_1 + p_2 = 1$ and $N_1 + N_2 = N$.

¹ See p. 73.

² See p. 43.

By dividing numerator and denominator by $N_2! = (N - N_1)!$ and setting $p_2 = 1 - p_1$ and $N_2 = N - N_1$, this may be written

$$P(N_1) = \frac{N(N-1) \cdots (N - N_1 + 1)}{N_1!} p_1^{N_1} (1 - p_1)^{N - N_1} \quad (2)$$

Suppose that p_1 is very small but that N is large enough so that Np_1 has a value of, say, at least .1. For example, if $p_1 = .003$, let $N = 1,000$, so that the value of Np_1 is 3. Represent Np_1 by that m , so $p_1 = \frac{m}{N}$. If this value of p_1 is put in the foregoing equation, then

$$P(N_1) = \frac{N(N-1) \cdots (N - N_1 + 1)}{N_1!} \left(\frac{m}{N}\right)^{N_1} \left(1 - \frac{m}{N}\right)^{N - N_1} \quad (3)$$

By distributing $\left(\frac{1}{N}\right)^{N_1}$ throughout the first N_1 factors and by separating $\left(1 - \frac{m}{N}\right)^{N - N_1}$ into its two components, Eq. (3) gives

$$P(N_1) = \frac{m^{N_1}}{N_1!} \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{N_1 - 1}{N}\right) \left(1 - \frac{m}{N}\right)^N \left(1 + \frac{m}{N}\right)^{-N_1} \quad (4)$$

But if N is large, as it must be to make Np_1 equal to .1 or more, the value of $\frac{1}{N}$ and $\frac{N_1 - 1}{N}$ will be negligible. The value of $\left(1 - \frac{m}{N}\right)^N$ will be approximately e^{-m} , and the value of $\left(1 + \frac{m}{N}\right)^{-N_1}$ will be practically 1 since $\frac{m}{N}$ is practically negligible. Hence

$$P(N_1) \doteq \frac{m^{N_1}}{N_1!} e^{-m} \quad (5)$$

or it may be written,

$$P(N_1) \doteq \frac{m^{N_1}}{N_1!} \exp [-m] \quad (6)$$

This is known as the Poisson distribution after the man who first developed it.

The mean of a Poisson distribution is m , which may be shown as follows:

First note that

$$\sum_0^N \frac{m^{N_1}}{N_1!} \exp [-m] = 1 \quad (7)$$

since the sum of the total probabilities of a distribution equals 1.

Next note that the mean of the Poisson distribution is by definition

$$\text{Mean } N_1 = \sum_0^N N_1 \frac{m^{N_1}}{N_1!} \exp [-m] \quad (8)$$

Note further that if N_1 in the numerator is canceled against the N_1 factor in $N_1!$ and if m is factored out of m^{N_1} , Eq. (8) becomes

$$\text{Mean } N_1 = m \sum_1^N \frac{m^{N_1-1}}{(N_1-1)!} \exp [-m] \quad (9)$$

But if $N_1 - 1$ is set equal to N'_1 and $N - 1$ to N' , the sum term becomes

$$\sum_0^{N'} \frac{m^{N'_1}}{N'_1!} \exp [-m]$$

which by Eq. (7) equals 1. Hence

$$\text{Mean } N_1 = m \quad (10)$$

Similarly, σ^2 of $N_1 = m$, which may be shown as follows:

By definition,

$$\sigma_{N_1}^2 = \sum N_1^2 \frac{m^{N_1}}{N_1!} \exp [-m] - m^2$$

But $N_1^2 = N_1(N_1 - 1) + N_1$; therefore,

$$\sigma_{N_1}^2 = \sum N_1(N_1 - 1) \frac{m^{N_1}}{N_1!} \exp [-m] + \sum N_1 \frac{m^{N_1}}{N_1!} \exp [-m] - m^2$$

The first term of the right-hand side of this expression equals

$$m^2 \sum \frac{m^{N_1-2}}{(N_1-2)!} \exp [-m]$$

which as above equals m^2 , and the second term equals m as demonstrated in the previous paragraph. Hence

$$\sigma_{N_1}^2 = m^2 + m - m^2$$

or

$$\sigma_{N_1}^2 = m \quad . \quad (11)$$

In the same manner it can be shown that $\mu_3 = m$, and

$$\mu_4 = m + 3m^2$$

Hence $\beta_1 = \frac{1}{m}$ and $\beta_2 = \frac{1}{m} + 3$.

These latter formulas suggest that, if N and hence m is large, the Poisson distribution approaches the normal form; for β_1 then equals approximately 0, and β_2 equals approximately 3. In fact, this approach to normality with increasing size of m may be mathematically demonstrated in a manner essentially the same as that used to demonstrate the approach of the binomial distribution to normality.

CHAPTER X

SAMPLING FROM CONTINUOUS NORMAL POPULATIONS I. VARIOUS SAMPLING DISTRIBUTIONS

The foregoing chapter was concerned with attributes that are qualitative, such as "defective" versus "nondefective," "for" versus "against," "black" versus "white," or attributes that have only discrete numerical values. This and the following chapter will be concerned with attributes that may theoretically vary continuously, such as the heights of individuals, yields of wheat, and the like.¹

The theory of sampling for a continuous variable has been most completely worked out for a normal population, *i.e.*, a population in which the probability, or relative frequency, of a given variable is measured by the normal probability curve. Because of this and because hypotheses of normal populations arise frequently in practical problems, the present chapter will develop the analysis in considerable detail.

The argument will apply strictly to a hypothetical infinite population. It might, for example, apply to the infinite set of electric-light bulbs that could be produced by a given process if that process were to be used indefinitely without modification. It might also apply to the infinite set of crops of a given variety that might be yielded by repeated farming of a given type of land with a given type of treatment. Again it might apply to the infinite set of white adult males living now and in the future. In fact, it might apply to the results of many types of repetitive or continuous physical, biological, or social processes. The argument will also be applicable without serious error to a large finite population that is distributed approximately in a normal manner, such as the heights of existing white adult males. In this instance the population will be presumed to be so large relative to the size of the sample that the withdrawal of the sample does not materially change the distribution of probabili-

¹ Cf. pp. 1-2.

ties in the population. Thus, if the probability of a case lying between 1.2 and 1.3 is .09, say, before the sampling is begun, it will be assumed to remain equal to .09 throughout the whole of the sampling process.

The analysis will assume that the process of sampling is a random one. This implies, as previously, that the results or rather the distribution of the results of repeated sampling from the given population can be predicted with reasonable accuracy by the use of an appropriate mathematical model. The principal task of the analysis will be to develop such a model.

The problem to be considered will be this: A random sample is obtained from a given normal population. Being normal the population may be specified by the mean and standard deviation; for the β_1 of all normal populations is 0, and the β_2 is equal to 3. Hence the problem will be to make inferences about the mean and standard deviation of the population from knowledge of the mean and standard deviation of the sample.

To solve this and similar problems, it will be necessary to consider what would happen under repeated random sampling from the given population. For this purpose, such statistics as the mean and variance are selected, and a study is made of how these statistics vary from sample to sample. Such a study seeks in particular to estimate the relative frequencies with which the selected statistics will assume various values among the infinite set of samples of given size that might be drawn from the given population (with replacements of the samples if the population is finite). That is, it seeks to derive the sampling distributions of the selected statistics.

To derive these sampling distributions by actually drawing a large number of random samples from the given population would be a tedious if not an impossible task in almost all cases. Fortunately, the derivation may be undertaken by a theoretical process similar to that by which the sampling distribution of percentages was derived in the preceding chapters. The steps in the theoretical argument are to find a mathematical model that appears to reproduce the conditions of sampling and then to derive the distributions of the selected statistics for this model by means of the probability calculus.

On the assumption that the process of sampling is a random one, it is argued that the actual distributions of statistics among

a large set of samples from the given population will tend to conform to the theoretical distributions worked out for the given mathematical model. The basis for this, it will be recalled,¹ is intuition and experience with respect to mass phenomena or what has been called the "law of large numbers." The task of the present chapter is to set up the appropriate mathematical model and to derive the theoretical sampling distributions for various statistics.

DISTRIBUTION OF SAMPLES OF 2 FROM A NORMAL POPULATION

For simplicity suppose that a sample contains only two cases, that is, $N = 2$. Of course, in practical work, samples rarely contain such a small number of cases. Much is to be gained in the theoretical analysis, however, by taking samples of 2. For the essentials of the argument are the same for these samples of 2 as they are for large samples, but the argument is much simpler and easier to comprehend. If the argument for samples of 2 is clearly understood, the generalization for samples of larger size is not very difficult.

In accordance with the assumption noted above, the population is assumed to be so large that the withdrawal of the sample does not materially change the distribution of probabilities in the population. Hence the withdrawal of two cases from a single population will be essentially the same as withdrawing one case from one population and another case from another population identical in all respects to the first. Similarly, repeated withdrawals of samples of 2 from a single population (with replacements of samples if the population is finite) will be essentially the same as repeated withdrawals of samples of 1 each from two identical populations. This suggests that the sampling distributions of means, variances, and other statistics of samples of 2 from a normal population may be derived from the following mathematical model:

The model is an arithmetical model instead of an algebraic one; this is intended to facilitate the exposition for the reader who does not have a ready knowledge of the calculus. However, algebraic equations will be given for all results obtained; and, for those who have the mathematical training, the translation of

¹ Cf. SMITH, J. G., and A. J. DUNCAN, *Elementary Statistics and Applications*, pp. 239-241. See also pp. 27-28.

the numerical argument into its algebraic counterpart is given in the appendix of this chapter.

The model supposes two populations identical in all respects; each population has 100,000 cases, which, when grouped in class intervals of 2, have relative frequencies of a normal frequency distribution. These relative frequencies are given in Table 25. The mean of each population is 100, and its standard deviation is 10. Now suppose that each case of population I, like that shown

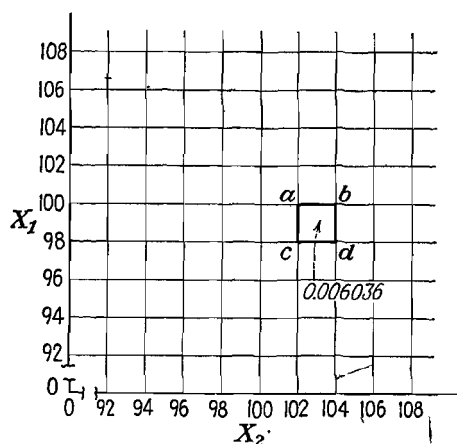


FIG. 63.—Probability of a joint sample $X_2 = 102-104$ and $X_1 = 98-100$.

in Table 25, is combined without restriction with each case of population II, also like that shown in Table 25, and suppose that the pairs of values so obtained are checked off on a scatter diagram such as that illustrated in Fig. 63. For example, if a case from population I (call it X_1) belongs to the interval 98– and a case from population II (call it X_2) belongs to the interval 102–, then this pair of cases (this sample of 2) will be checked off as belonging to the cell $abcd$ of Fig. 63.

If such pairs of cases, or samples of 2, are formed without restriction, as is supposed, then according to the multiplication theorem for independent probabilities, the probability (relative frequency) of a pair of cases belonging to any one cell will be equal to the probability (relative frequency) of a case belonging to the interval from which the first member of the pair was selected, times the probability (relative frequency) of a case belonging to the interval from which the second member of the pair was

TABLE 25.—DISTRIBUTION OF PROBABILITIES FOR A NORMAL POPULATION
($\bar{X} = 100$ and $\sigma = 10$)

Lower Limits of Class Intervals	Probabilities ¹ (Frequencies Relative to the Total of 100,000)
58—	.00002
60—	.00004
62—	.00009
64—	.00018
66—	.00035
68—	.00066
70—	.00121
72—	.00210
74—	.00354
76—	.00570
78—	.00885
80—	.01318
82—	.01887
84—	.02596
86—	.03431
88—	.04359
90—	.05320
92—	.06239
94—	.07033
96—	.07616
98—	.07926
100—	.07926
102—	.07616
104—	.07033
106—	.06239
108—	.05320
110—	.04359
112—	.03431
114—	.02596
116—	.01887
118—	.01318
120—	.00885
122—	.00570
124—	.00354
126—	.00210
128—	.00121
130—	.00066
132—	.00035
134—	.00018
136—	.00009
138—	.00004
140—	.00002

¹ These probabilities are derived by successive subtraction of the .2 intervals in the five-place normal probability area table in F. C. Mills and D. H. Davenport, *A Manual of Problems and Tables in Statistics*, pp. 199–203.

selected. For example, if the probability of a case belonging to the interval 102- in population II is .07616 and the probability of a case belonging to the interval 98- in population I is .07926, then the probability of a pair of cases belonging to the cell 98- for X_1 and 102- for X_2 , that is, cell $abcd$ of Fig. 63, is

$$(.07926)(.07616) = .006036.$$

The probability of any pair of cases among the set of all possible pairs of cases may thus be calculated by simple multiplication of two elementary probabilities. This has been done for the more probable pairs of cases, and the results are presented in Fig. 64.

It may appropriately be argued that this distribution of pairs of cases is a good prediction of the results that would be obtained by drawing one item at random from population I and another at random from population II, recording the results, and then replacing the cases and repeating the process. For if the process of selection is a random one, each case in each population will be drawn just about as often as every other case; and if the selection of the first case does not influence the selection of the second, then no particular pair of cases will tend to occur any more frequently than any other pair of cases. As a consequence, the relative frequencies of various types of samples will, in repeated sampling, be approximately the same as the relative frequencies of various types of combinations among the set of all possible combinations of one item each from the two populations.

According to the argument presented above, the latter will also be a good approximation to the relative frequencies of various types of samples of 2 from a single population. For in this case, the original population can be viewed as population I, and the population remaining after the first case has been drawn can be viewed as population II, the form of the latter being practically the same as that of population I. Relative frequencies in Fig. 64 therefore constitute a good prediction of the distribution of a very large number of samples of 2 from a normal population whose mean is 100 and whose standard deviation is 10.

Properties of the Distribution. *Its Circular Symmetry.* The most striking feature of the set of samples represented in Fig. 64 is the symmetrical distribution of the samples. They tend to

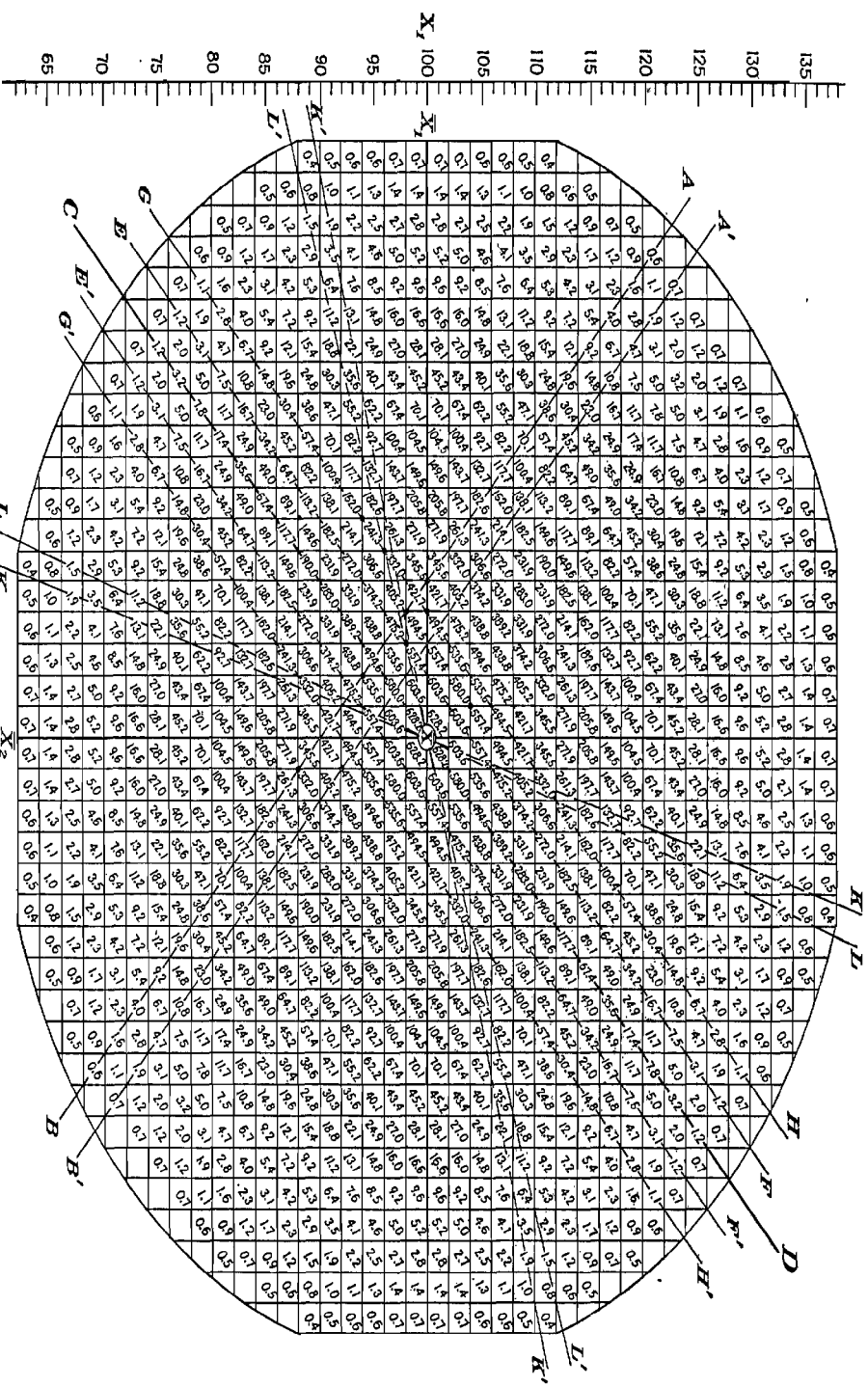


FIG. 64.—Probability distribution of all samples ($N = 2$) from a normal population with arithmetic mean of 100 and standard deviation of 10. Probabilities in 100,000ths.

cluster most closely around the point whose coordinates are both equal to the mean value, 100, and to thin out evenly in all directions from that point. In fact, it will be noted that cells lying equally distant from this central point have approximately equal probabilities.

If the cells had been made smaller, this circular symmetry would have been more evident. Careful study also shows that the probabilities in each row and each column tend to conform to a normal frequency distribution with the same mean and same standard deviation as the original population.

Figure 64 is typical of a large set of samples from a normal population, whatever the size of the sample. In every case the various samples cluster most closely around the point whose coordinates are all equal to the mean of the population, and in every case the distribution of samples around this point conforms to a symmetrically circular (or spherical) pattern.

Geometrical Measurement of the Mean of a Sample. Figure 64 designates a sample by indicating the interval to which case I belongs and the interval to which case II belongs. It may also be used to find the interval to which the mean of any sample belongs. This follows from the geometrical properties of the figure.

By definition the mean of any sample of two cases is

$$\bar{X} = \frac{X_1 + X_2}{2} = \frac{X_1}{2} + \frac{X_2}{2}.$$

For a given value of $\bar{X}$, this is the equation of a line in Fig. 64 that runs through the point $X_2 = 2\bar{X}$ on the X_2 -axis and the point $X_1 = 2\bar{X}$ on the X_1 -axis. All X_1X_2 sample combinations that lie on this line have the given mean value. For example, in Fig. 64 any sample point lying on line AB , such as points (100,90), (98,92), (94,96), (88,102), and (102,88), has a mean of 95. Coordinates are given in order X_1X_2 . In general, the plane of Fig. 64 may be covered with a set of parallel lines, such as line AB , any one of which is the locus of all sample combinations having the same mean. Geometrically, the slope of all these parallel lines is such that the increase in X_1 is matched by the decrease in X_2 , or vice versa, so that the mean of the two remains the same.

The parallel lines that could be drawn perpendicular to lines like AB (for example, like GH) represent samples such that, whenever

X_1 increases by a given quantity, X_2 increases also by the same given quantity. Thus, on line GH , if X_1 is 94, X_2 is 86, whereas, if X_1 is 96, X_2 is 88—as X_1 has increased by 2, X_2 has also increased by 2. Of this latter group of sample combinations one set is of special interest, *viz.*, that for which X_1 and X_2 have the same value, or, algebraically, $X_1 = X_2$. In Fig. 64 all samples lying on line CD , which passes through the origin and bisects the angle between the axes, have values of $X_1 = X_2$. In fact, line CD is the locus of all sample combinations in which $X_1 = X_2$. But in addition, when $X_1 = X_2$, it also follows that $\bar{X} = X_1 = X_2$. As a consequence, every point on line CD has a pair of equal coordinates whose value is the mean of those samples lying on the line perpendicular to CD through this point. This follows from the fact that on a line perpendicular to CD , such as AB , the mean is constant.

As a result of these geometrical properties, the mean of any sample in the diagram can be found immediately as follows: First locate the sample point on the diagram. Then proceed up or down from this point along a line perpendicular to the line CD until the intersection with CD is reached. The mean of the sample is either the X_1 or the X_2 coordinate (for $X_1 = X_2$) of this intersection point. For example, the sample point (84,106) lies on a line that intersects CD in a point whose X_2 coordinate is 95. Hence the mean of this sample is 95.

These properties of Fig. 64 also make it possible to find the interval to which any sample mean belongs. Thus suppose that intervals are marked off on the X_2 - (or X_1 -) axis, the corresponding points on the line CD are found by vertical (or horizontal) projection, and lines perpendicular to CD are drawn through these points. Then to find the interval to which the mean of a sample belongs it is necessary to find merely the pair of lines between which the sample lies. For example, the line perpendicular to CD through the point on CD whose X_2 coordinate is 95 is the line AB , and the line perpendicular to CD through the point whose X_2 coordinate is 97 is the line $A'B'$. Any sample lying between these two lines will have a mean lying between 95 and 97. This geometrical property of Fig. 64 will be found very useful in deriving the distribution of sample means.

Geometrical Measurement of the Variance and Standard Deviation of a Sample. Figure 64 may also be used to find the interval

within which the variance and standard deviation of a sample lies. First note that by definition the variance of a sample of two items is

$$\sigma^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{2}$$

With reference to Fig. 64 this may be interpreted as one-half the square of the distance¹ from the point whose sample values are both equal to $\bar{X}$ to the point whose sample values are X_1 and X_2 . For example, one-half the square of the distance from the point (102,88) to the point (95,95) is the variance of the sample 102,88. Likewise, one-half the square of the distance from the point (108,82) to the point (95,95) is the variance of the sample 108,82; a sample that has the same mean as 102,88 but a different variance. Similarly, one-half the square of the distance from the point (102,92) to the point (97,97) is the variance of the sample 102,92, a sample whose mean and variance both are different from those of the sample 102,88.

It follows from this that all sample points equally distant from the points representing their means have the same variances. Since the line CD is the locus of the points representing the means of the various samples,² it may be concluded that all sample points equally distant from the line CD (in a perpendicular direction) have the same variances. Thus, if intervals are marked off on some line perpendicular to CD and if lines are drawn parallel to CD through the end points of these intervals, the interval to which the variance of any sample belongs may be found by merely noting the pair of lines between which it lies. For example, the line EF is $2\sqrt{2}$ units distant from the line CD .³ All samples on this line therefore have variances equal to $\frac{1}{2}(2\sqrt{2})^2 = 4$. The line GH is $4\sqrt{2}$ units distant from line CD , and all samples on this line have variances equal to 16. The sample 98,104 lies between these two lines and hence has a variance lying between 4 and 16.

Since the standard deviation of a sample is equal to the square root of the variance, the above method of finding the

¹ It will be recalled that the distance between points (X_1, Y_1) and (X_2, Y_2) is

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}.$$

² See pp. 225-226.

³ Each side of a cell is equal to 2 units; therefore, the diagonal is equal to $2\sqrt{2}$ units.

variance of a sample also automatically gives its standard deviation. Thus the lines EF and GH include samples whose standard deviations lie between 2 and 4.

This geometrical method of finding the intervals to which the variance and standard deviation of a sample belong is very helpful in deriving the sampling distributions of these two statistics.

Geometric Measurement of the Sample Statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$.

As will be pointed out later, when the standard deviation of a

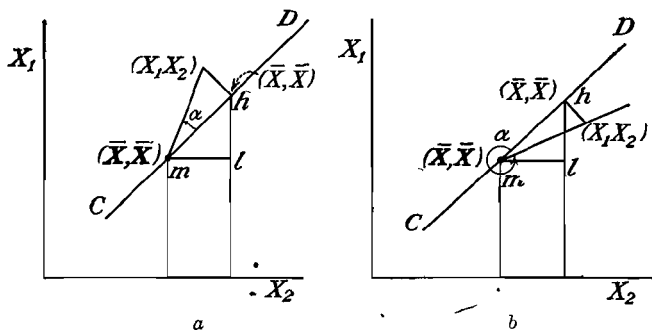


FIG. 65.

population is not known and the sample is small, the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ is the best to employ in testing hypotheses and determining confidence limits for the mean of the population. In this statistic, $\bar{X}$ represents the mean of the sample, $\bar{X}$ is the hypothetical value for the mean of the population, and $\bar{\sigma}$ is the maximum-likelihood estimate of the standard deviation of the population. The relationship between the value of $\bar{\sigma}$ and the standard deviation of the sample is $\bar{\sigma} = \sigma \sqrt{\frac{N}{N-1}}$. The derivation of this maximum-likelihood value is given in the next chapter.

In Fig. 64, the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ may be interpreted geometrically as follows: First, in Fig. 65a and 65b, note that $\sqrt{N}(\bar{X} - \bar{X})$, which for samples of 2 becomes $\sqrt{2}(\bar{X} - \bar{X})$, is equal to the distance from the point h , whose coordinates are both equal to $\bar{X}$, to the point m , whose coordinates are both

equal to $\bar{X}$.^{*} Again note that, for samples of 2, $\sigma = \sigma \sqrt{2}$ and that this latter is the perpendicular distance from the sample point (X_1, X_2) to the mean point $(\bar{X}, \bar{X})$.¹ Hence, if the sample point

lies above the line CD , as in Fig. 65a, the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$

is but the cotangent of the angle¹ that the line connecting that point with the central point $(\bar{X}, \bar{X})$ makes with CD . If the sample

point lies below CD , as in Fig. 65b, then the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$

equals minus the cotangent of the angle that the line connecting the sample point with the central point $(\bar{X}, \bar{X})$ makes with CD .

If lines are thus drawn through the central point $(\bar{X}, \bar{X})$ so that the cotangents of the angles they make with CD form regular intervals of, say .2, then to find the interval within which the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ lies for any sample it is neces-

sary merely to note between what pair of lines the sample point lies. For example, the line LL in Fig. 64 makes an angle with CD , the cotangent of which is 2, and the line KK makes an angle, the cotangent of which is 1.8. Hence, the sample 120,106, which lies between these two lines and above CD , has a value for the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ that lies between 1.8 and 2 (its exact

value is 1.857). Similarly, the lines $L'L'$ and $K'K'$ form angles with CD of which the cotangents are -2 and -1.8 . Hence the sample 120,106, which lies between these lines and below CD , has a value for $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ that lies between 1.8 and 2.

SAMPLING DISTRIBUTION OF THE MEAN

The foregoing properties of the distribution of samples of 2 from a normal population offer a ready means of obtaining the sampling distributions of various sample statistics. Consider first the sampling distribution of the mean.

By definition, the sampling distribution of sample means is a description of the relative frequencies with which samples

^{*} CD makes a 45-degree angle with the X_1 -axis, since it is the locus of all points whose two coordinates are identical. Hence in the right triangle hlm the distance $mh = \sqrt{2}$ times the distance ml .

¹ All angles are viewed as being measured counterclockwise from CD .

assume various mean values. In any concrete problem, such as is involved here, it will consist of a list of intervals together with the relative frequencies of samples whose means fall in those intervals. As pointed out above, it is possible from Fig. 64 to find the interval within which the mean of any sample of 2 lies. Hence to find the sampling distribution of the mean of two cases it is necessary only to lay off a given range of intervals for the mean and then determine from Fig. 64 the relative frequencies of samples lying in the various intervals. The exact procedure may be illustrated by the following example:

Suppose it is desired to find the relative frequency or probability of a mean lying between 95 and 97 (*i.e.*, from 95 up to but not including 97). To do this, first proceed from the points 95 and 97 on the X_2 -axis to the points vertically above on the CD line. Draw lines through these points perpendicular to CD . These are the lines AB and $A'B'$ of Fig. 64. The relative frequency, or probability, of samples having mean values lying between 95 and 97 is equal to the relative frequency, or probability, of samples lying between the lines AB and $A'B'$. This is equal to the probabilities of all cells completely included between these two lines; *i.e.*, it is equal to

$$\frac{535.6}{100,000} + \frac{494.5}{100,000} + \frac{421.7}{100,000} + \dots,$$

plus the prorated share of the probability of those cells only partly included between these two lines, that is,

$$\frac{1}{2} \cdot \frac{557.4}{100,000} + \frac{1}{2} \cdot \frac{475.2}{100,000} + \frac{1}{2} \cdot \frac{421.7}{100,000} + \frac{1}{2} \cdot \frac{494.5}{100,000} + \dots$$

The grand total is $\frac{9,570.8}{100,000}$, and this is accordingly the probability of a sample having a mean lying between 95 and 97.

When the procedure just described is applied to finding the probabilities of sample means lying within each of the class intervals 75-, 77-, etc., the results are those shown in Table 26. From this table and from the accompanying chart it can be seen that the mean of this distribution is the same as that of the mean of the population and that its variance is less. As a matter of fact, numerical calculation shows that the variance of this sampling distribution of the mean of two items is one-half the

variance of the population. Furthermore, the figure suggests that the sampling distribution of the mean has the form of a normal frequency curve.

These indications of the numerical analysis are verified by algebraic analysis.¹ This shows that in general the sampling

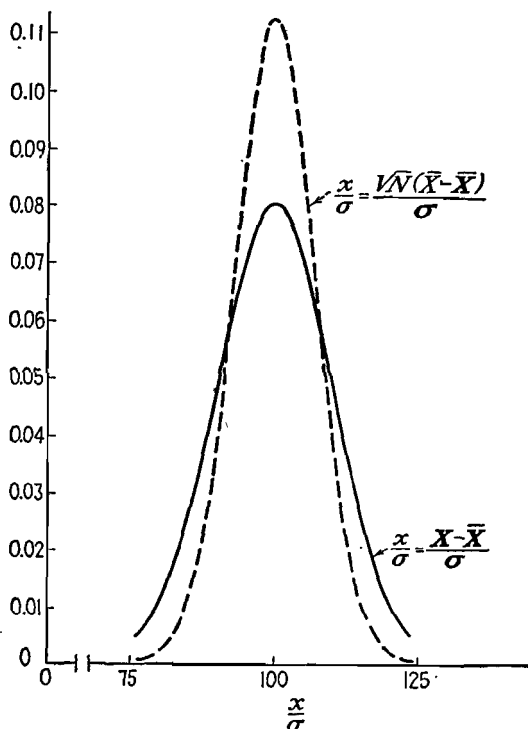


FIG. 66.—Comparison of a normal population with the sampling distribution of means ($N = 2$). Data from Table 26.

distribution of the mean is normal in form, that its mean is the mean of the population from which the samples have been drawn, and that its variance is $1/N$ th the variance of the population. The algebraic equation for the sampling distribution of the mean is thus

$$dP(\bar{X}) = \frac{1}{\sigma_{\bar{X}} \sqrt{2\pi}} \exp \left[-\frac{(\bar{X} - \bar{X})^2}{2\sigma_{\bar{X}}^2} \right] d\bar{X} \quad (1)$$

where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$

¹ See Appendix to this chapter (p. 251).

TABLE 26.—COMPARISON OF THE SAMPLING DISTRIBUTION OF THE MEAN
WITH THE DISTRIBUTION OF THE POPULATION

Lower limits of class intervals	Probabilities of	
	Sample mean	Population
(1)	(2)	(3)
75—	.00036	.00451
77—	.00093	.00714
79—	.00215	.01086
81—	.00456	.01585
83—	.00895	.02224
85—	.01619	.02999
87—	.02705	.03887
89—	.04177	.04839
91—	.05960	.05790
93—	.07856	.06658
95—	.09571	.07355
97—	.10773	.07808
99—	.11206	.07968
101—	.10773	.07808
103—	.09571	.07335
105—	.07856	.06658
107—	.05960	.05790
109—	.04177	.04839
111—	.02705	.03887
113—	.01619	.02999
115—	.00895	.02224
117—	.00456	.01585
119—	.00215	.01086
121—	.00093	.00714
123—	.00036	.00451

SAMPLING DISTRIBUTIONS OF THE VARIANCE AND STANDARD DEVIATION

The sampling distribution of sample variances gives the relative frequencies, or probabilities, of samples having various values for their variances. In numerical form, it lists certain intervals and records the relative frequencies of samples whose variances fall in these intervals.

The geometrical properties of Fig. 64 permit as ready a derivation of the distribution of the variances of samples of 2, as it did of the distribution of the means of samples of 2. For, as

pointed out above,¹ all sample points equally distant from the line CD (in a perpendicular direction) have the same variances. Thus a set of lines can be drawn parallel to the line CD and by computing the relative frequency, or probability, of the samples falling between various pairs of lines the relative frequency, or probability, of a sample having a given variance can be determined. The procedure may be illustrated by an example.

The sample 98,94 has a mean of 96 and a variance of 4. The point representing this sample in Fig. 64 lies on the line EF running parallel to line CD . According to the previous paragraph, all other sample points lying on the line EF have the same variance. The sample 100,92 has a mean of 96 and a variance of 16. The point representing this sample in Fig. 64 lies on the line GH , also running parallel to CD . All sample points lying on line GH therefore have a variance of 16. It also follows that all sample points lying between lines EF and GH have variances lying between 4 and 16. Furthermore, since all points on the line $E'F'$ of Fig. 64 are the same distance from line CD as those on line EF , they also have sample variances of 4, and likewise all points on the line $G'H'$ have sample variances of 16. Hence the total set of samples whose variances lie between 4 and 16 consists of all the samples lying between the lines EF and GH on the one hand and the lines $E'F'$ and $G'H'$ on the other hand. The relative frequency, or probability, of a sample having a variance between 4 and 16 is therefore the relative frequency, or probability, of a sample lying between EF and GH , plus the relative frequency, or probability, of a sample lying between $E'F'$ and $G'H'$. As in the case of the means, this total probability may be computed directly from Fig. 64 by adding the probabilities of all the cells and sections of cells included between these two pairs of lines. The result in this particular case is found to be 0.20510. It is suggested that the reader check this by carrying out the calculations himself.

By repeated applications of the foregoing procedure, the probabilities of samples having variances lying between other class limits may readily be computed. The results are summarized in Table 27 and pictured graphically in Fig. 67. This represents the sampling distribution of the variances of samples of 2. Owing to the small size of the sample, the shape of the

¹ See pp. 227-228.

distribution is unusual. A more common shape is shown by the distribution of samples of 11, which also is illustrated in Fig. 67.

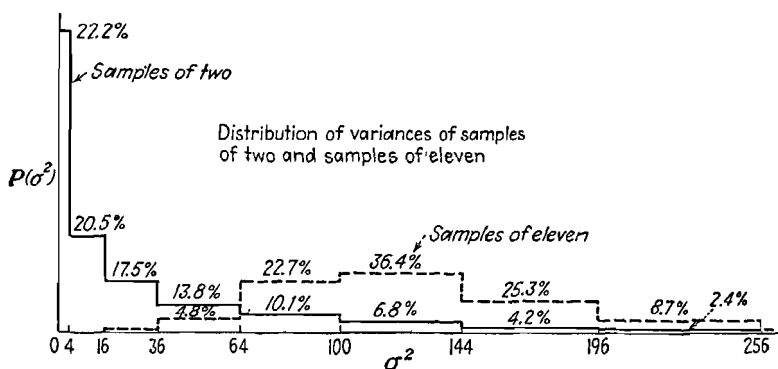


FIG. 67.—Effect of size of sample on the sampling distribution of variances.

Algebraic analysis¹ shows that the sampling distribution of variances of samples of N has the following equation:

$$dP(\sigma^2) = \frac{N^{\frac{N-1}{2}} (\sigma^2)^{\frac{N-3}{2}}}{2^{\frac{N-1}{2}} \left(\frac{N-3}{2}\right)! \sigma^{N-1}} \exp \left[-\frac{N\sigma^2}{2\sigma^2} \right] d\sigma^2 \quad (2)$$

For N greater than 2, this represents a curve that starts at 0, rises to a peak at $\sigma^2 = \frac{N-3}{N} \sigma^2$, and approaches 0 again as σ^2 goes to infinity. For small values of N the curve is thus very skewed. For large values of N , however, the curve is more symmetrical and almost normal in form, the mean of the curve being approximately the variance of the population σ^2 and its standard deviation being $\sigma^2 \sqrt{\frac{2}{N}}$.

For small samples, the sampling distribution of the variance is thus nonnormal, and use cannot be made of the normal probability tables in testing hypotheses and determining confidence limits. It can be shown, however, that if the unit of measurement is taken as $\frac{1}{N}$ times the variance of the population (that is, $\frac{\sigma^2}{N}$), then the distribution takes on the form of the χ^2

¹ See Appendix to this chapter (p. 263).

TABLE 27.—SAMPLING DISTRIBUTION OF THE VARIANCE AND STANDARD DEVIATION OF SAMPLES OF 2 FROM A NORMAL POPULATION WHOSE
 $\sigma = 10$
 $(N = 2)$

Values of		Probability
σ	σ^2	
0-	0-	.22193
2-	4-	.20510
4-	16-	.17515
6-	36-	.13820
8-	64-	.10078
10-	100-	.06790
12-	144-	.04228
14-	196-	.02432
16-	256-	.01291
18-	324-	.00632
20-	400-	.00285
22-	484-	.00118
24-	576-	.00042

TABLE 28.—SAMPLING DISTRIBUTION OF VARIANCES AND STANDARD DEVIATIONS OF SAMPLES OF 11 FROM A NORMAL POPULATION WHOSE
 $\sigma = 10$
 $(N = 11)$

Values of		Probability
σ	σ^2	
0-	0-	.00008
2-	4-	.00248
4-	16-	.04847
6-	36-	.22713
8-	64-	.36406
10-	100-	.25270
12-	144-	.08718
14-	196-	.01592
16-	256-	.00095
18-	324-	.00005
		.99902

distribution. More specifically, it can be shown that the quantity $\frac{\sigma^2}{\hat{\sigma}^2/N} = \frac{N\sigma^2}{\hat{\sigma}^2}$ has a sampling distribution that is of the form of the χ^2 distribution with n in the χ^2 equation equal to $N - 1$. Tables of the χ^2 distribution can therefore be used in computing probabilities of various sample variances. This use of the χ^2 tables will be explained more fully below.¹

The sampling distribution of the standard deviation of samples of 2 is found from Table 27 by taking the intervals 0-, 2-, 4-, 6-, 8-, the limits being the square root of the limits adopted for the variance. The result is shown in Fig. 68. The figure also gives the distribution of the standard deviation for samples of 11.

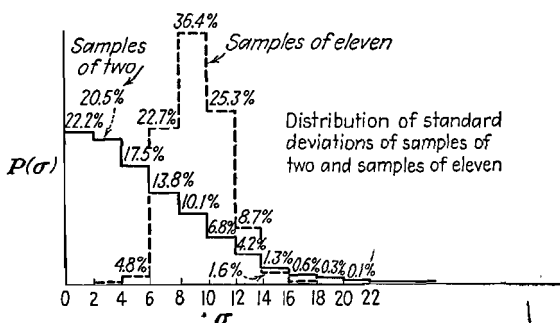


FIG. 68.—Effect of size of sample on sampling distribution of the standard deviation.

The equation for the sampling distribution of the standard deviation is

$$dP(\sigma) = \frac{N^{N-1} \sigma^{N-2}}{2^{\frac{N-3}{2}} \left(\frac{N-3}{2}\right)! \hat{\sigma}^{N-1}} \exp \left[-\frac{N\sigma^2}{\hat{\sigma}^2} \right] d\sigma \quad (3)$$

This equation is not used in practical analysis since it is easier to work with the variance σ^2 and to use tables of the χ^2 distribution as just explained.

SAMPLING DISTRIBUTION OF $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\hat{\sigma}}$

The sampling distribution of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\hat{\sigma}}$ gives the relative frequencies with which samples take on various values of this

¹ See pp. 284-289.

statistic. In numerical cases, it lists intervals of this statistic and gives the relative frequencies of samples belonging to these intervals.

It was pointed out above that if a line is drawn through a sample point connecting it with the central point $(\bar{X}, \bar{X})$ the value of the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ for the sample is equal to the cotangent of the angle that this line makes with CD if the sample point lies above CD or to minus the cotangent of this angle if the sample point lies below CD , all angles to be read counterclockwise.

The relative frequency, or probability, of samples having values of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ lying between certain limits is accordingly the relative frequency with which samples fall between the radii through the point $(\bar{X}, \bar{X})$ whose cotangents are equal to these limits. For example, in Fig. 64, the cotangent of the angle that LL makes with CD is 2.0, and the cotangent of the angle that KK makes with CD is 1.8; also, the cotangents of the angles that $L'L'$ and $K'K'$ make with CD are -2.0 and -1.8 , respectively. Hence, the samples for which $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ lies between 1.8 and 2.0 are the samples included between the lines LL and KK and lying above CD and the samples included between the lines $L'L'$ and $K'K'$ and lying below CD . Similarly, the samples for which $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ lies between -1.8 and -2.0 are the samples included between LL and KK and lying below CD and the samples included between $L'L'$ and $K'K'$ and lying above CD . Since various sample points are distributed with a circular symmetry about the point $(\bar{X}, \bar{X})$,¹ it follows that the relative frequency of samples included between LL and KK and lying above CD is proportional to the angle between these two lines. More specifically, it is equal to the ratio that this angle bears to 360 deg. Since in Fig. 64 angle $L\bar{X}D = 26.56$ deg and angle $K\bar{X}D = 29.05$ deg and hence the difference equals 2.49 deg, it follows that the relative frequency, or probability, of samples included between the lines LL and KK and lying above CD is

¹ See pp. 224-225.

equal to $\frac{2.49}{360} = .0069$. Owing to the symmetry of the probabilities represented by Fig. 64 this is also the probability of a sample included between $L'L'$ and $K'K'$ and lying below CD .

Hence the probability of a sample having a value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ between 1.8 and 2.0 is .0138; similarly, the probability of a sample having a value of that statistic between -1.8 and -2.0 is .0138.

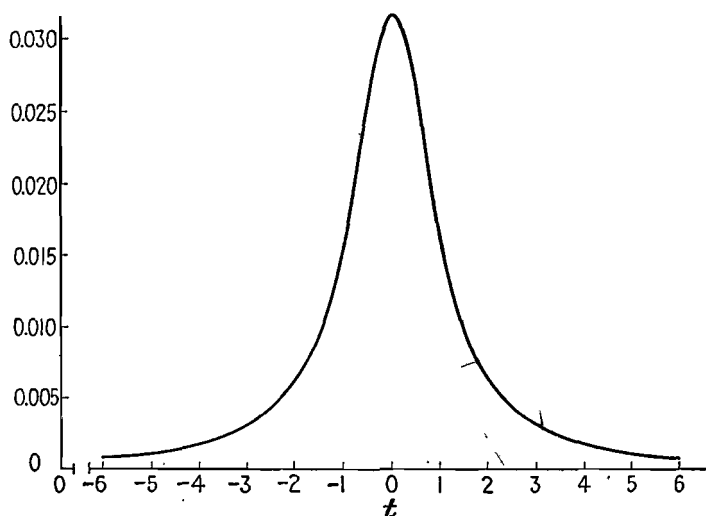


FIG. 69.—A t distribution with $N = n - 1 = 2 - 1 = 1$. Data in Table 29.

The sampling distribution $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ can therefore be readily obtained for samples of 2 as follows: Find the angles for which the cotangents are, say, 0.0, 0.05, 0.15, 0.25, 0.35, etc.; take the successive differences between these angles, and double the results; finally divide these results by 360. This quotient, for each successive interval, will be the probability of a sample having a value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ lying between 0 and 0.05, between 0.05 and 0.15, between 0.15 and 0.25, between 0.25 and 0.35, etc. Owing to the symmetry of the probabilities represented in Fig. 64, the probabilities of negative values of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ are the

TABLE 29.—SAMPLING DISTRIBUTION OF THE STATISTIC $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ FOR
 SAMPLES OF TWO FROM A NORMAL POPULATION
 ($\bar{X} = 100$ and $\sigma = 10$)

(1)	(2)	(3)	(4)	(5)	(6)
Cotangent of angle with CD line	Angle with CD line	Angle inter- vals, e.g., $\angle L\bar{X}K$	Twice angle intervals divided by 360 deg.	Class inter- vals in the statistic	Probabilities
-.05	92.862	5.724	.03180	-.05-	.03180
.00	90.000				
.05	87.138				
.15	81.469	5.669	.03149	.05-	.03149
.25	75.963	5.506	.03059	.15-	.03059
.35	70.710	5.253	.02918	.25-	.02918
.45	65.771	4.939	.02744	.35-	.02744
.55	61.191	4.580	.02544	.45-	.02544
.65	56.976	4.215	.02342	.55-	.02342
.75	53.130	3.846	.02137	.65-	.02137
.85	49.637	3.493	.01940	.75-	.01940
.95	46.470	3.167	.01759	.85-	.01759
1.05	43.603	2.867	.01593	.95-	.01593
1.15	41.010	2.593	.01440	1.05-	.01440
1.25	38.660	2.350	.01306	1.15-	.01306
1.35	36.528	2.132	.01184	1.25-	.01184
1.45	34.592	1.936	.01076	1.35-	.01076
1.55	32.829	1.763	.00979	1.45-	.00979
1.65	31.218	1.611	.00895	1.55-	.00895
1.75	29.743	1.475	.00819	1.65-	.00819
1.85	28.388	1.355	.00753	1.75-	.00753
1.95	27.150	1.238	.00688	1.85-	.00688
2.05	26.000	1.150	.00639	1.95-	.00639
2.15	24.944	1.056	.00587	2.05-	.00587
2.25	23.964	.980	.00544	2.15-	.00544
2.35	23.050	.914	.00508	2.25-	.00508
		.850	.00472	2.35-	.00472

TABLE 29.—SAMPLING DISTRIBUTION OF THE STATISTIC $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ FOR
SAMPLES OF TWO FROM A NORMAL POPULATION.—(Continued)

(1)	(2)	(3)	(4)	(5)	(6)
Cotangent of angle with CD line	Angle with CD line	Angle inter- vals, e.g., $\angle L\bar{X}K$	Twice angle intervals divided by 360 deg.	Class inter- vals in the statistic	Probabilities
2.45	22.200				
		.785	.00436	2.45-	.00436
2.55	21.415	.744	.00413	2.55-	.00413
2.65	20.671	.691	.00384	2.65-	.00384
2.75	19.980	.642	.00357	2.75-	.00357
2.85	19.338	.614	.00341	2.85-	.00341
2.95	18.724	.568	.00316	2.95-	.00316
3.05	18.156	.546	.00303	3.05-	.00303
3.15	17.610	.505	.00280	3.15-	.00280
3.25	17.105	.486	.00270	3.25-	.00270
3.35	16.619	.454	.00252	3.35-	.00252
3.45	16.165	.432	.00240	3.45-	.00240
3.55	15.733	.413	.00229	3.55-	.00229
3.65	15.320	.389	.00216	3.65-	.00216
3.75	14.931	.370	.00206	3.75-	.00206
3.85	14.561	.354	.00197	3.85-	.00197
3.95	14.207	.339	.00188	3.95-	.00188
4.05	13.870	.321	.00178	4.05-	.00178
4.15	13.549	.308	.00171	4.15-	.00171
4.25	13.241	.295	.00164	4.25-	.00164
4.35	12.946	.281	.00156	4.35-	.00156
4.45	12.665	.270	.00150	4.45-	.00150
4.55	12.395	.257	.00143	4.55-	.00143
4.65	12.138	.250	.00139	4.65-	.00139
4.75	11.888				

same as the probabilities of positive values. The sampling distribution of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ is thus symmetrical about zero.

The calculations here indicated are carried out in Table 29 and the results pictured graphically in Fig. 69.

Algebraic analysis shows that, in general, the sampling distribution of the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ is of the form of the t (or Student's) distribution, with the n in the equation equal to $N - 1$. For samples of any size the equation for the t distribution is as follows:¹

$$dP \left[t = \frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma} \right] = \frac{\left(\frac{n-1}{2} \right)!}{\sqrt{n\pi} \left(\frac{n-2}{2} \right)! \left(1 + \frac{t^2}{n} \right)^{\frac{n+1}{2}}} dt \quad (4)$$

where $n = N - 1$.

As might have been inferred from the numerical analysis, the sampling distribution of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ is independent of the hypothetical values assumed for the mean or standard deviation of the population and varies only with N , the size of the sample. For large values of N the curve is approximately normal with a mean of 0 and a variance approximately equal to 1. In these cases the normal probability table can be used in place of the t table.

SAMPLING DISTRIBUTIONS OF OTHER STATISTICS

In estimating the properties of a normal population the sampling distributions of the mean, $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$, and the variance (or standard deviation) are the most important. For a normal distribution is characterized by its mean and standard deviation and these are best estimated by the above sampling distributions.

It is also possible to estimate the mean of the population from the median of a sample, and in the absence of knowledge of the sample mean this would be the next best statistic. As in the case of the mean, the sampling distribution of the median

¹ See Appendix to this chapter (p. 266).

is normal, and its mean is the mean (= median) of the population. The standard error of the median, however, equals $\frac{1.2533}{\sqrt{N}}$ σ , which is about $1\frac{1}{4}$ times as large as the standard error of the mean. It is this greater sampling error that makes the median less "efficient" than the mean in setting up confidence limits for the mean (= median) of the population. For the confidence interval derived from the median will, for a given confidence coefficient, be $1\frac{1}{4}$ times larger than that derived from the mean and will therefore not give as good an estimate of the population mean, or median.

In instances where time and economy of effort are important, it is also possible to estimate the variance (or standard deviation) of a normal population from the sample range.¹ It has been found possible to derive the mean range, the standard error of the range, and the upper and lower .001, .005, .010, .025, .050, and .100 points of the sampling distribution of the range for samples of 2 to 20 cases.² These data are reproduced in Table XIII of the Appendix. The unit for the table is the standard deviation of the population, and the table can thus be used to estimate the population standard deviation from the sample mean. The procedure is discussed on pages 294-296.

The sample statistics $\sqrt{\beta_1}$ and β_2 are important as measures of departure from normality. If the population is normal, the sampling distributions of these statistics tend to normality with increasing size of the sample. Their means are 0 and 3, and their standard deviations are approximately $\sqrt{6/N}$ and $\sqrt{24/N}$, respectively. Hence, if a sample is large, departures from normality may be tested by treating $\frac{\sqrt{\beta_1}}{\sqrt{6/N}}$ and $\frac{\beta_2 - 3}{\sqrt{24/N}}$ as normally distributed variates.

Because of the complexity of the exact formulas for the means and standard deviations of the sampling distributions of $\sqrt{\beta_1}$ and β_2 , R. A. Fisher³ suggests the use of certain "k

¹ The range is the difference between the largest and smallest cases in a sample.

² L. H. C. Tippett, E. S. Pearson, and H. O. Hartley have been mainly responsible for deriving these data. See footnote to Table XIII in the Appendix.

³ *Statistical Methods for Research Workers*, Appendix, Chap. III.

statistics" in place of the moments of a sample distribution and certain "*g* statistics" in place of the sample β statistics. The k statistics¹ are defined as follows:

$$\begin{aligned} k_1 &= \frac{1}{N} \Sigma x \\ k_2 &= \frac{\Sigma x^2}{N-1} \\ k_3 &= \frac{N}{(N-1)(N-2)} \Sigma x^3 \\ k_4 &= \frac{N}{(N-1)(N-2)(N-3)} \left[(N-1) \Sigma x^4 - 3 \frac{(N-1)}{N} (\Sigma x^2)^2 \right] \end{aligned}$$

The *g* statistics bear the same relationship to the k statistics as the β statistics do to the sample moments, *viz.*,

$$g_1^2 = \frac{k_3^2}{k_2^3} \quad g_2 = \frac{k_4}{k_2^2}$$

For large samples the sampling distributions of the *g* statistics tend to be normal in form, with means of zero and standard deviations equal exactly to

$$\begin{aligned} \sigma_{g_1} &= \frac{6N(N-1)}{(N-2)(N-1)(N-3)} \\ \sigma_{g_2} &= \frac{24N(N-1)^2}{(N-3)(N-2)(N-3)(N-5)} \end{aligned}$$

The use of the sampling distribution of β_1 and β_2 or g_1 and g_2 to test departures from normality will be discussed in the next chapter.

For small samples, the sampling distributions of $\sqrt{\beta_1}$ and β_2 have been approximated by "fitting" Pearsonian curves having the same moments as those of the exact sampling distributions. For samples of 25 to 100, the 5 per cent and 10 per cent limits of the sampling distribution of $\sqrt{\beta_1}$ have been determined

¹ The mean values of the k 's are the "cumulants" of the population. Cf. pp. 83-84.

approximately by P. Williams.¹ These are reproduced in Table X of the Appendix.²

It is interesting to compare the limits of this table with those obtained by the assumption that $\sqrt{\beta_1}$ has a normal sampling distribution with a mean of zero and a standard deviation of $\sqrt{6/N}$. For example, for samples of 100, the 5 per cent point given by the table is 0.389, while the assumption of a normal sampling distribution gives 0.403. This suggests that for samples of 100 or more a normal sampling distribution gives fairly good results.

Approximate values for the upper and lower 1 per cent and 5 per cent points of the sampling distribution of β_2 have also been computed³ and are reproduced in Table XI of the Appendix.⁴ By comparison, the 5 per cent points derived from the assumption that β_2 is normally distributed with a mean of 3 and a standard deviation of $\sqrt{24/N}$ are 2.19 and 3.81 for samples of 100 and 2.64 and 3.36 for samples of 500. The table values are 2.35 and 3.77 for samples of 100 and 2.57 and 3.37 for samples of 500, which shows that for samples of 500 or more the assumption of a normal sampling distribution gives results that would appear to be fairly reliable. For samples of 100, the assumption of a normal sampling distribution gives a lower limit that is somewhat below that given by the table.

For practical use, Tables X and XI of the Appendix have been put in chart form.⁵ This permits ready interpolation for values not listed in the tables.

To test departure from normality with respect to kurtosis, tables have also been constructed for the sampling distribution of the statistic

$$a = \frac{\text{A.D. (calculated from the mean)}}{\sigma}$$

Whereas the sampling distribution of β_2 is quite skewed for $N < 200$ and the accuracy of the tabled probability levels for

¹ "Note on the Sampling Distribution of $\sqrt{\beta_1}$, where the Population Is Normal," *Biometrika*, Vol. 27 (1935), pp. 269-271.

² See p. 480.

³ PEARSON, E. S., "A Further Development of Tests for Normality," *Biometrika*, Vol. 22 (1930-1931), pp. 239-249.

⁴ See p. 480.

⁵ GEARY, R. C., and E. S. PEARSON, *Tests of Normality*, pp. 13-15.

β_2 cannot be determined without further investigation, the approximation involved in the calculation of probability levels for a is very close, even when N is as small as 10.* For this reason the use of a in place of β_2 to test existence of kurtosis is recommended. Probability levels for a have been worked out¹ and are given in Table XII of the Appendix. The use of this table will be illustrated in the next chapter.

SUMMARY

The fundamental basis of any sampling analysis consists in a comparison of a sample with the set of all possible samples that may be derived from some hypothetical population. The basis of the comparison will vary from problem to problem. The present chapter, which is concerned with sampling from a normal population, discussed four principal sample measurements that might be used to make this comparison. These are the mean of the sample, the variance and standard deviation of the sample, and the sample statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$.

When all possible samples from a normal population are considered with reference to their mean values, it is found that the set of sample means forms a normal frequency distribution the mean of which is the mean of the population and the variance of which is the variance of the population divided by N . This distribution of the sample means is called the "sampling distribution of the mean." It is to be noted that its shape and position depend on the values of the population mean and variance.

When the set of all possible samples is described in terms of the variances of the samples, it is found that the set of sample variances make up a skewed frequency distribution, which, if σ^2/N is taken as the unit of measurement, is of the form of the χ^2 distribution, with n in the χ^2 formula equal to $N - 1$. This is the "sampling distribution of the variance." The shape and position of this distribution depends only on the value of the population variance and is independent of the value of the population mean. If the sample is large, say 30 or more, the

* See *ibid.*, p. 2.

¹ GEARY, R. C., "Moments of the Ratio of the Mean Deviation to the Standard Deviation for Normal Samples," *Biometrika*, Vol. 28 (1936), pp. 295-307.

sampling distribution of the variance is almost normal in form, the mean of the distribution being approximately equal to the variance of the population and the variance of the distribution being equal to the variance of the population times $2/N$.

When the set of all possible samples is described in terms of the sample values of the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$, it is found that these sample values make up a frequency distribution which is of the form of the t distribution, with n equal to $N - 1$. This is the "sampling distribution" of the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$. Since the form and shape of the t distribution depend only on the value of n (here equal to $N - 1$), it follows that the sampling distribution of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ is independent of the values of the population mean and variance. For large samples, say 30 or more, the t distribution is almost normal, with a mean of zero and a standard deviation of approximately unity.

Attention was also called to the sampling distributions of such statistics as the median and the range, β_1 and β_2 , g_1 and g_2 , and A.D./ σ .

APPENDIX

It is the purpose of this appendix to derive the sampling distributions of the mean, variance, and $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ of a sample of 2 cases and a sample of N cases from a normal population. The argument will be the same as in the text but will be algebraic and hence general, instead of arithmetical and particular.

SPECIAL CASE $N = 2$

Distribution of All Possible Samples of Two from a Normal Population. *Derivation.* Let the first case in a sample be represented by X_1 and the second by X_2 . Since the population is normal, the probability of X_1 lying in the interval X_1 to $X_1 + dX_1$ is

$$dP(X_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(X_1 - \bar{X})^2}{2\sigma^2}\right] dX_1 \quad (1)$$

and the probability of X_2 lying in the interval X_2 to $X_2 + dX_2$ is

$$dP(X_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(X_2 - \bar{X})^2}{2\sigma^2} \right] dX_2 \quad (2)$$

If samples of two are drawn at random from the population, the law of large numbers suggests that the relative frequency with which samples will have one case in the interval X_1 to $X_1 + dX_1$ and the other in the interval X_2 to $X_2 + dX_2$ may be predicted by the joint probability of X_1 and X_2 , that is, by the product of Eqs. (1) and (2); thus

$$dP(X_1, X_2) = \frac{1}{\sigma^2 2\pi} \exp \left[-\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{2\sigma^2} \right] dX_1 dX_2 \quad (3)$$

Equation (3) is the mathematical formula for the distribution of all possible samples of two from a normal population. It is the model that any large number of actual random samples of two will tend to approximate.

Geometrical Properties. If values of X_1 and X_2 for any sample are taken as the coordinates of a point in a plane, then the set of all possible samples will be represented by a cluster of points in this plane; and their distribution over the plane will be given by equation (3). The factor $dX_1 dX_2$ of this equation will represent the area of the plane containing the specified sample values and

$$\frac{1}{\sigma^2 2\pi} \exp \left[-\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{2\sigma^2} \right]$$

will represent the "density" of sample points in this area. Since the density factor increases as X_1 and X_2 approach $\bar{X}$, it follows that the greatest concentration of sample points is in the immediate neighborhood of the point $\bar{X}, \bar{X}$. Furthermore, since

$$(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$$

measures the square of the distance from $\bar{X}, \bar{X}$ it follows that all sections of the plane equally distant from $\bar{X}, \bar{X}$ will have the same density of points and that the density decreases as the distance from $\bar{X}, \bar{X}$ increases. The distribution of all samples of two thus has a circular symmetry about the central point, $\bar{X}, \bar{X}$.

The point $\bar{X}, \bar{X}$, it will be noted, lies on the line through the origin that bisects the angle between the axes. This line (line CD of Fig. 64, or Figs. 65a and 65b, if projected to pass through the point where $X_1 = 0$ and $X_2 = 0$) is the locus of all samples that

have two equal values and thus has the equation $X_1 = X_2$ or $X_1 - X_2 = 0$.

Now the mean of any sample ($N = 2$) is $(X_1 + X_2)/2 = \bar{X}$. But this is the equation of a line that cuts the X_1 axis at $2\bar{X}$ and the X_2 axis at $2\bar{X}$, and is perpendicular to the line CD , cutting that line at $\bar{X}, \bar{X}$. Such a line is line AB of Fig. 64. All points on line AB have the same mean. It then follows that the mean of any sample can be found by finding where the line through it, perpendicular to CD , cuts CD . Hence variation in mean values from sample to sample can be represented by variation along CD .

Also, since the variance of a sample X_1, X_2 is by definition

$$\sigma^2 = [(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2]/2$$

and since $(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$ represents the square of the distance from the sample point X_1, X_2 to the mean point $\bar{X}, \bar{X}$, it follows that all sample points equidistant from their mean points have the same variances. But all mean points $\bar{X}, \bar{X}$ lie on the line CD ; hence all points equidistant from line CD have the same variance.¹ Loci of equal variances are accordingly lines, like lines EF and $E'F'$ of Fig. 64, that are parallel to line CD . Variation in sample variances is thus measured by variation perpendicular to these lines.

Distribution of All Possible Samples in Terms of Their Means and Variances. From the foregoing analysis, it follows that the plane of Fig. 64 can be divided into cells indicating variation in $\bar{X}$ and in σ^2 or σ , in lieu of cells indicating variation in X_1 and X_2 . This might be done as follows:

Select any dX_1 interval on the X_1 axis and draw perpendicular lines through each end of this interval; likewise select any dX_2 interval on the X_2 axis (equal to dX_1) and draw lines through the end points perpendicular to the X_2 axis. Let these two pairs of parallel lines mark out the cell $abcd$ of Fig. 70. The top and bottom of this cell will be dX_2 , and the sides dX_1 . This cell is like cell $abcd$ in Fig. 63 on page 222. Next draw lines through a and c perpendicular to CD . These will mark off an interval on CD that will equal $d\bar{X} \sqrt{2}$; since $\bar{X} = (X_1 + X_2)/2$,

$$d\bar{X} = (dX_1 + dX_2)/2,$$

and because $dX_1 = dX_2$, $d\bar{X} = dX_1$ or dX_2 . Hence, in Fig. 70,

¹ See p. 247.

$fg = ab \sqrt{2} = dX_2 \sqrt{2} = d\bar{X} \sqrt{2}$. Finally, draw lines through b and d parallel to CD and call the perpendicular distance between them $d\sigma \sqrt{2}$; the variance, it will be recalled from page 227, equals one-half the square of the distance from CD .

From the cell $abcd$ there has thus been derived the new cell $efgh$ of size $d\bar{X} \sqrt{2} d\sigma \sqrt{2} = 2d\bar{X}d\sigma$. When the whole of the

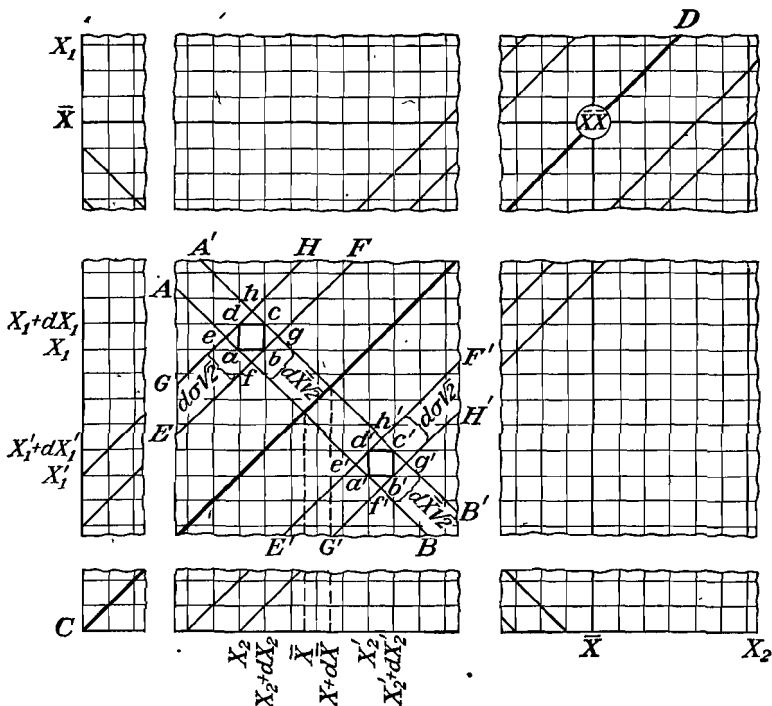


FIG. 70.

plane has been cut up into cells like $efgh$, the plane will become a grid based on $\bar{X}$ and σ intervals instead of X_1 and X_2 intervals.

In Eq. (3) the factor that measures the variation in density of sample points from one part of the plane to the other is the quantity $\frac{1}{\sigma^2 2\pi} \exp \left[-\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{2\sigma^2} \right]$. But the numerator of the exponential can be evaluated as follows:¹

¹ For $N\sigma^2 = \Sigma d^2 - NC^2$, in which $d = X_i - \bar{X}$ and $C = \bar{X} - \bar{X}$. This is merely a version of the "short formula" for finding the standard deviation. In this application, it is to be remembered, $N = 2$.

$$\begin{aligned}(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 &= (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + 2(\bar{X} - \bar{X})^2 \\ &= 2\sigma^2 + 2(\bar{X} - \bar{X})^2\end{aligned}$$

This indicates that it is possible also to measure the variation in density of sample points in terms of $\bar{X}$ and σ .

This now permits the complete expression of the distribution of all samples in terms of $\bar{X}$ and σ . For from the foregoing it follows that the probability of a sample having a mean lying between $\bar{X}$ and $\bar{X} + d\bar{X}$ and a σ lying between σ and $\sigma + d\sigma$ is equal to the density factor for the areas containing the specified values of $\bar{X}$ and σ times the size of the area. The density factor

is $\frac{1}{\sigma^2 \sqrt{2\pi}} \exp \left[-\frac{2\sigma^2 + 2(\bar{X} - \bar{X})^2}{2\sigma^2} \right]$. The area factor is double

the area of square $efgh$ in Fig. 70; i.e., double $2d\bar{X}d\sigma$ or $4d\bar{X}d\sigma$. The doubling of the area of $efgh$ results from the fact that there is another square, $e'f'g'h'$, of the same size on the opposite side of line CD that also contains samples having the same mean and standard deviation as samples in square $efgh$. Since probability equals density factor times area, it follows that the probability of a sample having an $\bar{X}$ between $\bar{X}$ and $\bar{X} + d\bar{X}$ and a σ between σ and $\sigma + d\sigma$ is equal to

$$\begin{aligned}dP(\bar{X}, \sigma) &= \frac{1}{\sigma^2 \sqrt{2\pi}} \exp \left[-\frac{2\sigma^2 + 2(\bar{X} - \bar{X})^2}{2\sigma^2} \right] 4d\bar{X}d\sigma \\ &= \frac{1}{\sqrt{2\pi} \sigma / \sqrt{2}} \exp \left[-\frac{2(\bar{X} - \bar{X})^2}{2\sigma^2} \right] d\bar{X} \cdot \frac{2\sqrt{2}}{\sigma \sqrt{2\pi}} \exp \left[-\frac{2\sigma^2}{2\sigma^2} \right] d\sigma\end{aligned}\quad (4)$$

or, since $2\sigma d\sigma = d\sigma^2$

$$\begin{aligned}dP(\bar{X}, \sigma) &= \frac{1}{\sqrt{2\pi} \sigma / \sqrt{2}} \exp \left[-\frac{2(\bar{X} - \bar{X})^2}{2\sigma^2} \right] d\bar{X} \cdot \frac{\sqrt{2}}{\sigma \sqrt{2\pi}} \\ &\quad \exp \left[-\frac{2\sigma^2}{2\sigma^2} \right] d\sigma^2\end{aligned}\quad (4')$$

Equation (4') is the equation for the distribution of all sample points in terms of their means and variances.

Distribution of All Sample Means. Equation (4) gives the probability of a sample falling in either square $efgh$ or square $e'f'g'h'$ of Fig. 70. It is the probability of a sample having a mean between $\bar{X}$ and $\bar{X} + d\bar{X}$ and a standard deviation between

σ and $\sigma + d\sigma$. To find the probability of a sample having a mean between $\bar{X}$ and $\bar{X} + d\bar{X}$ and a standard deviation of any value, it is merely necessary, by the addition theorem, to sum probabilities like Eq. (4) for all values of the standard deviation. This can be done with reference to Fig. 64 by summing up the probabilities of all cells lying between AB and $A'B'$. Algebraically this sum is found by integrating Eq. (4) with respect to σ . Thus

$$dP(\bar{X}) = \frac{1}{\sqrt{2\pi} \sigma / \sqrt{2}} \exp \left[-\frac{2(\bar{X} - \bar{X})^2}{2\sigma^2} \right] d\bar{X} \int_0^\infty \frac{2\sqrt{2}}{\sigma \sqrt{2\pi}} \exp \left[-\frac{2\sigma^2}{2\sigma^2} \right] d\sigma$$

But the integral is of the form

$$2 \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

where $z = \sqrt{2}\sigma/\sigma$ and is equal to 1, since it is double one-half the area under the standard normal curve. Hence

$$\begin{aligned} dP(\bar{X}) &= \frac{1}{\sqrt{2\pi} \sigma / \sqrt{2}} \exp \left[-\frac{2(\bar{X} - \bar{X})^2}{2\sigma^2} \right] d\bar{X} \\ &= \frac{1}{\sigma_{\bar{X}} \sqrt{2\pi}} \exp \left[-\frac{(\bar{X} - \bar{X})^2}{2\sigma_{\bar{X}}^2} \right] d\bar{X} \end{aligned} \quad (5)$$

in which $\sigma_{\bar{X}} = \sigma/\sqrt{2}$.

This shows that, for a normal population, the distribution of means of samples of two is normal in form with a mean equal to the mean of the population and a variance equal to the variance of the population divided by two.

Distribution of All Sample Variances. The same process may be used to find the distribution of all sample variances. In this instance, the probabilities of all cells lying between lines GH and EF and lines $G'H'$ and $E'F'$ are summed. Algebraically the distribution is found by integrating (4') with reference to $\bar{X}$. Thus

$$dP(\sigma^2) = \frac{\sqrt{2}}{\sigma \sqrt{2\pi}} \exp \left[-\frac{2\sigma^2}{2\sigma^2} \right] d\sigma^2 \cdot \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi} \sigma_{\bar{X}}} \exp \left[-\frac{(\bar{X} - \bar{X})^2}{2\sigma_{\bar{X}}^2} \right] d\bar{X}$$

But the integral gives the total area under the normal curve and thus equals 1. Hence

$$dP(\sigma^2) = \frac{1}{\sigma \delta \sqrt{\pi}} \exp \left[-\frac{2\sigma^2}{2\delta^2} \right] d\sigma^2 \quad (6)$$

If χ^2 is set equal to $2\sigma^2/\delta^2$, and it is noted that $\sqrt{\pi} = (-\frac{1}{2})!$, this becomes

$$dP(\chi^2) = \frac{e^{-\chi^2/2} (\chi^2)^{-\frac{1}{2}}}{\sqrt{2} (-\frac{1}{2})!} d\chi^2$$

which shows that for samples of two from a normal population the sampling distribution of $2\sigma^2/\delta^2$ has the form of a χ^2 distribution with $n = 2 - 1 = 1$.¹

Distribution of All Sample Values of $t = \frac{\sqrt{2}(\bar{X} - \bar{X})}{\bar{\sigma}}$. To find the distribution of sample values of $t = \frac{\sqrt{2}(\bar{X} - \bar{X})}{\bar{\sigma}}$, first note that by definition $\bar{\sigma} = \sigma \sqrt{\frac{N}{N-1}}$ or when $N = 2$, $\bar{\sigma} = \sigma \sqrt{2}$. Hence for $N = 2$, $t = \frac{\bar{X} - \bar{X}}{\sigma}$, $\sigma^2 t^2 = (\bar{X} - \bar{X})^2$ and $\sigma dt = d\bar{X}$. Substituting these values in (4') gives

$$dP(t, \sigma) = \frac{\sigma}{\sqrt{2\pi} \sigma \sqrt{2}} \exp \left[-\frac{\sigma^2 t^2}{\delta^2} \right] dt \cdot \frac{\sqrt{2}}{\sigma \delta \sqrt{2\pi}} \exp \left[-\frac{\sigma^2}{\delta^2} \right] d\sigma^2$$

Combining terms yields

$$dP(t, \sigma) = \frac{dt}{\pi} \cdot \frac{1}{\delta^2} \exp \left[-\frac{\sigma^2}{\delta^2} (1 + t^2) \right] d\sigma^2$$

which may be put in the form

$$dP(t, \sigma) = \frac{dt}{\pi(1 + t^2)} \exp \left[-\frac{\sigma^2}{\delta^2} (1 + t^2) \right] d \left[\frac{\sigma^2}{\delta^2} (1 + t^2) \right]$$

To get the distribution of t , integrate this for all values of σ^2 from 0 to ∞ . Thus,

$$dP(t) = \frac{dt}{\pi(1 + t^2)} \int_0^\infty e^{-y} dy, \text{ where } y = \frac{\sigma^2}{\delta^2} (1 + t^2)$$

¹ Cf. equation for the χ^2 curve, page 111.

But the integral equals $-e^{-y} \Big|_0^{\infty}$ which equals 1. Hence

$$dP(t) = \frac{dt}{\pi(1+t^2)}$$

which is seen to be of the form of the t distribution with

$$n = N - 1 = 2 - 1 = 1.$$

Hence $\frac{\sqrt{2}(\bar{X} - \bar{X})}{\bar{s}}$ is distributed like t with $n = 1$.¹

THE GENERAL CASE $N > 2$

Before proceeding to the general case, it will be well to prepare the ground by reviewing several mathematical functions and relationships.

N-dimensional Geometry. Those who have had coordinate geometry will remember that a linear equation in two variables may be represented by a straight line drawn on the coordinate plane. If the equation is put in the form $aX_1 + bX_2 + c = 0$, $-c/b$ is the X_2 intercept, $-c/a$ is the X_1 intercept, and $-a/b$ is the slope of the line with the X_1 -axis. The "direction ratios" of the line, i.e., factors that are proportional to the cosines of the angles the line makes with the X_1 - and X_2 -axis, are $-a$ and b . Thus a line that is perpendicular to $aX_1 + bX_2 + c = 0$ would be one whose slope equals $+b/a$ and whose direction ratios are a and b . In particular, a line through the origin that made equal angles with the axes would be $X_1 - X_2 = 0$, and a line perpendicular to this line would be $X_1 + X_2 = c$. Similarly, in three dimensions, a linear equation of the form

$$aX_1 + bX_2 + dX_3 + c = 0$$

represents a plane, and a, b, d are the direction ratios of a line perpendicular to the plane. If this is generalized, it may be said that a linear equation of the form

$$aX_1 + bX_2 + \dots + kX_N + c = 0$$

represents a hyperplane, or flat space, in N dimensions and $a, b, \dots, k$ are the direction ratios of a line perpendicular to this hyperplane. Just as a line in a two-dimensional plane is said to have one dimension (i.e., length) and a plane in three

¹ Cf. equation for the t distribution, p. 111.

dimensions is said to have two dimensions (*i.e.*, length and breadth), a hyperplane in N dimensions is said to have $N - 1$ dimensions. It is to be noted that this use of N -dimensional space is merely a convenient way of generalizing certain algebraic relationships that have concrete counterparts in two and three dimensions. There can be no real N -dimensional figures.

This generalization of two- and three-dimensional relationships may be extended to other figures. Thus, in two dimensions, $(X_1 - a)^2 + (X_2 - b)^2 = r^2$ is the equation for a circle with center at (a, b) and radius equal to r ; in three dimensions, $(X_1 - a)^2 + (X_2 - b)^2 + (X_3 - c)^2 = r^2$ is a sphere with center at (a, b, c) and a radius equal to r . If these relationships are generalized, it may be said that

$$\dots (X_1 - a)^2 + (X_2 - b)^2 + \dots + (X_N - k)^2 = r^2$$

represents a hypersphere in N -dimensional space with center at $(a, b, \dots, k)$ and radius equal to r . The circumference of a circle is of one dimension (being only a line), the surface of a sphere is of two dimensions, and the surface of a hypersphere in N dimensions will be of $N - 1$ dimensions. These N -dimensional notions will be useful in the analysis that follows.

The Gamma Function. The integral

$$\Gamma(m) = \int_0^{\infty} e^{-x} x^{m-1} dx, \quad m > 0 \quad (7)$$

is known as the "gamma function." Integration by parts shows that the integral on the right equals

$$[-e^{-x} x^{m-1}]_0^{\infty} + \int_0^{\infty} e^{-x} (m-1) x^{(m-1)-1} dx \quad (8)$$

But the first term of Eq. (8) equals 0 when $x = 0$ and also when $x = \infty$ and the second term is seen to be equal to $(m-1)\Gamma(m-1)$. Hence we have the relationship

$$\Gamma(m) = (m-1)\Gamma(m-1) \quad (9)$$

By Eq. (9), $\Gamma(m-1) = (m-2)\Gamma(m-2)$, etc.; therefore, if m is a positive integer,¹

$$\Gamma(m) = (m-1)(m-2)(m-3) \dots 1 = (m-1)! \quad (10)$$

¹ The last term will be $\Gamma(1) = \int_0^{\infty} e^{-x} dx = [-e^{-x}]_0^{\infty} = e^0 = 1$.

Equation (10) holds for positive integral values of m . For fractional values of m , it is taken as the definition of $m!$. Thus, in general, $m!$ is defined by

$$m! = \Gamma(m + 1) \quad (10')$$

In statistical theory, $\Gamma(\frac{1}{2})$ is of particular interest; for if m is set equal to $\frac{1}{2}$ in Eq. (7), it becomes

$$\Gamma(\frac{1}{2}) = \int_0^{\infty} e^{-x} x^{-\frac{1}{2}} dx$$

By setting $x = y^2/2k^2$, this may be put in the form

$$\begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \int_0^{\infty} e^{-\frac{y^2}{2k^2}} \left(\frac{y^2}{2k^2}\right)^{-\frac{1}{2}} d\left(\frac{y^2}{2k^2}\right) \\ &= \int_0^{\infty} e^{-\frac{y^2}{2k^2}} \frac{k\sqrt{2}}{y} \frac{2y}{2k^2} dy \\ &= \frac{\sqrt{2}}{k} \int_0^{\infty} e^{-\frac{y^2}{2k^2}} dy \end{aligned}$$

If the right member is multiplied by $\sqrt{2\pi}/\sqrt{2\pi}$, this becomes

$$\Gamma\left(\frac{1}{2}\right) = 2\sqrt{\pi} \int_0^{\infty} \frac{1}{k\sqrt{2\pi}} e^{-\frac{y^2}{2k^2}} dy \quad (11)$$

But the integral is recognized as the sum of half the area of a normal frequency curve and is thus equal to $\frac{1}{2}$. Hence, Eq. (11) reduces to¹

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \quad (12)$$

In conclusion, it may be noted that

$$\int_0^x e^{-x} x^{m-1} dx$$

is known as the "incomplete gamma function." Tables of this function have been computed by Karl Pearson and his staff and have been published by the Cambridge University Press.

Distribution of All Possible Samples. *Derivation.* The first step in deriving the sampling distribution of means, variances,

¹ The integral $\int_0^{\infty} e^{-y^2} dy$ can be evaluated without making use of knowledge of the normal curve. See, for example, John F. Kenney, *Mathematics of Statistics* (1939), Part II, pp. 35-37.

etc., is to obtain the derivation of all possible samples of N from a normal population. This may be accomplished by direct application of the multiplication theorem.

The probability of a case lying between X_1 and $X_1 + dX_1$ is

$$dP(X_1) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(X_1 - \bar{X})^2}{2\sigma^2} \right] dX_1 \quad (13)$$

Similarly, the probability of a case lying between X_2 and $X_2 + dX_2$ is

$$dP(X_2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(X_2 - \bar{X})^2}{2\sigma^2} \right] dX_2 \quad (13')$$

and the same sort of equation holds for the probability of $X_3, X_4, \dots, X_N$.

If samples of N are drawn at random from an infinite normal population, the law of large numbers suggests that the relative frequency, or probability, of a sample in which one case lies between X_1 and $X_1 + dX_1$, another between X_2 and $X_2 + dX_2$, a third between X_3 and $X_3 + dX_3$, etc., will be predicted by the joint probability of $X_1, X_2, X_3, \dots, X_N$, that is, by

$$dP(X_1, X_2, \dots, X_N) = \frac{1}{(\sigma \sqrt{2\pi})^N} \exp \left[-\frac{\sum (X_i - \bar{X})^2}{2\sigma^2} \right] (dX_1)(dX_2) \dots (dX_N) \quad (14)$$

Equation (14) is the equation for the distribution of all possible samples of N from a normal population and is the model that any large number of samples of N will tend to approximate.

Geometrical Representation and Properties. If the sample values of $X_1, X_2, \dots, X_N$ are represented by a point in N -dimensional space with coordinates $X_1, X_2, \dots, X_N$, then Eq. (14) represents a cluster of sample points with center at $\bar{X}, \bar{X}, \bar{X}, \dots, \bar{X}$ (see Fig. 71a).

Geometrically Eq. (14) says that in any N dimensional cell whose sides extend from X_1 to $X_1 + dX_1$, X_2 to $X_2 + dX_2$, $\dots, X_N$ to $X_N + dX_N$ the density of the sample is given by

$$\frac{1}{(\sigma \sqrt{2\pi})^N} \exp \left[-\frac{\sum (X_i - \bar{X})^2}{2\sigma^2} \right]$$

and the probability of a sample falling in this cell is equal to the product of this density factor times the volume of the cell, viz.,

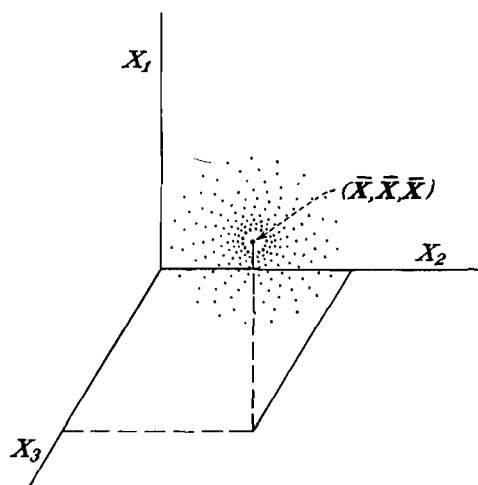


FIG. 71a.

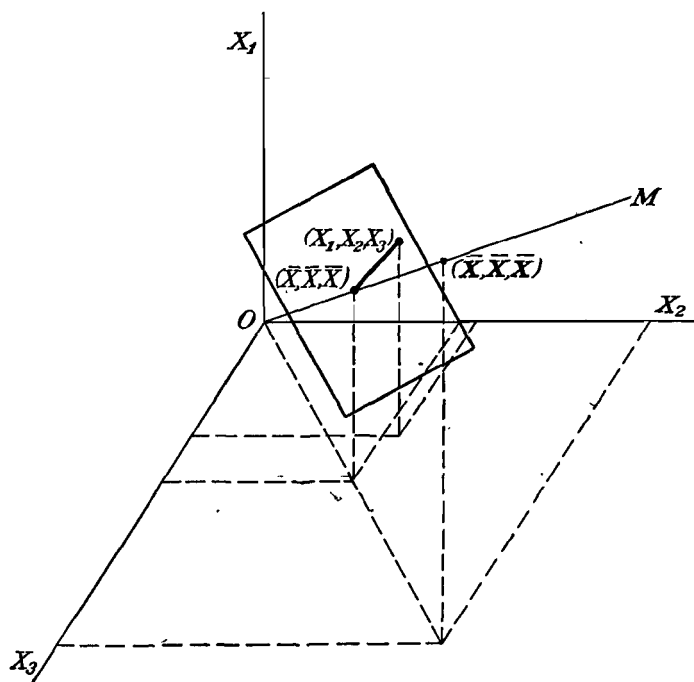


FIG. 71b.

$dX_1, dX_2, dX_3 \dots dX_N$. From this interpretation it follows that the density of sample points is the greatest when $X_1, X_2, \dots, X_N$ all equal $\bar{X}$, that is, when $X_1, X_2, \dots, X_N$ all lie at the center of the cluster. Since $\Sigma(X_i - \bar{X})^2$ equals the square of the distance of a sample point from the center point $\bar{X}, \bar{X}, \dots, \bar{X}$, Eq. (14) also shows that the density of sample points is constant for all cells lying on a hypersphere with center at $\bar{X}, \bar{X}, \dots, \bar{X}$ and radius equal to $\sqrt{\Sigma(X_i - \bar{X})^2}$.

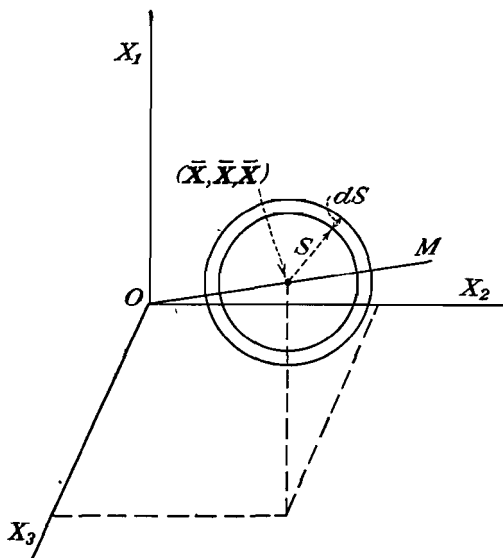


FIG. 72.

The center point $\bar{X}, \bar{X}, \dots, \bar{X}$ lies on the line OM , Fig. 71b, through the origin whose coordinates are equal. All samples with identical values lie on this line. The mean of any sample is

$$\frac{X_1 + X_2 + X_3 + \dots + X_N}{N} = \bar{X}$$

For a given value of $\bar{X}$ this defines a hyperplane that is perpendicular to the line OM since its direction ratios are all equal. In other words, all samples that have the same mean lie on a plane that is perpendicular to line OM and cut this line in the sample point $\bar{X}, \bar{X}, \dots, \bar{X}$. It follows that variation in the mean of a sample can be represented by variation along the line OM .

The variance of a sample is $\sigma^2 = \frac{\Sigma(X_i - \bar{X})^2}{N}$. Thus $N\sigma^2$ represents the square of the distance of a sample point from the point lying on OM whose coordinates are all equal to the mean of the sample. All samples with the same variance thus lie on a hypercylinder whose axis is the line OM . It follows that variation in the variance of a sample can be represented by variation in the square of the perpendicular distance of a sample point from the line OM .

Distribution of All Possible Samples in Terms of Their Means and Variances. It is suggested by the foregoing analysis that the distribution of all possible samples may be given in terms of the means and variances of the samples provided that the proper element of volume is found. As indicated above, the probability of a sample is the same for all "cubical cells" lying on the surface of a hypersphere, the probability being proportional to $\exp \left[\frac{-\Sigma(X_i - \bar{X})^2}{2\sigma^2} \right]$. If $\Sigma(X_i - \bar{X})^2$ is set equal to S^2 , this means that the density of sample points is constant throughout a shell with center at $\bar{X}, \bar{X}, \dots, \bar{X}$, radius equal to S , and thickness equal to dS . But¹

$$\begin{aligned}\Sigma(X_i - \bar{X})^2 &= \Sigma(X_i - \bar{X})^2 + N(\bar{X} - \bar{X})^2 \\ &= N\sigma^2 + N(\bar{X} - \bar{X})^2\end{aligned}$$

Hence the probability of samples for all cells lying in a given shell is proportional to

$$\exp \left[-\frac{N\sigma^2}{2\sigma^2} \right] \exp \left[\frac{-N(\bar{X} - \bar{X})^2}{2\sigma^2} \right]$$

that is, to the product of a function depending on the variance of a sample by a function depending on the mean of the sample. If the shell in question is cut by a hyperplane through the point $\bar{X}, \bar{X}, \dots, \bar{X}$ on OM and perpendicular to OM and another through the point $\bar{X} + d\bar{X}, \bar{X} + d\bar{X}, \dots, \bar{X} + d\bar{X}$ and also perpendicular to OM , these planes will cut off a shell of one less

¹ This is merely a "short" equation for obtaining the sum of the squares of the deviations from the mean of a sample by finding the sum of the squares of the deviations from an arbitrary origin (here the mean of the population) and the difference between the sample mean and this origin.

dimension¹ throughout which the density of sample points continues to be proportional to

$$\exp \left[-\frac{N\sigma^2}{2\sigma^2} \right] \exp \left[\frac{-N(\bar{X} - \bar{X})^2}{2\sigma^2} \right]$$

Within an error dependent upon infinitesimals of higher order, this second shell may be approximated by a "rectangular" shell whose inner radius is $\sqrt{\sum (X_i - \bar{X})^2} = \sqrt{N} \sigma$, whose outer radius is $\sqrt{N} (\sigma + d\sigma)$, and whose width is $d\bar{X}$ (see Fig. 74A,

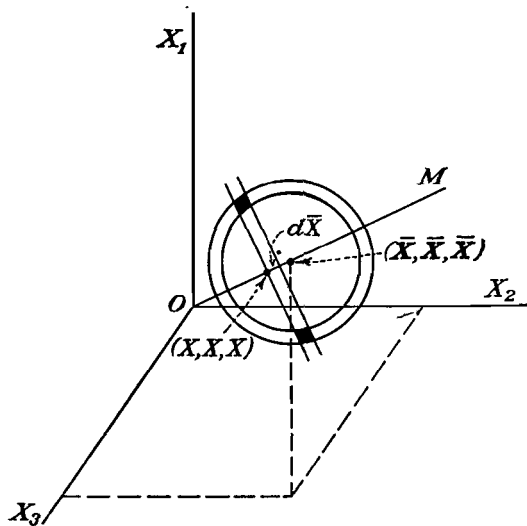


FIG. 73.

B, C).² The proportionate frequency of points in this second shell will be approximated by the product of the density factor

$$\exp \left[\frac{-N\sigma^2}{2\sigma^2} \right] \exp \left[\frac{-N(\bar{X} - \bar{X})^2}{2\sigma^2} \right]$$

¹ In three dimensions, the first shell is a true shell, the second is a "ring" with width approximately equal to $d\bar{X}$ and thickness equal to $d\sigma$ (see Fig. 74A, *B, C*).

² The argument is thought out in three dimensions but is generalized to N dimensions. The diagrams illustrate the three-dimensional argument. Figure 74A shows half the original shell. Figure 74B shows the rectangular approximation. Figure 74C shows that the areas of the cross sections are practically the same, although of different shapes.

times the volume of the shell, or some factor proportional to this volume. This relative frequency will represent the relative frequency, or probability, of all points whose means lie between $\bar{X}$ and $\bar{X} + d\bar{X}$ and whose standard deviations lie between σ and $\sigma + d\sigma$. It will thus give an expression for the distribution of all possible samples in terms of their means and standard deviations.

Since the radius of the shell in question is proportional to σ , its volume will be proportional to $(\sigma)^{N-2} d\sigma d\bar{X}$. For a hyperplane of N dimensions will intersect an N -dimensional hypersphere in another hypersphere of $N - 1$ dimensions (*e.g.*, a plane will intersect a sphere in a circle), and the generalized surface area

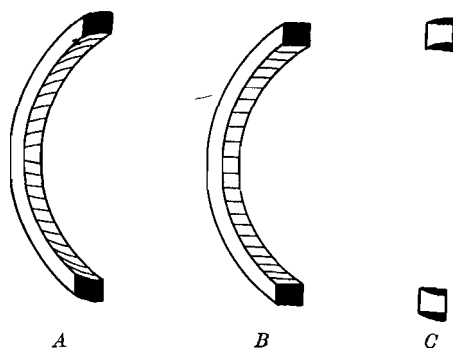


FIG. 74.

of the latter will be proportional to the radius raised to the $(N - 2)$ th power. For example, the surface of a three-dimensional sphere is proportional to r^2 and the circumference of a circle to r . The total volume will equal this surface area multiplied by the width of the shell $d\bar{X}$ and its thickness $d\sigma$. Multiplying the volume of the shell by the density factor thus gives the relative frequency of the samples lying in the shell as proportional to

$$\sigma^{N-2} \exp \left[\frac{-N\sigma^2}{2\sigma^2} \right] \exp \left[\frac{-N(\bar{X} - \bar{\mathbf{X}})^2}{2\sigma^2} \right] d\sigma d\bar{X}$$

But $d\sigma^2 = 2\sigma d\sigma$ or $d\sigma = \frac{d\sigma^2}{2\sigma}$; therefore, the above may be written

$$\frac{1}{2} (\sigma^2)^{\frac{N-3}{2}} \exp \left[\frac{-N\sigma^2}{2\sigma^2} \right] \exp \left[\frac{-N(\bar{X} - \bar{\mathbf{X}})^2}{2\sigma^2} \right] d\sigma^2 d\bar{X}$$

Thus it is to be concluded that the relative frequency of a sample having a mean lying between $\bar{X}$ and $\bar{X} + d\bar{X}$ and a variance lying between σ^2 and $\sigma^2 + d\sigma^2$ is

$$dP(\bar{X}, \sigma^2) = K \exp \left[\frac{-N(\bar{X} - \bar{X})^2}{2\sigma^2} \right] d\bar{X} (\sigma^2)^{\frac{N-3}{2}} \exp \left[\frac{-N\sigma^2}{2\sigma^2} \right] d\sigma^2 \quad (15)$$

where K is some constant independent of σ^2 and $\bar{X}$.

Distribution of All Sample Means. If σ^2 is given a particular value, Eq. (15) gives the distribution of sample means; *i.e.*, it gives the relative frequencies of samples with different mean values, but all the same σ^2 . It shows that these relative frequencies are proportional to

$$\exp \left[\frac{-(\bar{X} - \bar{X})^2}{2\sigma^2/N} \right] d\bar{X}$$

which is obviously the form of a normal frequency distribution with mean at $\bar{X}$ and variance equal to σ^2/N . It also shows that the form of the distribution is the same no matter what the value of σ^2 chosen. In other words, the distribution of sample means is independent of the value of the sample variance. Hence, if the relative frequencies are summed for all values of σ^2 , the relative frequencies of various mean values will be given in general by

$$dP(\bar{X}) = K_1 \exp \left[\frac{-(\bar{X} - \bar{X})^2}{2\sigma^2/N} \right] d\bar{X}$$

To obtain K_1 , integrate the expression above from $-\infty$ to $+\infty$ and set it equal to 1, the total frequency. Thus

$$\int_{-\infty}^{\infty} K_1 \exp \left[\frac{-(\bar{X} - \bar{X})^2}{2\sigma^2/N} \right] d\bar{X} = 1$$

or setting $y = \frac{\bar{X} - \bar{X}}{\sigma/\sqrt{N}}$ and noting that $dy = \frac{d\bar{X}}{\sigma/\sqrt{N}}$,

$$\frac{\sigma\sqrt{2\pi}}{\sqrt{N}} K_1 \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy = 1$$

But the integral is obviously the area under a normal curve of unit variance and hence is 1. Thus $K_1 = \frac{1}{\sqrt{2\pi} \sigma/\sqrt{N}}$. Hence

$$dP(\bar{X}) = \frac{1}{\sqrt{2\pi} \delta / \sqrt{N}} \exp \left[\frac{-(\bar{X} - \bar{X})^2}{2\delta^2/N} \right] d\bar{X} \quad (16)$$

which is the distribution of all sample means.

Distribution of All Sample Variances. If $\bar{X}$ is given a particular value, Eq. (15) gives the distribution of sample variances, i.e., the relative frequencies of samples with different σ^2 's, but all the same $\bar{X}$'s. It shows that these relative frequencies are proportional to

$$(\sigma^2)^{\frac{N-3}{2}} \exp \left[\frac{-N\sigma^2}{2\delta^2} \right] d\sigma^2$$

and thus indicates that the distribution of sample variances is independent of the value of the sample mean. Hence, if the relative frequencies are summed for all sample mean values, the relative frequencies of various variances will be given in general by

$$dP(\sigma^2) = K_2 (\sigma^2)^{\frac{N-3}{2}} \exp \left[\frac{-N\sigma^2}{2\delta^2} \right] d\sigma^2$$

To obtain K_2 , integrate this expression from 0 to $+\infty$ and set it equal to 1, the total frequency. Thus

$$\int_0^\infty K_2 (\sigma^2)^{\frac{N-3}{2}} \exp \left[\frac{-N\sigma^2}{2\delta^2} \right] d\sigma^2 = 1$$

But if $N\sigma^2/\delta^2$ is set equal to χ^2 , this may be put in the form

$$\frac{(\delta^2)^{\frac{N-1}{2}} 2^{\frac{N-1}{2}} K_2}{N^{\frac{N-1}{2}}} \int_0^\infty e^{-\frac{\chi^2}{2}} \left(\frac{\chi^2}{2} \right)^{\frac{N-3}{2}} d \left(\frac{\chi^2}{2} \right) = 1$$

But the integral equals $\Gamma \left(\frac{N-1}{2} \right)$; therefore,

$$K_2 = \frac{N^{\frac{N-1}{2}}}{2^{\frac{N-1}{2}} \Gamma \left(\frac{N-1}{2} \right) (\delta^2)^{\frac{N-1}{2}}}$$

and

$$dP(\sigma^2) = \frac{N^{\frac{N-1}{2}} (\delta^2)^{\frac{N-1}{2}}}{2^{\frac{N-1}{2}} \Gamma \left(\frac{N-1}{2} \right) (\delta^2)^{\frac{N-1}{2}}} \exp \left[\frac{-N\sigma^2}{2\delta^2} \right] d\sigma^2 \quad (17)$$

If $N\sigma^2/\delta^2$ is set equal to χ^2 , this takes the form of a χ^2 distribution with $n = N - 1$. Since $d\sigma^2 = 2\sigma d\sigma$, the distribution of σ is

$$dP(\sigma) = \frac{N^{\frac{N-1}{2}} (\sigma)^{N-2}}{2^{\frac{N-3}{2}} \Gamma\left(\frac{N-1}{2}\right) (\delta^2)^{\frac{N-1}{2}}} \exp\left[-\frac{N\sigma^2}{2\delta^2}\right] d\sigma \quad (18)$$

Distribution of Sample Values of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$. Since the distribution of sample variances is independent of the distribution of sample means, and vice versa, their joint distribution is equal to the product of their individual distributions, shown by Eqs. (16) and (17). The exact form for Eq. (15) is thus given by the product of Eqs. (16) and (17), or

$$\begin{aligned} dP(\bar{X}, \sigma^2) &= \frac{1}{\sqrt{2\pi} \delta / \sqrt{N}} \exp\left[-\frac{(\bar{X} - \bar{X})^2}{2\delta^2/N}\right] d\bar{X} \\ &\quad \frac{N^{\frac{N-1}{2}} (\sigma^2)^{\frac{N-3}{2}}}{2^{\frac{N-1}{2}} \Gamma\left(\frac{N-1}{2}\right) (\delta^2)^{\frac{N-1}{2}}} \exp\left[-\frac{N\sigma^2}{2\delta^2}\right] d\sigma^2 \\ &= \frac{N^{\frac{N}{2}} (\sigma^2)^{\frac{N-3}{2}}}{2^{\frac{N}{2}} \sqrt{\pi} \Gamma\left(\frac{N-1}{2}\right) (\delta^2)^{\frac{N}{2}}} \\ &\quad \exp\left[-\frac{N\sigma^2}{2\delta^2} - \frac{(\bar{X} - \bar{X})^2}{2\delta^2/N}\right] d\sigma^2 d\bar{X} \quad (15') \end{aligned}$$

The distribution of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ may be obtained from Eq. (15') as follows: First set the symbol $t = \frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ and note that $\bar{\sigma} = \sigma \sqrt{\frac{N}{N-1}}$. Therefore, $t = \frac{\sqrt{N-1}(\bar{X} - \bar{X})}{\sigma}$, while

$$dt = \frac{\sqrt{N-1}}{\sigma} d\bar{X}.$$

Substituting $\frac{\sigma^2 t^2}{N-1}$ for $(\bar{X} - \bar{X})^2$ and $\frac{\sigma}{\sqrt{N-1}} dt$ for $d\bar{X}$ in Eq. (15') yields

$$dP(t, \sigma^2) = \frac{N^{\frac{N}{2}} (\sigma^2)^{\frac{N-2}{2}}}{2^{\frac{N}{2}} \sqrt{N-1} \sqrt{\pi} \Gamma\left(\frac{N-1}{2}\right) (\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{N\sigma^2}{2\sigma^2} - \frac{N}{N-1} \frac{\sigma^2 t^2}{2\sigma^2} \right] d\sigma^2 dt$$

which may be put in the form

$$dP(t, \sigma^2) = \frac{N^{\frac{N}{2}} (\sigma^2)^{\frac{N-2}{2}}}{2^{\frac{N}{2}} \sqrt{N-1} \sqrt{\pi} \Gamma\left(\frac{N-1}{2}\right) (\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right] d\sigma^2 dt$$

Further manipulation shows that the following is the equivalent of the foregoing:

$$dP(t, \sigma^2) = \left[\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right]^{\frac{N-2}{2}} \exp \left[-\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right] d \left[\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right] \frac{dt}{\sqrt{N-1} \sqrt{\pi} \Gamma\left(\frac{N-1}{2}\right) \left(1 + \frac{t^2}{N-1} \right)^{\frac{N}{2}}}$$

Here t is taken as a constant in $d \left[\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right]$. If the foregoing is summed or integrated for σ^2 for a given value of t , the probability of that t is given by

$$dP(t) = \frac{dt}{\sqrt{N-1} \sqrt{\pi} \Gamma\left(\frac{N-1}{2}\right) \left(1 + \frac{t^2}{N-1} \right)^{\frac{N}{2}}} \int_0^\infty \left[\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right]^{\frac{N-2}{2}} \exp \left[-\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right] d \left[\frac{N\sigma^2}{2\sigma^2} \left(1 + \frac{t^2}{N-1} \right) \right]$$

But the integral is of the form $\int_0^\infty e^{-x} X^{\frac{N-2}{2}} dX$, which equals

$\Gamma\left(\frac{N}{2}\right)$. Hence, the distribution of t is

$$dP(t) = \frac{\Gamma\left(\frac{N}{2}\right) dt}{\sqrt{N-1} \sqrt{\pi} \Gamma\left(\frac{N-1}{2}\right) \left(1 + \frac{t^2}{N-1}\right)^{\frac{N}{2}}}$$

or if n is set equal to $N - 1$,

$$dP(t) = \frac{\Gamma\left(\frac{n+1}{2}\right) dt}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right) \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

which is recognized as the form of the t distribution. Hence $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ is distributed like t with $n = N - 1$.

CHAPTER XI

SAMPLING FROM CONTINUOUS NORMAL POPULATIONS II. USES OF THE SAMPLING DISTRIBUTIONS

SAMPLING DISTRIBUTION OF THE MEAN

Used to Test a Hypothesis. Problem When Population Variance Is Known. The practical use of the sampling distributions discussed in the preceding chapter may be explained by reference to several examples. Consider first an illustration of how the sampling distribution of the mean is used to test a hypothesis. Suppose it is claimed that a new process of manufacturing electric light bulbs will increase their mean length of life to 1,100 kilowatt-hours. A manufacturer tries out this new process and produces a sample batch of 100 bulbs for which the mean length of life proves to be 1,090. The new process, it may be supposed, does not change the variability in length of life from bulb to bulb, and therefore the standard deviation in length of life may be taken as that of the old process. Suppose this is known to be 200 kilowatt-hours. The question to which the manufacturer seeks an answer is therefore this: Is it reasonable to infer that the given sample is from a population in which the mean is 1,100 kilowatt-hours? This is the question that will now be discussed in some detail.

Coefficient of Risk. Before answering the question the manufacturer must first determine what degree of risk he is willing to undergo in rejecting the hypothesis when it is true. Suppose he is willing to make this mistake once in twenty times on the average; his coefficient of risk will then be .05.

Region of Rejection. The next step is to select a "region of rejection," which on the basis of the given hypothesis will mark off the unreasonable or unacceptable samples from the others, the probability of this subset of samples being just .05. Since the sampling distribution of the mean is normal in form, the argument is essentially the same as that pertaining to percentages (see Chap. IX).

If the manufacturer is indifferent as to what mean value other than the hypothetical mean value is the true one, he will do best to distribute his .05 region of rejection equally at both ends of the sampling distribution. If he wishes to reject the hypothesis more often when the true mean value is actually below the hypothetical mean value than when it is above this value, he will do best to put the .05 region entirely at the lower end of the distribution. If he wishes to reject the hypothesis more often when the true mean value is actually above the hypothetical mean value than when it is below this value, he will do best to put his .05 region entirely at the upper end of the sampling distribution.

In the present instance, a region of rejection lying entirely at the lower end of the distribution would appear to be the best region to adopt; for the manufacturer would not care if the claims of the new process were more than substantiated. He would care only if they fell short of being true.

Testing the Hypothesis. Since .05 of the area of a normal frequency curve lies below 1.645 σ from the mean and since the standard deviation of the sampling distribution of the mean is equal to the standard deviation of the population divided by $\sqrt{N}$, the .05 region of rejection lying entirely at the lower end of the distribution would consist of all samples whose means lay below 1,100 (the hypothetical population mean) minus

$$(1.645) \frac{200}{\sqrt{100}},$$

that is, below $1,100 - 32.9 = 1067.1$.

A graph of the sampling distribution of the mean and this region of rejection is shown in Fig. 75. Since the actual sample value was 1,090, it does not fall in this region of rejection and the hypothesis is not rejected. Even though the sample failed to have a mean as high as that claimed for the new process, the manufacturer will not reject the claim that the new process will make bulbs averaging 1,100 kilowatt-hours of life. That is, the amount by which the sample average fell short of the figure claimed was not large enough to warrant the rejection of the claim.

Other Examples. Other cases might be conceived of in which an upper region of rejection or an evenly distributed region

of rejection would be the proper region to employ. For example, if a process of cigarette manufacture is claimed to yield cigarettes of a low burning temperature, an upper region of rejection would be appropriate. Or a manufacturer of automobile tires may wish to copy the tires of a competitor as closely as possible, and in this case he might be equally eager to avoid turning out tires that were markedly worse or markedly better than the competing brand; here an equally distributed region of rejection would be in order.

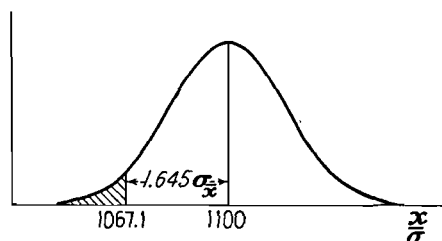


FIG. 75.

Confidence Limits for the Population Mean. Problem When the Population Variance Is Known. In many cases it is the aim of the statistical investigation to determine confidence limits for the mean of the population from which a sample has been drawn, rather than to test any particular hypothesis. Again this may be done in a manner that is very similar to the process of determining confidence limits for percentage figures. The following discussion will relate to the electric-light bulb example described in the preceding section. As before, it will be assumed that the population variance is known.

The Confidence Coefficient. Confidence limits, it will be recalled, are the end points of a certain range of values that may be said to include the true value with a given degree of probability. This degree of probability is called the "confidence coefficient" associated with the given "confidence interval." Suppose in the present instance that the manufacturer of electric-light bulbs adopts a confidence coefficient of .95; that is, he wishes to be able to say that there is a probability of .95 that the confidence interval he establishes covers the true mean value.

The Confidence Interval. As previously pointed out, innumerable confidence intervals may be set up that will all have a

confidence coefficient of .95. Suppose the manufacturer is as much concerned with underestimating the true mean value as with overestimating it. Then he will proceed as follows:

He will select a lower value such that if it is the true mean value the probability of getting the given sample mean or some higher value will be just .025. The process is illustrated in Fig. 76A. Next he will select some upper value such that if it is the true value the probability of getting the given sample

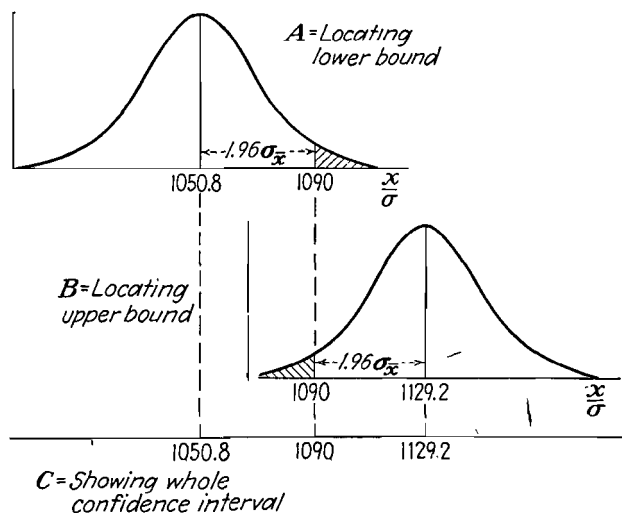


FIG. 76.

mean or some lower value will be just .025. This is illustrated in Fig. 76B.

The two values so determined will mark the upper and lower confidence limits for the population mean, as illustrated graphically in Fig. 76C. The distance between either limit and the sample mean value will be equal to 1.96 times the standard deviation of the sampling distribution of the mean, that is, 1.96 times the "standard error" of the mean, as it is called. For the sampling distribution of the mean is normal, and the probability of a sample value exceeding the true mean value by 1.96 σ is just .025; and the same is true of a sample value falling short of the true mean value by 1.96 σ . Since the standard error of the mean is equal to the standard deviation of the population, divided by the square root of N , the confidence

limits for the population mean may be found in this instance by simply laying off $\pm 1.96 \frac{\sigma}{\sqrt{N}}$ from the sample mean value.

For the given problem, these limits would be equal to

$$1,090 \pm 1.96 \frac{200}{\sqrt{100}},$$

or 1,129.2 and 1,050.8 kilowatt-hours. This interval is symmetrical with respect to the sample mean value and does not tend to overestimate the true value more than it tends to underestimate it.

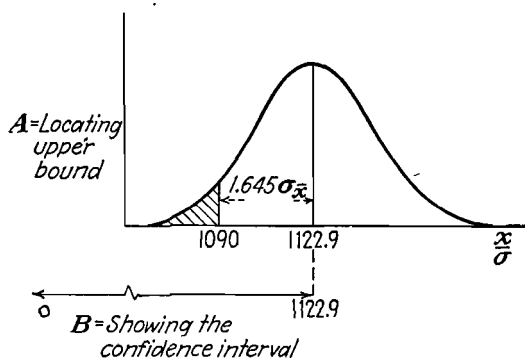


FIG. 77.

If the manufacturer is concerned solely with the possibility of overestimating the true mean value, he may desire only an upper confidence limit. The confidence interval will presumably run from zero to this upper bound. If the upper bound is selected as that value such that if it is the true mean value the probability of getting the sample mean value or a lower value is just .05, then this confidence interval may be said to have a probability of .95 of covering the true mean value. Since the probability of a normally distributed variate falling short of its true mean value by 1.645σ is just .05, the upper limit of the confidence interval may be found by adding 1.645 times the standard error of the mean to the sample mean value. This is illustrated in Fig. 77. For the given problem, this yields

$$1,090 + 1.645 \frac{200}{\sqrt{100}} = 1,122.9. \text{ It may be said, then, that}$$

there is a probability of .95 that the range 0-1,122.9 covers the true mean value.

If the manufacturer wished not to underestimate the true mean value, he would reverse the foregoing process. He would subtract 1.645 times the standard error of the mean from the sample mean value to obtain a lower bound for the true mean value. This is illustrated in Fig. 78 and would be

$$1,090 - 1.645 \frac{200}{\sqrt{100}} = 1,057.1.$$

He could then say that there was a probability of .95 that the range 1057.1- ∞ covered the true mean value.

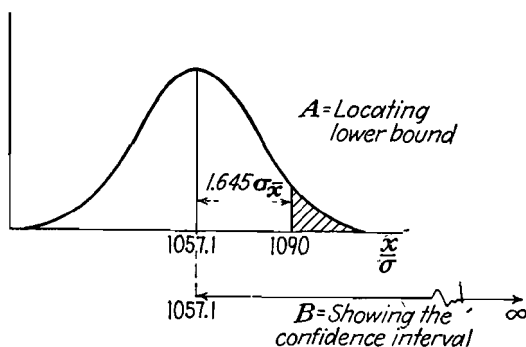


FIG. 78.

Relationship between Confidence Coefficient, Confidence Interval, and N . It is to be noted that the confidence limits set up in this way depend on the confidence coefficient adopted. The smaller this coefficient, the closer the limits to the sample value, and vice versa.¹ In other words, one can always narrow the range that is presumed to cover the true value if he is willing to increase the risk of its not doing so. Since the standard error of the mean is equal to the standard deviation of the population, divided by the square root of N , it also is to be noted that, the larger the sample, the closer the confidence limits to the sample value. This is merely another way of saying that, the larger the size of the sample, the more confidence one can have that the true mean value is somewhere in the immediate neighborhood

¹ This of course, is true only of the upper bound in Fig. 77 and the lower bound in Fig. 78.

of the sample mean value.¹ These points should be carefully noted.

Maximum-likelihood Estimate of the Population Mean. The Problem When Population Variance Is Known. In addition to setting up a range that might reasonably be presumed to include the population mean, the manufacturer of electric-light bulbs may wish to have a single estimate of the population mean that could be viewed as the "best" estimate to be made of this population parameter. As pointed out in the preview of theory, one way of making such a single estimate is to take that value for the population mean that makes the probability of the sample mean a maximum.² An estimate so made is a maximum-likelihood estimate.

Since the sampling distribution of the mean is normal, it follows that the probability of a sample mean is the greatest when the sample mean has the same value as the population mean.³ If, therefore, in estimating the population mean from the sample mean, the former is taken equal to the latter, then the probability of the given sample mean is a maximum.

In the problem under discussion the mean of the sample of electric bulbs was 1,090 kilowatt-hours. The maximum-likelihood estimate of the population mean is therefore $\bar{X} = 1,090$ kilowatt-hours.

SAMPLING DISTRIBUTION OF $\frac{\sqrt{N}(\bar{X} - \bar{\mu})}{\sigma}$

Used to Test a Hypothesis. The Problem When the Population Variance Is Unknown. The analysis of the preceding section was based on the assumption that the variance of the population is known. If the variance of the population is not known and the hypothesis to be tested does not prescribe any value for this parameter, the sampling distribution of the statistic $\frac{\sqrt{N}(\bar{X} - \bar{\mu})}{\sigma}$ should be used.⁴

¹ A more precise statement is this: The larger the value of N , the greater the probability that a given finite confidence interval, however small, will include the true value.

² See p. 180.

³ For then $P(\bar{X})$ would equal $\frac{1}{\sigma_{\bar{X}} \sqrt{2\pi}}$, which is the highest value it can have.

⁴ This is called a "composite hypothesis"; for it supposes that the sample

For purposes of illustration suppose that the electric-light bulb example is modified as follows: An inventor offers a new process of manufacturing electric-light bulbs that will, he claims, increase the mean length of life to 1,100 kilowatt-hours. In the absence of any prolonged experience with the new process, the inventor makes no claims regarding the variability in length of life from bulb to bulb, and there is no reason a priori to believe that the variability of bulbs manufactured by the new process will be the same as that of bulbs manufactured by the old. The latter, it will be recalled, was the assumption of the preceding section. Furthermore, suppose the manufacturer to whom the process is offered is interested, not in the variability in length of life of the bulbs, but only in their mean length of life. To test the claim of the inventor he manufactures a batch of 10 bulbs by the new process and finds that the mean length of life of these 10 bulbs is 1,090 kilowatt-hours.¹

In view of this result, is it reasonable to accept the hypothesis that the new process will in general produce bulbs whose mean length of life is 1,100 kilowatt-hours? Note that nothing is said here about the variance in length of life. To put the question another way, is it reasonable to assume that this sample could have come from a population in which the mean length of life was 1,100 kilowatt-hours and the variance was any value whatever?

Since the hypothesis does not prescribe any particular value for the variance of the population and its value is not known, the proper statistic to use in testing the given hypothesis is $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\hat{\sigma}}$; for in this statistic, only the hypothetical value of the population mean, $\bar{X}$, enters. The quantity $\hat{\sigma}^2$, it will be recalled, is the maximum-likelihood estimate of the population variance that is made from the sample; it is equal to the variance of the sample times $\frac{N}{N-1}$. For the problem illustrated, let the value of the sample variance be $(180)^2$ kilowatt-hours.²

is not from some one specific population but from any one of a group of populations all of which have the same mean.

¹ A small sample is used here purposely (see p. 284).

² This, it is recalled, is calculated from the sample data by the formula $\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{N}$.

Coefficient of Risk. In order to answer the question that has been posed, the manufacturer must adopt a definite coefficient of risk that will measure the chance he is willing to take of rejecting the hypothesis when it is actually true. As before, suppose he adopts a coefficient of risk equal to .05.

Testing the Hypothesis with a Symmetrical Region of Rejection. Next the manufacturer must adopt a definite region of rejection with a probability of .05. Such a region will be attained if from among all the possible sample values of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$, as represented by the sampling distribution of this statistic, he chooses a special subset such that the probability of a sample falling in this subset is just .05. Since the various possible

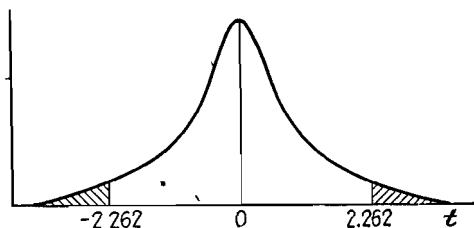


FIG. 79.

sample values of this statistic are distributed in the form of a t distribution, with $n = N - 1$, the manufacturer will secure a .05 region of rejection if he includes in the region all values of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ that are numerically greater than 2.262.* The

value 2.262 is found by consulting a table of the t distribution, in which it is seen that for $n = 9$ (that is, $10 - 1$) the probability of getting an absolute value of t that is equal to or greater than 2.262 is just .05. As shown in Fig. 79, this is a symmetrical region of rejection because the t distribution is symmetrically distributed about its mean of zero.

If the manufacturer adopts this symmetrical region of rejection, he can test the given hypothesis as follows: This hypothesis is that $\bar{X} = 1,100$. The sample mean $\bar{X}$ is 1,090, $N = 10$, and the sample standard deviation is 180. The last gives

$$\sigma = 180 \sqrt{\frac{10}{9}} = 189.7;$$

* See pp. 110-111, 474.

For these values the sample value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$ is

$$\frac{\sqrt{10}(1,090 - 1,100)}{189.7} = -.167.$$

This is greater than -2.262 and less than $+2.262$, and the sample value does not therefore fall in the region of rejection. This is illustrated by Fig. 79. The sample mean, although less than $1,100$, is not sufficiently less than this quantity to cast doubt on its being the population mean.

Other Regions of Rejection. The region of rejection adopted in the foregoing instance was a symmetrical region with respect to the t distribution. This would be the proper region to adopt

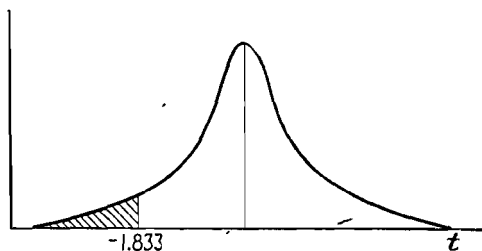


FIG. 80.

if the manufacturer were indifferent to whether the true mean value of the new process was above or below the claimed value. Since it is reasonable to suppose that he would be more concerned if the true mean value were less than $1,100$ than if it were above this amount, a region that contains sample values all at the lower end of the distribution is more appropriate for testing the given hypothesis. Such a region would be given by those values of t that (for $n = 9$) are equal to or less than -1.833 . For the t table shows that the probability of a t less than -1.833 is just .05.* A picture of this appropriate region and the location of the sample value with reference to it is shown in Fig. 80.

In another case a region of rejection lying all at the upper end of the distribution might be the more appropriate. It is suggested that the reader himself think of a case in which this might be true and work out an illustrative problem.

* The table gives the probability of an absolute value of t equal to or greater than $|1.833|$ as .10; hence the probability of a t equal to or less than -1.833 is .05.

Case Where Population Variance Is Known Compared with Case Where Population Variance Is Not Known. The difference between the case in which the standard deviation of the population is assumed to be known and the present case, in which it is not assumed to be known, may be made clear by several diagrams. These diagrams show how the various regions of rejection look in

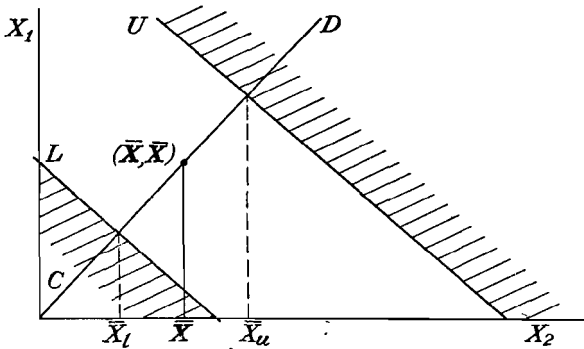


FIG. 81a.

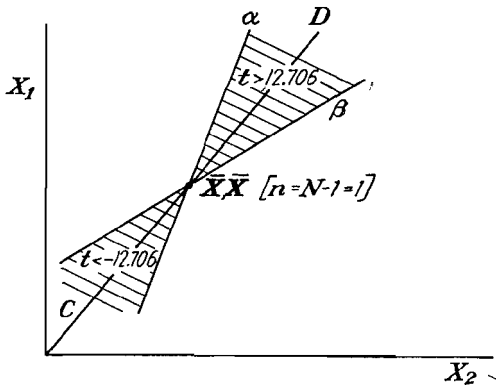


FIG. 81b.

terms of Fig. 64 (page 224). Consider first the symmetrical regions of rejection. In the case in which the standard deviation of the population is known, a hypothesis will be rejected if the sample falls in a region in which the mean of the sample deviates from the hypothetical population mean by more than a given amount, say 1.96 σ , for a .05 symmetrical region. In terms of Fig. 64 this means that all samples falling outside the lines L and U , shown in Fig. 81a, which is a miniature reproduction of Fig. 64

omitting the probabilities, will lead to the rejection of the hypothesis that the population mean is $\bar{X}$. This is illustrated by the crosshatched portions of Fig. 81a.

In the case in which the standard deviation of the population is not known the hypothesis is rejected if the sample falls in a region for which $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\hat{\sigma}}$ is numerically less than the .025 value of t . In terms of Fig. 64, this means that all samples lying between the lines α and β , Fig. 81b, which is also a miniature reproduction of Fig. 64 omitting the probabilities, will lead to the

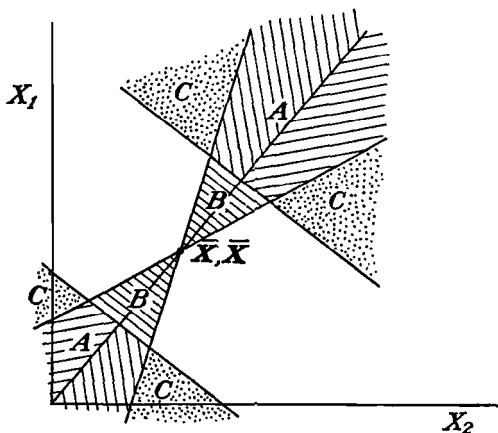


FIG. 81c.

rejection of the hypothesis that $\bar{X}$ is the mean of the population. This is illustrated by the crosshatched portion of Fig. 81b.

Although the regions are symmetrical in both instances, it is to be noted that they do not include the same set of samples. Some samples are common to each, but there are other samples that are included in only one of the two regions.

As shown by Fig. 81c, samples falling in regions marked A would lead to the rejection of the hypothesis in either instance. Samples falling in regions C would lead to rejection of the hypothesis only when the standard deviation is known, while samples falling in regions B would lead to rejection of the hypothesis only when the standard deviation of the population is estimated from the sample.

The reason why the latter samples lead to rejection of the hypothesis despite the fact that their mean values are relatively

close to the population mean is that their standard deviations, and hence the estimates of the population standard deviation, are so small that the difference between the hypothetical population mean and the sample mean, small as it actually is, appears to be relatively large. Likewise, although the means of samples falling in regions *C* differ greatly from $\bar{X}$, these samples do not

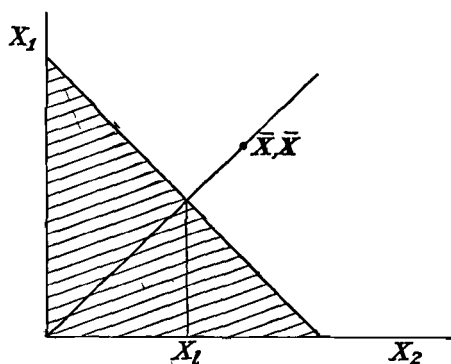


FIG. 82a.

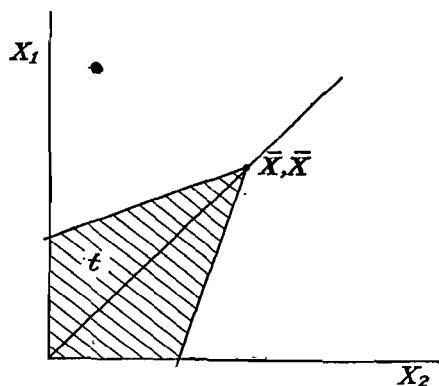


FIG. 82b.

lead to rejection of the hypothesis when the standard deviation of the population is not known, despite the great difference between their mean values and that of the population mean, simply because the standard deviations of these samples (and hence the estimates of the population standard deviation) make this difference seem relatively small. Similar remarks apply to the one-sided regions shown in Fig. 82a to c.

Confidence Limits for the Population Mean. Problem When the Population Variance Is Unknown. The determination of confidence limits for the population mean when the population variance is not known proceeds much the same as in the case when the variance is known. The essential difference is that the t distribution is used in place of the normal distribution.

Suppose that a manufacturer finds that the mean length of life of 10 electric-light bulbs manufactured by some process is 1,090 kilowatt-hours and that the standard deviation of this sample is 180 kilowatt-hours. Nothing is known about the standard deviation in length of life of bulbs producible by the

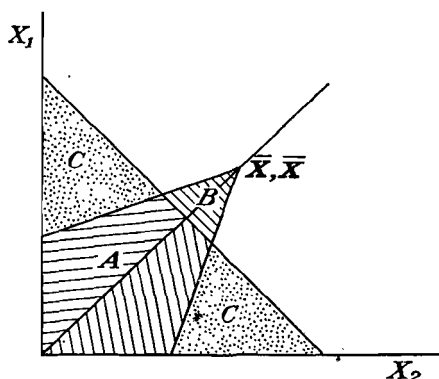


FIG. 82c.

new process other than the information provided by the sample. Under these conditions the manufacturer wishes to determine limits within which the average length of life of bulbs producible by the new process might in general be expected to lie. That is, he wishes to set up confidence limits for the mean of the population of bulbs producible by the new process. In setting up these limits he wants a confidence coefficient of .95, say.

Determination of Confidence Limits. If the manufacturer is indifferent as to whether he overestimates or underestimates the true mean length of life, he will adopt limits that are symmetrical with reference to the given sample mean. In this case, the upper limit may be found by determining a value of $\bar{X}$ such that the probability of getting the sample value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ or a lower value is just .025, and the lower limit

may be found by determining the value of $\bar{X}$ that makes the probability of getting the sample value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ or a higher value just equal to .025. For if this procedure is always adopted in setting up symmetrical confidence limits, the limits so determined will include the population mean 95 times out of 100 on the average.¹ Since the sampling distribution of

$$\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$$

is a t distribution, with n in the t formula equal to $N - 1$, and since, for $n = 10 - 1 = 9$, the .025 points of the t distribution are ± 2.262 , these upper and lower limits for $\bar{X}$ can be found by setting $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma} = \pm 2.262$ and solving for $\bar{X}$. For the given problem this yields the following,

$$\frac{\sqrt{10}(1,090 - \bar{X})}{180\sqrt{\frac{10}{9}}} = \pm 2.262$$

or $\bar{X} = 1,225.72$ as the upper limit and 954.28 as the lower limit. The analysis is illustrated by Fig. 83.

If the manufacturer wishes not to overestimate the mean of the population and does not care whether it is underestimated or not, he will seek only an upper bound for his confidence interval; in other words, he will seek an interval that runs from zero to this upper bound. For a confidence coefficient of .95, such an upper bound may be found by determining the value of $\bar{X}$ that makes the probability of the sample value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ or a lower value just equal to .05. Since the lower .05 point of the t distribution (for $n = 10 - 1 = 9$) is -1.833 , the upper bound for $\bar{X}$ in the present instance is given by

$$\frac{\sqrt{10}(1,090 - \bar{X})}{180\sqrt{\frac{10}{9}}} = -1.833$$

which gives the value 1,199.98 for $\bar{X}$. The whole confidence

¹ See pp. 174-179.

interval is thus 0 to 1,199.98. The analysis is illustrated in Fig. 84.

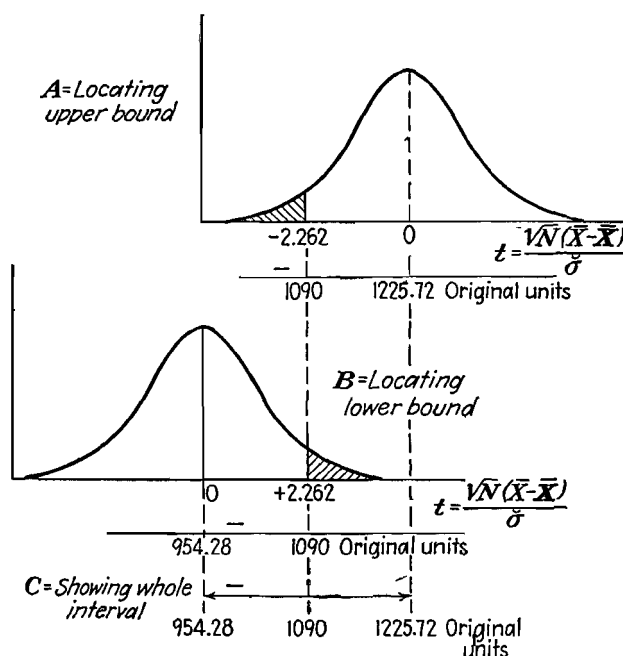


FIG. 83.

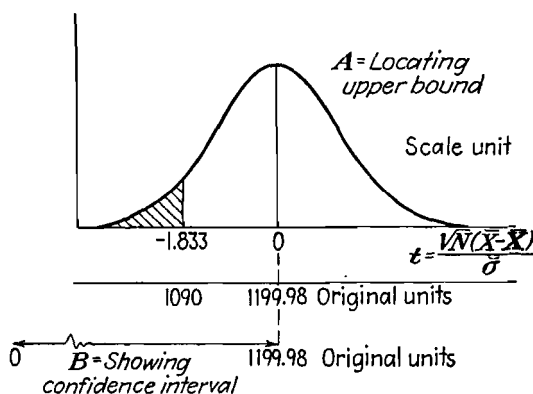


FIG. 84.

It might happen in cases of this kind that the manufacturer desires not to underestimate the mean of the population. In

such an instance he may wish only to determine a lower bound for the population mean value, the upper "bound" presumably being infinity. For a confidence coefficient of .95, a lower bound of this kind can be found by determining the value of $\bar{X}$ that makes the probability of the sample value of $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\hat{\sigma}}$ or a higher value just equal to .05. Since the upper .05 point of

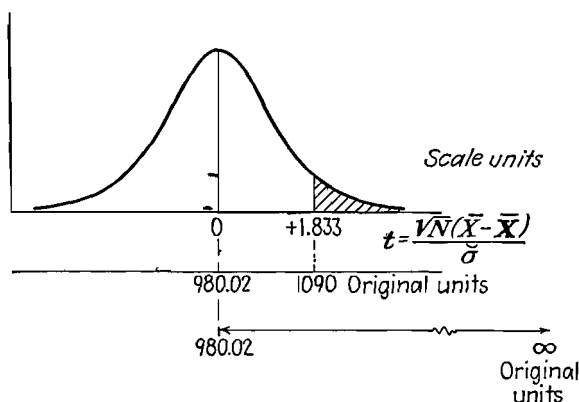


FIG. 85.

the t distribution (for $n = 10 - 1 = 9$) is $+1.833$, the lower bound for $\bar{X}$ in the given problem is yielded by the equation

$$\frac{\sqrt{10}(1,090 - \bar{X})}{180\sqrt{\frac{10}{9}}} = +1.833$$

which gives $\bar{X} = 980.02$. The whole confidence interval is thus 980.02 to ∞ . The analysis is illustrated in Fig. 85.

Maximum-likelihood Estimate of the Population Mean. Problem When the Population Variance Is Unknown. When the variance of the population is not known, the maximum-likelihood estimate of the population mean is based on the t distribution instead of the normal distribution. The procedure, however, is exactly the same. The optimum estimate of $\bar{X}$ is the value of $\bar{X}$ that makes the probability of the sample $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\hat{\sigma}}$ a maxi-

mum. Since the sampling distribution of this statistic is a t -distribution the peak of which always occurs at zero, the given

sample $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ will have the greatest probability of occurrence when $\bar{X}$ is such that $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma} = 0$, that is, when $\bar{X} = \bar{X}$. As in the previous case, therefore, the optimum estimate of the population mean is the value of the sample mean.

Use of Normal Curve with Large Samples. The foregoing procedure for testing hypotheses, finding confidence limits, and determining optimum estimates when the variance of the population is not known will give exact results whether the sample is large or small. If the sample is large, however, say 30 or more, approximate results may be obtained by using the normal curve in place of the t distribution. That is, for large samples, $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\sigma}$ becomes an $\frac{x}{\sigma}$ that can be looked up in a normal table. The basis for the procedure is that when N is large the t distribution is almost identical with the standard normal curve so that the latter can be used in its place. It is for this reason that the t table does not give values for $n > 30$.

SAMPLING DISTRIBUTION OF THE VARIANCE

Testing a Hypothesis. *The Problem.* The foregoing sections were concerned with the mean of the population. Consider now a problem that is concerned with the variance. Suppose that the inventor of the new process for the manufacture of electric-light bulbs claims that his process will reduce the variability in length of life of the bulbs. The process, it will be assumed, does not change the mean length of life but merely makes possible a more uniform and hence a more dependable product.

To make the problem concrete, suppose that the inventor claims that the standard deviation for the new electric-light bulbs is 180 kilowatt-hours, which represents, it will be supposed, a considerable reduction from the standard deviation of bulbs now in use. To test this claim a manufacturer produces 10 bulbs by the new process and finds that the standard deviation of this sample is 190 kilowatt-hours.¹ The question is: In view of the sample result, is the claim of a population standard deviation

¹ A small sample is taken to show that the analysis can be applied to small as well as to large samples. For a simpler method applicable only to large samples; see pp. 289-290.

of 180 kilowatt-hours a reasonable one? Or, to put this in terms of variances, with a sample variance of 36,100 square kilowatt-hours is the claim of a population variance of 32,400 square kilowatt-hours a tenable hypothesis?

Coefficient of Risk. Suppose, as in the case of the mean, that the manufacturer is willing to run the chance of rejecting hypotheses of this kind 5 times out of 100 when they are actually true. His coefficient of risk is thus .05.

Regions of Rejection. To assure a coefficient of risk of .05 the manufacturer must select from the whole set of samples that might be drawn from the given hypothetical population a subset of samples the probability of which is just .05. If he rejects the hypothesis whenever the given sample is found to belong to this special subset, he will attain a coefficient of risk equal to .05. This subset will be his region of rejection.

Since the manufacturer is interested here in the variance of the electric-light bulbs producible by the new process, the set of all possible samples and the chosen region of rejection can most appropriately be described in terms of the sample variance. When the set of all possible samples is so described, the result is the sampling distribution of the variance. This, as already noted, is a skewed distribution which centers roughly around the variance of the population—its mean is actually $\frac{N-1}{N}$ times σ^2 .^{*} If the unit of measurement is taken as σ^2/N , the distribution assumes the form of a χ^2 distribution, with n in the χ^2 equation equal to $N-1$.[†]

Since the manufacturer wishes the variance in bulbs to be as small as possible, he would like to reduce to a minimum the chance of accepting the inventor's claim when in actuality the variance of the new process is above the hypothetical figure. The most appropriate region of rejection for him to select is consequently the region comprising the upper .05 tail of the sampling distribution. A χ^2 table shows that, for

$$n = 10 - 1 = 9,$$

this comprises all values of χ^2 greater than 16.919, that is, all

^{*} For mean of $N\sigma^2/\sigma^2$ is $N-1$, since mean χ^2 is n . See $\dagger$.

[†] See pp. 111-112, 263-263.

samples whose variances are greater than $16.919\sigma^2/10$, or $1.6919\sigma^2$. This is illustrated in Fig. 86.

Testing the Hypothesis. The hypothesis to be tested in the given instance is that $\sigma^2 = 32,400$. The region of rejection thus constitutes all samples whose variances exceed

$$1.6919(32,400) = 54,818.$$

The given sample variance is 36,100, which is far from this region of rejection. Hence the given hypothesis is not rejected; the claim of the inventor is not disproved.

Other Examples. In testing hypotheses concerning the population variance, the statistician sometimes wishes to minimize

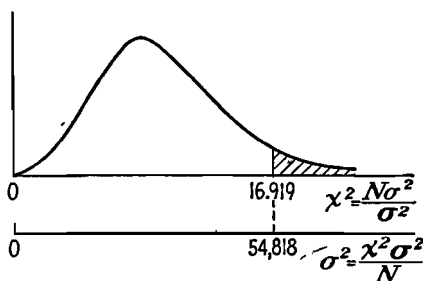


FIG. 86.

the risk of accepting the hypothesis when in fact the true variance is below the hypothetical figure being tested. In a college course, for example, it might be claimed that a new type of examination will give a better spread in grades and hence be a better examination for grading the students. In testing the hypothetical variance claimed for the new type of examination, the statistician would want to minimize the risk of accepting this claim when in fact the true variance in grades is less than this figure. In such a case, the region of rejection would most appropriately be chosen to constitute the lower 5 per cent of the sampling distribution.

For $N = 10$, for example, this would constitute all values of χ^2 less than 3.325, or all samples whose variances are less than $3.325\sigma^2/10$. This is illustrated in Fig. 87.

Still other cases might occur in which the statistician is indifferent as to whether the true variance is above or below the hypothetical variance being tested. In these cases, the most appropriate region of rejection would be one in which the total

probability of .05 is equally distributed at both ends of the distribution. Unfortunately, the χ^2 table has not been constructed so that the points marking the .025 tails of the distribution can be obtained without interpolation. If the coefficient of risk were set at .04, however, tails of .02 each could readily be determined from the existing table. Thus, if $n = 9$, a .04 region of

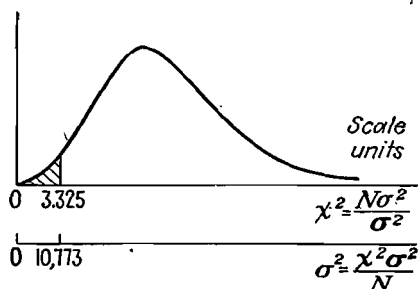


FIG. 87.

rejection could be taken to constitute all values of χ^2 less than 2.532 and all values of χ^2 greater than 19.697. This region of rejection would comprise all samples whose variances are less than $2.532\sigma^2/10$ and greater than $19.697\sigma^2/10$. Such a region of rejection would afford an even balance between the risk of accepting the hypothesis when the true variance is less than the hypo-

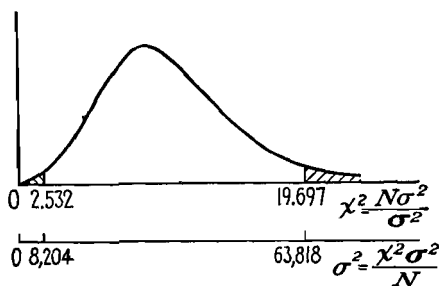


FIG. 88.

thetical figure and the risk of accepting it when the true variance is greater than this amount. This is illustrated in Fig. 88.

Confidence Limits for the Population Variance. *The Problem.* In many cases no particular hypothesis regarding the population variance is to be tested. Instead, it is merely desired to determine a range of values within which, on the basis of a given

sample, the variance of the population may reasonably be expected to lie. It is with the determination of such confidence limits that the present section is concerned.

To make the problem concrete suppose that 10 electric-light bulbs are manufactured by the new process referred to in the foregoing example and that the standard deviation in length of life of this sample is found to be 190 kilowatt-hours. This means a variance of 36,100 square kilowatt-hours. Suppose further that the manufacturer who is testing this new process wants to run no greater chance than 4 out of 100 that the confidence limits set up will fail to cover the true value of the variance. That is, he wishes confidence limits whose confidence coefficient is .96.*

Determination of Confidence Limits. If the manufacturer is interested in both an upper and a lower bound for the population variance, he can set up confidence limits as follows: He can obtain an upper limit for σ^2 by finding the value that makes the quantity $N\sigma^2/\hat{\sigma}^2$ just equal to the lower .02 point of the χ^2 distribution. For if this value is the true value, then the probability of the sample σ^2 or some lower value will be just .02. Similarly, a lower bound for σ^2 can be obtained by finding the value of $\hat{\sigma}^2$ that makes the quantity $N\sigma^2/\hat{\sigma}^2$ just equal to the upper .02 point of the χ^2 distribution. For if this value is the true value, then the probability of getting the sample or some higher value will be just .02. The values of $\hat{\sigma}^2$ so determined will constitute the confidence limits for this parameter and the confidence interval so marked off will have a probability of .96 of covering the true value.

For the given data these upper and lower limits for $\hat{\sigma}^2$ are determined as follows: For $n = 10 - 1 = 9$, the lower .02 point of the χ^2 distribution is 2.532. The upper limit for $\hat{\sigma}^2$ is thus given by $N\sigma^2/\hat{\sigma}^2 = 2.532$. This yields

$$\frac{(10)(36,100)}{\hat{\sigma}^2} = 2.532,$$

or $\hat{\sigma}^2 = 142,575$. Similarly, for $n = 9$, the upper .02 point of the χ^2 distribution is 19.697 and the lower limit for $\hat{\sigma}^2$ is given

* The confidence coefficient is put at this figure instead of the usual .95 because the χ^2 table gives the .02 points at each end instead of the usual .025 points.

by $N\sigma^2/\delta^2 = 19.697$, or $(10)(36,100)/\delta^2 = 19.697$. This yields $\delta^2 = 18,328$. The confidence interval is thus 18,328–142,575, and it may be said that there is a probability of .96 that this interval covers the true variance.

The foregoing interval was an unbiased confidence interval in that the probability of failing to cover the population value because the interval was set too low was equal to the probability of failing to cover the population value because the interval was set too high. If the manufacturer had wished to set up a confidence interval that had only an upper bound, he could have found that upper bound by setting $N\sigma^2/\delta^2$ equal to the .95 point of the χ^2 distribution for $n = N - 1$. For the given data this yields $10(36,100)/\delta^2 = 3.325$, or $\delta^2 = 108,571$. Hence the manufacturer might say that the range 0–108,571 has a probability of .95 of covering the true value. That is, in cases of this kind the manufacturer would go wrong only 5 per cent of the time in assuming that the population variance was equal to or below the upper bound. It will be noted here that the confidence interval established in this instance had a confidence coefficient of .95 instead of .96 as in the previous example. The former was adopted because the χ^2 table lacks a .96 point.

In conclusion, it should be noted once again that, the larger the sample, the narrower the confidence interval. If a sample of 20 had been taken instead of a sample of 10, the limits for δ^2 would have been 21,433 and 84,277. Thus, the larger the sample, the greater the assurance that the sample variance is near the true variance. It must be remembered, however, that a larger sample may involve increased expense.

SPECIAL METHOD FOR LARGER SAMPLES

As indicated in the preceding chapter, when the sample is large the sampling distribution of the variance is approximately normal, its mean being practically the mean of the population

(more exactly, $\frac{N}{N-1} \sigma^2$), and its standard deviation being $\sigma^2 \sqrt{\frac{2}{N}}$. Hence for large samples, say samples greater than 30, any hypothetical value of σ^2 may be tested by noting the value of $\frac{\sigma^2 - \delta^2}{\delta^2 \sqrt{2/N}}$. If a symmetrical region of rejection is adopted

with a probability of .05, then all values of $\frac{\sigma^2 - \delta^2}{\delta^2 \sqrt{2/N}}$ lying outside of ± 1.96 will lead to a rejection of the hypothesis. If the region of rejection is put at one end, the hypothesis will be rejected if $\frac{\sigma^2 - \delta^2}{\delta^2 \sqrt{2/N}}$ is greater than 1.645 in one instance or if it is less than -1.645 in another instance. Similarly, confidence limits with a coefficient of .95 may be determined by setting $\frac{\sigma^2 - \delta^2}{\delta^2 \sqrt{2/N}} = \pm 1.96$ if both an upper and lower bound is desired, or to 1.645 if a lower bound only is desired, or to -1.645 if an upper bound only is desired.

To illustrate this special method for large samples, suppose that the batch of electric-light bulbs of the previous sample had numbered 98 instead of 10. Let the sample variance be 36,100 as before, and consider the hypothesis that the true variance is 32,400 (that is, $\delta = 180$). If the region of rejection is taken as the upper .05 tail of the normal curve, then this hypothesis

may be tested by finding the value of $\frac{\sigma^2 - \delta^2}{\delta^2 \sqrt{2/N}}$. For the given

data this is equal to $\frac{36,100 - 32,400}{32,400 \sqrt{\frac{2}{98}}} = .8$. Since this is less than 1.645, the sample does not fall in the region of rejection and the hypothesis is not rejected.

To determine confidence limits for the population variance in this instance, the procedure is as follows: Let the confidence coefficient be .96 so that the result will be comparable with that of the preceding section. Also, let the upper bound of the confidence interval be determined so that if the population variance has a value equal to this upper bound then the probability of the sample result or a lower value is just .02, and let the lower bound be determined in the same manner. Then these

upper and lower limits will be given by $\frac{36,100 - \delta^2}{\delta^2 \sqrt{\frac{2}{98}}} = \pm 2.054$.

This yields $\delta^2 = 27,911$ and $\delta^2 = 51,090$ as the limits of the confidence interval. It will be noted that this interval is still smaller than that for which $N = 10$ or $N = 20$.

Maximum-likelihood Estimate of the Population Variance.
So far maximum-likelihood estimates of population parameters

have been found to be the sample values of the corresponding statistics. Thus the maximum-likelihood estimate of the percentage of "favorable" cases in the population is the percentage of "favorable" cases in the sample. Similarly, whether or not the variance of the population is known, the maximum-likelihood estimate of the mean of the population is the mean of the sample. In the present instance, however, the maximum-likelihood estimate of the population variance that can be made independently of the sample mean¹ is not the sample variance but the sample variance multiplied by $\frac{N}{N-1}$. The derivation of this result is as follows:

By definition the maximum-likelihood estimate of a population parameter is the value of the parameter that will make the

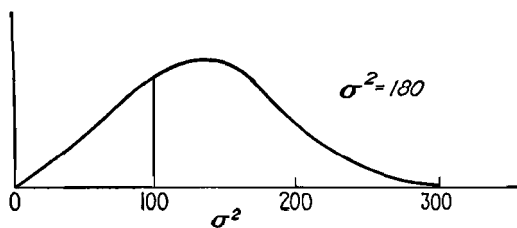


FIG. 89a.

probability of the given sample a maximum. For an assigned value of the population variance the probability of a sample having any given variance may be determined from the sampling distribution of the variance. The equation for this distribution is Eq. (3) of Chap. X. If a particular value is given to σ^2 in this equation, it will yield the probability of getting samples with various values of σ^2 . If, however, the value of a given sample variance is substituted in this equation, then the formula yields the probability of getting this particular sample for various possible values of σ^2 . Figure 89a, for example, shows the probability of a sample variance of 100 when the population variance is 200, the size of the sample being 11. Figure 89b shows (for $N = 11$) the probability of a sample variance of 100 when the population variance is 80, and Fig. 89c shows this

¹ When the mean and standard deviation are estimated jointly then the maximum-likelihood estimate of the population variance is also the variance of the sample.

probability when the population variance is equal to $\frac{N\sigma^2}{(N-1)}$, that is, to 110. Further calculations will show that the probability of the given σ^2 is greater when σ^2 has this last value than when it has any other value.¹ This is indicated by the curve shown in Fig. 90. The optimum estimate of σ^2 when $\sigma^2 = 100$ and $N = 11$ is thus 110.

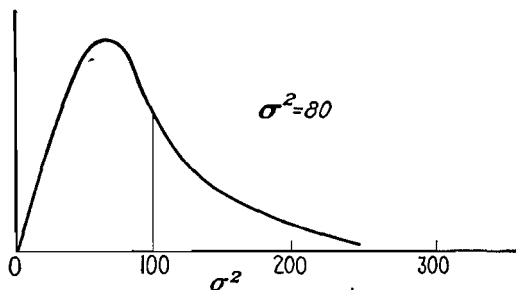


FIG. 89b.

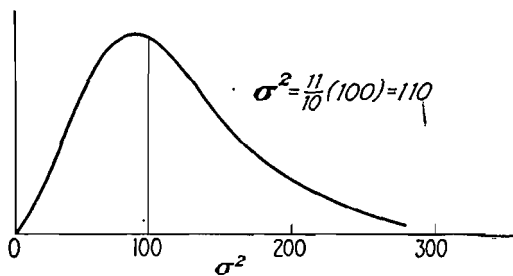


FIG. 89c.

The algebraic counterpart of this geometrical explanation is as follows: For a given σ^2 , Eq. (3) of Chap. X gives the probability of that σ^2 as a function of σ^2 . If this probability is to be a maximum for σ^2 , then the derivative of the function with respect to σ^2 must be zero at this maximizing value. To simplify

¹ For simplicity the argument speaks of "the probability of the given σ^2 "; but a more precise expression would be "the probability of a sample variance lying between the given σ^2 and the given $\sigma^2 + d\sigma^2$." Again, "the probability of a sample variance of 100" is merely a short way of saying "the probability of a sample variance lying between 100 and $100 + d\sigma^2$." Although the true probability is an area, it is only the ordinate that changes from case to case in this argument and is therefore all that needs to be considered here.

matters the logarithm of the probability is differentiated instead of the probability itself, but this does not alter the result since the probability will be a maximum when its logarithm is a maximum. Thus, from Eq. (3), Chap. X,

$$\log [P(\sigma^2)] = -\frac{N\sigma^2}{2\sigma^2} - \frac{N-1}{2} \log \sigma^2 + \text{terms not involving } \sigma^2.$$

Differentiating this with respect to σ^2 and setting the result equal to 0 yields the following:

$$\frac{d \log [P(\sigma^2)]}{d\sigma^2} = \frac{N\sigma^2}{2(\sigma^2)^2} - \frac{N-1}{2\sigma^2} = 0$$

or

$$\sigma^2 = \frac{N}{N-1} \sigma^2 \quad (1)$$

It may seem strange at first thought that the optimum value of σ^2 is not the value that makes the given σ^2 fall at the mode

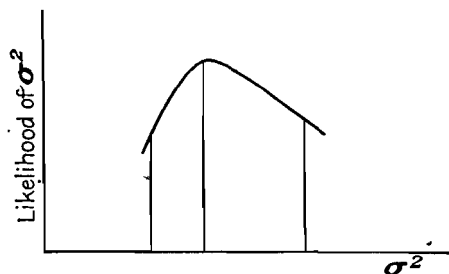


FIG. 90.

or peak of the sampling distribution. On more careful examination, however, it is seen that the sampling distribution of σ^2 changes not only its position but also its shape with changes in the value assigned to σ^2 . The height at the mode of one curve may thus be less than the height at some other point on another curve. In the present instance it turns out that the height of the curve given by $\sigma^2 = \frac{N}{N-1} \sigma^2$ at the point σ^2 is greater than the maximum height of the curve whose mode occurs at σ^2 .

The explanation of this somewhat unexpected result for the maximum-likelihood estimate of σ^2 lies in the fact that the

variance of a sample is always measured from the mean of the sample and not the mean of the population. Since the mean itself varies from sample to sample, any estimate of the population variance that ignores this variation in the mean will tend to underestimate its true value. It will be seen in Chap. XIV that, when the mean of the population and the variance of the population are estimated jointly, the joint maximum-likelihood estimates are $\bar{X} = \bar{X}$ and $\hat{\sigma}^2 = \sigma^2$.

SAMPLING DISTRIBUTION OF THE RANGE

Table XIII of the Appendix¹ giving pertinent facts regarding the sampling distribution of the range may be used to estimate the population standard deviation in a quick and ready manner; for the values given in the table are for $w = \frac{X_n - X_1}{\delta}$, and also the mean values for w in samples of varying size are given. On the assumption that a given sample range is close to the mean range for samples of the given size, it follows that $\frac{X_n - X_1}{w}$ will give a rough estimate of δ . Suppose, for example, that the range of a sample of 10 cases is 50, then $50/3.078$ or

$$50(.325) = 16.25$$

is a crude estimate of δ . Furthermore, by using the .025 points of Table XIII it will be seen that the chances are 95 out of 100 that the interval from $50/4.79$ to $50/1.67$, that is, from 10.44 to 29.94, includes the standard deviation of the population.

If all the data are available, good results involving little labor can be obtained by breaking up a large sample into a number of smaller samples of the same size, calculating the range for each of these samples, taking the average of these sample ranges, and using this average range to establish limits for the population standard deviation. Suppose, for example, that 10 samples of eight cases each showed ranges of 30, 24, 16, 28, 11, 20, 22, 35, 17, 25. The mean of these 10 sample ranges is 22.8. Table XIII shows that, if the population standard deviation is δ , the mean range for samples of 8 is 2.847δ . It also shows that the standard error of the range is $.820\delta$. Hence the stand-

¹ See p. 482. Note X_n = largest, X_1 = smallest case.

ard error of a mean of 10 sample ranges would be $.820/\sqrt{10}$. In the present instance, therefore, the .95 confidence interval for σ is given by $\frac{\sqrt{10}(22.8 - 2.847\sigma)}{.820\sigma} = \pm 1.96$ (the latter being the upper and lower .025 probability points of the normal curve). The confidence interval for σ is accordingly 6.8 to 9.7.

The sampling distribution of the range may also be used for quick analysis in certain problems in which the variation in means of particular groups is being tested for significance.¹ For example, suppose that five classes of 12 students each have average grades of 77.25, 77.83, 70.92, 69.92, and 74.00 and the ranges of individual grades for the five classes are 31, 34, 24, 44, and 44 points. On the basis of the variation of grades indicated by the ranges, is the variation in mean grades more than might reasonably be attributed to chance? This is the question that may readily be answered as follows:

The mean of the five ranges is a mean range of 35.4 points. From Table XIII of the Appendix it is seen that for groups of 12 the mean range is 3.258 σ . Hence $35.4/3.258 = 10.86$ may serve as a rough estimate of the population standard deviation σ . Now, if $\sigma = 10.86$, then means of 12 grades should have a $\sigma = 10.86/\sqrt{12} = 3.135$ and Table XIII of the Appendix shows that for groups of five the mean range for a variable whose σ is 3.135 is $2.326\sigma = 2.326 \times 3.135 = 7.292$. For the five mean grades given above the range is 7.91 and is thus just about what would be expected on the basis of chance. Hence the analysis of the variation in mean grades based on the ranges of grades in the individual classes indicates that the variation in means is apparently no greater than may reasonably be attributed to chance. This is identical with the conclusion reached in Chap. XVII after a more refined analysis of variance. The rougher analysis presented here is especially valuable as a preliminary analysis for throwing out those cases in which the variation in means is shown by the rougher analysis to be clearly not due to chance.

The various percentage points of the sampling distribution of the range may also be useful in setting up charts to control

¹ See SMITH, J. G., and A. J. DUNCAN, *Elementary Statistics and Applications*, Chap. XV, for discussion of correlation ratio. Also, see Chap. XVIII below on Analysis of Variance.

the quality of industrial mass production. An interesting discussion of the use of the range in industrial quality control may be found in reports issued by the British Standards Institution.¹

Use of Sampling Distributions of β_1 , β_2 (or g_1 , g_2) and a to Test for Departure from Normality. As indicated in the previous chapter, the sampling distributions of β_1 and β_2 (or g_1 and g_2 if the worker prefers these more refined statistics) and of $a = A.D./\sigma$ may be used to test departure from normality. For illustrative purposes consider the data on the weights of 300 Princeton freshmen² discussed in Chap. VII.

For these data,

$$\sqrt{\beta_1} = .637, \quad \beta_2 = 4.6720, \quad g_1 = .635, \quad g_2 = 1.7320$$

and $a = .7640$. Tests of departure from normality may be carried out as follows.

For samples of 300,

$$\sigma_{\sqrt{\beta_1}} \doteq \sqrt{\frac{6}{N}} = .141$$

$$\sigma_{\beta_2} \doteq \sqrt{\frac{24}{N}} = .283$$

and the ratios of $\sqrt{\beta_1}$ and $\beta_2 - 3$ to their standard errors are

$$\frac{\sqrt{\beta_1}}{\sigma_{\sqrt{\beta_1}}} = \frac{.637}{.141} = 4.5$$

$$\frac{\sqrt{\beta_2 - 3}}{\sigma_{\beta_2}} = \frac{4.6720 - 3}{.283} = 5.9$$

Both these indicate a marked departure from normality in that the probability of either $\sqrt{\beta_1}$ or $\beta_2 - 3$ exceeding their standard errors by as much as 4.5 to 5.9 times is very small. For samples as large as 300, g_1 and g_2 and their standard errors are almost identical with $\sqrt{\beta_1}$ and $\beta_2 - 3$ and their standard errors. In the case in hand, $\sigma_{g_1} = .141$, $\sigma_{g_2} = .281$, $g_1/\sigma_{g_1} = .635/.141 = 4.5$, and $g_2/\sigma_{g_2} = 1.7320/.281 = 6.2$; therefore, the conclusions are

¹ See DUDGING, B. P., and W. J. JENNETT, *The Application of Statistical Methods to Quality Control*, British Standards Institution, No. 600 (1942); and PEARSON, E. S., *The Application of Statistical Methods to Industrial Standardization and Quality Control*, British Standards Institution, No. 600 (1935).

² See pp. 138, 141.

not modified by the use of the more accurate standard errors of g_1 and g'_2 . Table XI of the Appendix also shows¹ that the sample value of $\beta_2 = 4.6720$ is considerably beyond the upper 1 per cent point, and Table XII of the Appendix shows² that the sample value of $\alpha = .7640$ is below the lower 1 per cent point for α . Hence all the tests indicate a distinct departure from normality.

¹ See p. 480.

² See p. 481.

CHAPTER XII

SAMPLING FLUCTUATIONS IN CORRELATION STATISTICS

SAMPLING DISTRIBUTION OF THE CORRELATION COEFFICIENT r

Correlation coefficients have sampling distributions that are different from any yet discussed. Thus, if a large number of samples are taken at random from a normal bivariate population,¹ and if a frequency distribution is made of the sample values of r_{12} , the shape of this distribution will be found to conform to the equation²

$$dP(r_{12}) = \frac{(1 - \mathbf{r}_{12}^2)^{N-1}}{\pi(N-3)!} (1 - r_{12}^2)^{\frac{N-4}{2}} \cdot \frac{d^{N-2}}{d(\mathbf{r}_{12}r_{12})^{N-2}} \left(\frac{\cos^{-1}(-\mathbf{r}_{12}r_{12})}{\sqrt{1 - \mathbf{r}_{12}^2 r_{12}^2}} \right) dr_{12} \quad (1)$$

where $\mathbf{r}_{12}$, in boldface type, refers to the population correlation coefficient and $\frac{d^{N-2}}{d(\mathbf{r}_{12}r_{12})^{N-2}} \left(\frac{\cos^{-1}(-\mathbf{r}_{12}r_{12})}{\sqrt{1 - \mathbf{r}_{12}^2 r_{12}^2}} \right)$ means the $N - 2$

¹ A sample from a monovariate or univariate population consists of a group of individual values; a sample from a bivariate population consists of a group of pairs of values. For example, a sample from the univariate population consisting of white male heights might be represented by the measurements 68, 70, 69, 74, 70, 71, 72, 71, 68, 69 inches, each of which represents the height of a particular individual. A sample from the bivariate population consisting of white male heights and weights might be represented by pairs of measurements:

68	70	69	74	70	71	72	71	68	69	inches
140	150	149	206	165	183	190	175	154	142	pounds

each pair consisting of the height and weight of a particular individual. It is to be noted that each of the above samples is said to be a sample of 10, although the second contains 20 measurements. They are each a sample of 10 in the sense that measurements are made only of 10 individuals.

² Cf. FISHER, R. A., "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample," *Metron*, Vol. 1 (1920-1921), Part 4, pp. 1-32.

derivative of the expression in parentheses with respect to r_{12} . It is to be especially noted that the shape of this distribution depends not only on N , the size of the samples, but also on the population coefficient of correlation, r_{12} . For small values of r_{12} the distribution is practically symmetrical;¹ for large absolute values of r_{12} , however, the distribution is very skewed. This is illustrated in Fig. 91. The character of the sampling distribution of r is in part related to the fact that r_{12} and r_{12} must always be between -1 and $+1$. Hence if r_{12} is close to $+1$, say .94, then the range of values above r_{12} that might possibly be taken by the sample r_{12} is very small compared with the range

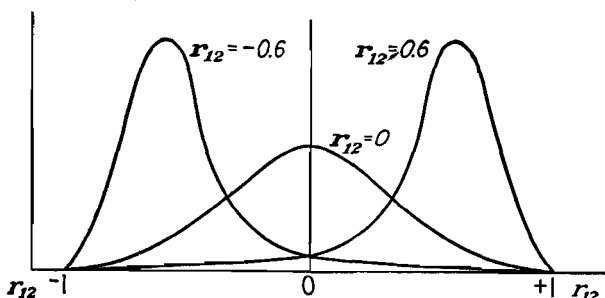


FIG. 91.—Sampling distributions of r_{12} , for $r_{12} = -0.6$, $r_{12} = 0$, and $r_{12} = 0.6$.

of values that might be taken by r_{12} below r_{12} , and vice versa, if r_{12} is close to -1 . Even for large values of N the distribution of r_{12} remains very skewed when r_{12} is close to $+1$ or -1 .

Fortunately it is possible to “transform” the sampling distribution of r_{12} into a more useful form. Thus, R. A. Fisher has shown that if z_{12} is defined as $\frac{1}{2}[\log_e(1 + r_{12}) - \log_e(1 - r_{12})]$, that is, as $\tanh^{-1} r_{12}$ (read “inverse hyperbolic tangent of r_{12} ”), then for all but very small values of N the sampling distribution of z_{12} is practically normal in form, with a mean approximately equal to the population value, z_{12} , that is, to

$$\frac{1}{2}[\log_e(1 + r_{12}) - \log_e(1 - r_{12})],$$

and a standard deviation equal approximately to $1/\sqrt{N - 3}$.*

¹ For $r_{12} = 0$ it is perfectly symmetrical.

* The distribution of z may be obtained by substituting

$$r_{12} = \frac{e^{2z} - 1}{e^{2z} + 1}$$

in Eq. (1). The mean of the distribution of z is always, except when $r_{12} = 0$,

Accordingly, if this "z transformation" is used, a sampling distribution can be obtained that for all but very small samples is practically independent of the population correlation coefficient and is moreover approximately normal in form. The z transformation thus makes it unnecessary to use elaborate tables of the distribution of r_{12} in making estimates and testing hypotheses regarding the population correlation coefficient.

Again it is to be noted that all the foregoing depends on the assumption that the population of X_1 and X_2 values is distributed in the form of a joint normal frequency distribution.¹

USE OF THE SAMPLING DISTRIBUTION OF $z = \tanh^{-1} r$

In testing hypotheses and setting up confidence limits regarding correlation coefficients the most practicable procedure is to make use of the z transformation. For, as pointed out above, the sampling distribution of z is practically normal in form, and tables of the normal distribution are generally available. Since z is normally distributed, all the previous discussion regarding any normally distributed statistic applies here, such as, for example, the discussion of the sampling fluctuations of the mean when the population standard deviation is known. The following discussion will therefore be primarily devoted to illustrating the use of the z transformation itself.

Testing a Hypothesis. In Smith and Duncan's *Elementary Statistics and Application* the simple correlation coefficient between the standing of Mount Holyoke students in first-semester English and their standing in second-semester English was found to be $r_{12} = .89576$. A table of hyperbolic tangents (see Appendix, Table XIV) shows that the angle for which'

slightly greater in absolute value than the population value of z_{12} , the "bias" being given roughly by $\frac{r_{12}}{2(N-1)}$.

A more accurate formula that indicates the closeness of the approximation for the standard deviation is as follows:

$$\sigma_{z_{12}}^2 = \frac{1}{N-3 + \frac{8 + (N-3)r_{12}^2}{2N+2-r_{12}^2}}$$

¹ For nonnormal cases see discussion, p. 451.

.89576 is the hyperbolic tangent (it will be recalled that

$$r_{12} = \tanh z_{12},$$

for $z_{12} = \tanh^{-1} r_{12}$) is approximately 1.4506. If tables are not available, z may be calculated from the equation

$$\begin{aligned} z_{12} &= \frac{1}{2} [\log_e (1 + r_{12}) - \log_e (1 - r_{12})] \\ &= \frac{1}{2(.43429)} [\log_{10} (1 + r_{12}) - \log_{10} (1 - r_{12})] \end{aligned}$$

For the given case this yields $z = 1.45049$, which checks to three places with the value derived from the table.¹ Here the sample value of z_{12} will be taken as that given by the table, *viz.*,

$$z_{12} = 1.4506.$$

Suppose it is argued that the true correlation coefficient between first- and second-semester English grades is .9900. How does this hypothesis fare with reference to the sample coefficient of .89576? To test this hypothesis convert both these correlation coefficients to their corresponding z values. The table of hyperbolic tangents indicates that $z = \tanh^{-1} .9900 = 2.6465$ and, as previously, $z = \tanh^{-1} .89576 = 1.4506$. Hence the hypothetical z is 2.6465, and the sample z is 1.4506. The standard deviation of $z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{81-3}} = .11323$. Since z is normally distributed, symmetrical regions of rejection would be given by the hypothetical z plus or minus $1.96\sigma_z$, or in this particular case by $2.6465 \pm .22193$, or 2.8684 and 2.4246. Since the sample value falls below 2.4246, the hypothesis would not be accepted.

Confidence Interval for r_{12} . In order to determine a confidence interval for r_{12} from the given sample, it is necessary merely to determine a confidence interval for the population z_{12} and convert this into a value for r_{12} . Thus, suppose that symmetrical confidence limits are desired (symmetrical in the sense that failure to cover the true value is equally probable at one end and at the other). For z_{12} these are given by the sample $z_{12} \pm 1.96\sigma_z$. In the present instance,

¹ The difference is due to the difference in decimals carried, methods of interpolation, etc., and not to any difference in definition.

$$\delta_z = \frac{1}{\sqrt{N} - 3} = \frac{1}{\sqrt{81} - 3} = .11323, \text{ and}$$

$$1.96\delta_z = .22193.$$

Accordingly, the upper confidence limit for z_{12} is

$$1.4506 + .22193 = 1.6725,$$

and the lower confidence limit for z_{12} is

$$1.4506 - .22193 = 1.2287.$$

Interpolating in a table of hyperbolic tangents shows that the hyperbolic tangent of 1.6725 is .93188 and the hyperbolic tangent of 1.2287 is .84220. These values may also be obtained from the inverse of the equation for the z transformation given above, *viz.*,

$$r_{12} = \frac{e^{2z_{12}} - 1}{e^{2z_{12}} + 1}$$

Thus for the upper limit, $2z_{12}$ equals 3.3450; $\log_{10} e^{3.3450}$ equals 3.3450(.43429) = 1.45270; and $e^{3.3450}$ equals the antilogarithm of this, or 28.359. Consequently, the upper limit for r_{12} equals $27.359/29.359 = .93188$, which checks the interpolation in the table. Similar calculations would check the .84220 lower limit for r_{12} .

In summary, the confidence limits for r_{12} are .84220 and .93188; it may thus be concluded that the range .84220-.93188 covers the true value of r_{12} with a probability of .95. To put it another way, any hypothetical value of r_{12} that is below .84220 or above .93188 is deemed an unreasonable value in view of the sample value $r_{12} = .89576$.

Confidence limits could also be set up and hypotheses could be tested in such a way that the region of rejection all came at one end. All this would be a duplication of the discussion of Chap. XI, however, and need not be repeated here. As mentioned above, once r_{12} has been transformed into z_{12} , all the analysis pertaining to normally distributed variables can be applied.

Optimum Estimate of the Population Correlation Coefficient.

An estimate of the population correlation coefficient that is independent of the estimates of the means and variances of the population can be made from the sampling distribution of

r_{12} . The independent maximum likelihood estimate of r_{12} is obtained by so choosing r_{12} that the probability of the sample r_{12} , as indicated by Eq. (1), is a maximum.¹ This yields the approximate result²

$$\check{r}_{12} = r_{12} - \frac{r_{12}(1 - r_{12}^2)}{2(N - 1)} \quad (2)$$

That is, the sample r_{12} is somewhat too high as an estimate of the population correlation coefficient and needs to be corrected by the subtraction of the quantity $\frac{r_{12}(1 - r_{12}^2)}{2(N - 1)}$. For example, if $r_{12} = .89576$ and $N = 81$, then the maximum-likelihood estimate of r_{12} is $.89576 - \frac{(.89576)(1 - .89576^2)}{2(81 - 1)} = .89465$.

When r_{12} is estimated jointly with the means and variances of the population, it is found that the maximum-likelihood estimate of r_{12} is the sample r_{12} .

SAMPLING DISTRIBUTION OF OTHER CORRELATION COEFFICIENTS

Partial Correlation Coefficients. The analysis that has been described in the preceding section for a simple correlation coefficient can be applied with very little modification to partial correlation coefficients. The only change required in the formulas is to subtract from N the number of variables held constant. As before, the practicable procedure is to take $z = \tanh^{-1} r_{ij.k} \dots$ and to treat z as if it were normally distributed. The mean of the distribution of z is again approximately the value of z corresponding to the population correlation coefficient, $r_{ij.k} \dots$, that is, $z = \tanh^{-1} r_{ij.k} \dots$; but the standard deviation of z is now the reciprocal of the square root of $N - m - 3$, where m is the number of variables held constant.

¹ See pp. 181-182.

² This is obtained by taking the derivative of expression (1) with respect to r_{12} and setting the result equal to zero. The formula is only a first approximation. Better approximations may be found in Karl Pearson's *Tables for Statisticians and Biometricians*, Vol. II, p. 253. Another approximate equation that is commonly used is

$$\check{r}^2 = \frac{r^2(N - 1) - 1}{N - 2}$$

To illustrate, consider the determination of confidence limits for the $r_{12.34}$ of the Mount Holyoke data. For these it was found that $r_{12.34}$ was .80440. For this value of r , the value of z is 1.1110. The standard deviation of z in this instance is

$$\frac{1}{\sqrt{81 - 2 - 3}} = .1147.$$

Accordingly, symmetrical .95 confidence limits for z are

$$1.1110 + 1.96(0.1147)$$

and $1.1110 - 1.96(0.1147)$, or 1.3358 and 0.8862. The $r_{12.34}$ limits corresponding to these two z limits are, respectively, .8706 and .7095.

Hypotheses regarding $r_{12.34}$ can be tested and estimates can be made in the same manner as hypotheses and estimates pertaining to simple correlation coefficients except that the number of variables held constant (m) must always be subtracted from N in all the equations used.

Multiple Correlation Coefficients. When the population multiple correlation coefficient is zero, the statistic

$$\frac{R^2/(k - 1)}{(1 - R^2)/(N - k)}$$

has a sampling distribution of the form of the F distribution, with $n_1 = k - 1$ and $n_2 = N - k$. Here R is the multiple correlation coefficient, k is the number of regression statistics $a_{1.234}, b_{12.3}, \dots, b_{13.2}, \dots$, etc., in the regression equation, and N is the size of the sample. Hence the F distribution can be used to test whether a given multiple correlation coefficient is significantly different from zero.

For the Mount Holyoke data, for example, it was found that $R^2_{4.123} = .2529$, and the question arises whether this is significantly different from zero. To answer this the hypothesis that the population value is zero (the "null hypothesis" as it may be called) is set up and tested as follows: The number of independent variables is three, and the number of regression statistics is one more than this, or four. Hence, $k = 4$ and

$$\frac{R^2}{k - 1} = \frac{.2529}{3} = .0843.$$

Since $N = 81$, $\frac{1 - R^2}{N - k} = \frac{.7471}{77} = .00970$. Therefore,

$$\frac{R^2/(k - 1)}{(1 - R^2)/(N - k)} = \frac{.0843}{.0097} = 8.69$$

The F table (see Appendix, page 476) shows that, for $n_1 = 3$ and $n_2 = 60$, the .05 point is 2.758 and, for $n_1 = 3$ and $n_2 = \infty$, the .05 point is 2.605. Accordingly, the .05 point for $n_1 = 3$ and $n_2 = 77$ lies between 2.758 and 2.605. Since the sample value is much greater than this, the null hypothesis is obviously rejected and the multiple correlation coefficient must be declared significantly different from zero. That is, there must be some multiple correlation between the variables concerned.

When the population correlation coefficient is not zero, the distribution of the multiple correlation coefficient is more complicated. This was worked out by R. A. Fisher in 1928.¹ Using Fisher's analysis, Mordecai Ezekiel has drawn a series of charts that give the lower confidence limits for the population multiple correlation coefficient for samples of varying size and varying numbers of independent variables. The confidence coefficient used was .95. These charts are to be found in the Appendix to Ezekiel's book on *Methods of Correlation Analysis*. For $N = 75$ and $k = 4$, for example, Ezekiel's Chart C shows that, if the sample multiple correlation is .57, then the lower confidence limit for the population value is .40. For the Mount Holyoke data the various multiple correlation coefficients are $R_{1.234} = .9056$, $R_{2.134} = .8983$, $R_{3.124} = .6326$, and $R_{4.123} = .5029$. For each of these, $N = 81$, and $k = 4$. Ezekiel's Chart C shows that the lower confidence limit for these are approximately .85, .85, .49, and .32, respectively. In other words, the chance is .95 that the ranges .85-1, .85-1, .49-1, and .32-1 cover the population value in each case.

The maximum-likelihood estimate of the multiple correlation coefficient is given by the approximate formula

$$\hat{R}^2 = \frac{R^2(N - 1) - (k - 1)}{N - k}$$

¹ *Proceedings of the Royal Society of London, Series A* (1928), Vol. 121, pp. 654-673.

Thus the maximum-likelihood estimate of $R_{1,234}^2$ is

$$\frac{.9056(80) - 3}{77} = .9019$$

Correlation Ratio. If the correlation ratio for the population is zero, then the statistic $\frac{\eta_{12}^2/(k-1)}{(1-\eta_{12}^2)/(N-k)}$ also has a sampling distribution of the form of an F distribution with $n_1 = k-1$ and $n_2 = N-k$. Here k stands for the number of means entering into the calculation of η_{12} .

To illustrate the use of the sampling distribution of η_{12}^2 consider again the Mount Holyoke data. For these, $\eta_{12}^2 = .82124$, $k = 11$, $N = 81$, and thus

$$\frac{.82124/10}{.17873/70} = \frac{.082124}{.002553} = 32.168$$

To test the hypothesis that the true correlation ratio is 0, take a coefficient of risk equal to .05, and take the upper .05 tail as the region of rejection. For this case, $n_1 = 11 - 1 = 10$, and $n_2 = 81 - 11 = 70$. For $n_1 = 8$ and $n_2 = 60$, the F table shows that the upper .05 point is 2.097; and, for $n_1 = 8$ and $n_2 = \infty$, the upper .05 point is 1.938. For $n_1 = 12$ and $n_2 = 60$, the upper .05 point is 1.918; and, for $n_1 = 12$ and $n_2 = \infty$, the .05 point is 1.752. Interpolation for $n_1 = 10$ and $n_2 = 70$ is unnecessary since it is obvious that 32.168 is far beyond the upper .05 point for this problem and the correlation ratio is certainly significantly different from zero.

Correlation Index. If a curve such as a parabola is fitted to the original data and the correlation index for the population is zero, then the statistic $\frac{I^2/(k-1)}{(1-I^2)/(N-k)}$ is likewise distributed like F with $n_1 = k-1$ and $n_2 = N-k$, k being the number of regression statistics in the equation for the curve. When logarithms or reciprocals are used to make a distribution linear, the correlation between the logarithms or reciprocals can be treated as any ordinary linear correlation coefficient.

Test for Linearity. Since a correlation ratio and a correlation index are always greater than a correlation coefficient calculated for the same data, the question may be asked: How can it be determined when a distribution is really linear and when non-

linear? This question can be answered by a sampling test. Thus the null hypothesis is set up that the population is linear and this hypothesis tested in the light of the difference between the correlation coefficient and the correlation ratio (or the correlation index) for the given set of sample data.

If the population is a normal bivariate distribution in which the progression of means is actually linear, then the statistic

$$\frac{(\eta^2 - r^2)/(k - 2)}{(1 - \eta^2)/(N - k)} \quad \text{or the statistic} \quad \frac{(I^2 - r^2)/(k - 2)}{(1 - I^2)/(N - k)}$$

is distributed like F with $n_1 = k - 2$ and $n_2 = N - k$, where k is the number of means from which the correlation ratio is computed or in the case of I the number of regression statistics in the equation of the curve.

For the Mount Holyoke example a correlation index was not computed; but the correlation ratio was calculated, and according to the above procedure a test for linearity can be made as follows:

$$\begin{array}{llll} r_{12}^2 = .80239 & \eta_{12}^2 = .82123 & k = 11 & N = 81 \\ & & n_1 = 9 & n_2 = 70 \end{array}$$

Hence,

$$\frac{\frac{.82123 - .80239}{9}}{\frac{.17877}{70}} = \frac{.00209}{.00255} = .8196$$

The .05 point in the F distribution for $n_1 = 9$ and $n_2 = 70$ is between 1.752 and 2.097. Thus .8196 is well within that limit, and correlation is presumably linear.

PART III

Advanced Sampling Problems

CHAPTER XIII

SAMPLING FROM A DISCRETE MANIFOLD POPULATION

Chapter IX was concerned with sampling from a discrete twofold population. It is the purpose of the present chapter to extend the discussion of this earlier chapter to a discrete manifold population, a population that is divided into more than two classes. The theory of sampling from a discrete manifold population has widespread application, for in many discrete populations the cases are grouped into several classes. Furthermore, the cases of any continuously distributed population can be, and commonly are, arbitrarily grouped into a finite number of classes.

The theoretical argument pertaining to manifold populations is fundamentally an extension of the argument pertaining to a twofold population. The ensuing analysis will accordingly pursue the same line of attack as that of the earlier chapter. The argument will be presented in full; but those parts that are identical with the earlier argument will be reviewed only briefly, and attention will primarily be centered on the complications that arise from a manifold instead of a twofold division of the population.

THEORETICAL ARGUMENT

Fundamental Assumptions. The fundamental assumptions are the same as in the simpler case. It is first assumed that the sample is small relative to the population from which it is drawn. It can therefore be assumed that the percentages of the cases belonging to the various classes of the population are not changed by the withdrawal of the sample values. This will, of course, be only approximately true if the population is actually finite and will be perfectly true only for an infinite population.

Although the sample is assumed to be small relative to the population, it is nevertheless assumed to be sufficiently large to permit the use of certain approximations. In some of the theoretical illustrations very small samples are used for the sake of simplicity, but in the discussion of the practical application of the argument it is assumed that the samples are large in the aggregate. As a general rule, it is assumed that the product of the population percentage of any class times the size of the sample (that is, $p_i N$) is at least equal to 5.

A second fundamental assumption is that the method of sampling is a random one. This implies once again that the results of repeated sampling by the selected method can be predicted by the calculation of probabilities for some mathematical model. The appropriate model will be described in the next section.

In order to make the analysis concrete, suppose the problem is that of estimating the percentages of various types of religious adherents in a given locality. The alternatives are taken to be Catholic, Protestant, and other religious denominations, including atheists, and the problem will be to estimate by means of a random sample the percentages of these various religious adherents in the given population.

The Sampling Distribution of Percentages. In making inferences about the percentages of various attributes in a manifold population, the sample statistics that naturally offer themselves for this purpose are the sample percentages of these attributes. Subsequently, another statistic will be described that is more commonly used in large samples; but for the moment attention will be centered upon the sample percentages. To make use of these percentages in testing hypotheses, etc., the sampling distribution of the percentages must be derived. This will be the principal task of the present section.

Derivation. The analysis of Chap. IX suggests that the sampling distribution of percentages can most conveniently be derived from the following mathematical model: Suppose 10 packs of cards in each of which there are p_1 per cent red, p_2 per cent white, and p_3 per cent blue cards. Let all possible combinations of N cards each be made by combining without restriction each card in any given pack with a single card from each of the other packs. Since the probability of a red card in

each group is p_1 , the probability of a white card is p_2 , and the probability of a blue card is p_3 ; and since combinations are formed without restriction so that the selection of any card from any one pack does not affect the probabilities in other packs, then by the multiplication theorem the probability of a combination having, for example, the first three cards red, the second five white; and the last two blue, is $p_1^3 p_2^5 p_3^2$.

The probability just indicated is true for any combination, however, in which there are three red, five white, and two blue cards. The total number of such combinations is given by the combinatorial formula¹

$$C_{N_1 \cdot N_2 \cdot N_3}^N = \frac{N!}{N_1! N_2! N_3!}$$

which, for the above example, equals

$$\frac{10!}{3! 5! 2!} = 2,520$$

Accordingly, the total probability of such a combination would be

$$\frac{10!}{3! 5! 2!} p_1^3 p_2^5 p_3^2$$

For N packs of cards and for combinations containing N cards, the probability of a combination containing N_1 red, N_2 white, and N_3 blue cards would be

$$\frac{N!}{N_1! N_2! N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3}$$

where $N_1 + N_2 + N_3 = N$.

The last probability may be taken as a good prediction of the relative frequency with which samples of N would have N_1 Catholics, N_2 Protestants, and N_3 other denominations in the whole set of samples of size N that might, by repeated sampling (with replacements²), be drawn at random from the given population. This is true because, if the process of sampling is random,

¹ See pp. 24-26.

² Since the population is finite, each case would have to be "put back" so as to be eligible for withdrawal on the next draw. This need only be true, however, for the theoretical argument. In practice it is assumed that for very large populations the effect of not making replacements is so slight that it can be ignored. Cf. pp. 187-188.

it is reasonable to suppose that every particular combination of persons will appear as frequently as every other combination so that the relative frequency of samples of N having N_1 Catholics, N_2 Protestants, and N_3 other denominations will tend to be the same as the relative frequency of combinations of N_1 red, N_2 white, and N_3 blue cards among the set of all possible combina-

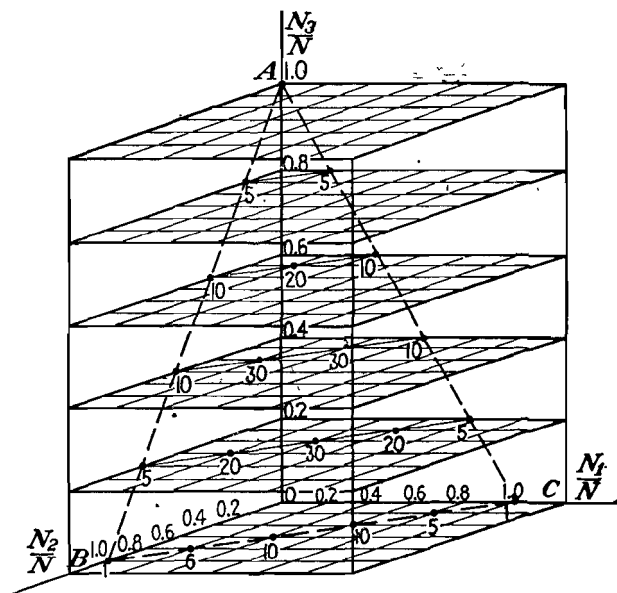


FIG. 92.—The multinomial distribution represented by the equation

$$P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) = \frac{5!}{N_1!N_2!N_3!} \cdot \left(\frac{1}{3}\right)^5.$$

(Probabilities are expressed in terms of $\frac{1}{27}$ ds.)

tions of N cards that might be made by selecting a card from each of N different packs. The formula for the sampling distribution of percentages from a threefold population is thus

$$P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) = \frac{N!}{N_1!N_2!N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3}$$

or, for a manifold population,

$$P\left(\frac{N_1}{N}, \frac{N_2}{N}, \dots, \frac{N_k}{N}\right) = \frac{N!}{N_1!N_2! \dots N_k!} p_1^{N_1} p_2^{N_2} \dots p_k^{N_k} \quad (1)$$

It will be recognized that this is merely a generalization of the formula for the binomial distribution. Its technical name is the "multinomial distribution" since it is the equation also for the terms of a multinomial expansion, that is,

$$(p_1 + p_2 + \cdots + p_k)^N.$$

The Multinomial Distribution. In the case of the binomial distribution, a sample could be described by a single percentage figure. For $\frac{N_1}{N} + \frac{N_2}{N} = 1$; and as soon as $\frac{N_1}{N}$ was specified, the type of sample was fully designated. A graph of the binomial distribution could therefore be reduced to an ordinary two-dimensional graph. For a multinomial distribution, however, this is not possible, for there are at least two independent percentages that characterize each sample. A graph of a multinomial distribution must therefore be multidimensional.

Illustration of Symmetrical Multinomial Distribution. A concrete illustration of a very simple multinomial distribution is shown in Fig. 92. This is a graph of the multinomial distribution for which $N = 5$, $k = 3$, and $p_1 = p_2 = p_3 = \frac{1}{3}$. Accordingly, the graph represents the distribution of probabilities of various types of samples of size 5 drawn at random from a three-fold population in which the percentages of cases in the three classes are all equal. The equation for this particular distribution is

$$\begin{aligned} P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) &= \frac{5!}{N_1!N_2!N_3!} \left(\frac{1}{3}\right)^{N_1} \left(\frac{1}{3}\right)^{N_2} \left(\frac{1}{3}\right)^{N_3} \\ &= \frac{5!}{N_1!N_2!N_3!} \left(\frac{1}{3}\right)^5 \end{aligned}$$

and the values of $P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right)$ given by this equation for various values of $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ are recorded in Table 30.—

Since there are three classes, the graph of this particular multinomial distribution is drawn in three dimensions. Each point in the three-dimensional diagram represents a particular set of values for $\frac{N_1}{N}$, $\frac{N_2}{N}$, $\frac{N_3}{N}$. These combinations of $\frac{N_1}{N}$, $\frac{N_2}{N}$, $\frac{N_3}{N}$ values which are possible under the conditions of the problem

TABLE 30.—THE MULTINOMIAL DISTRIBUTION REPRESENTING THE SAMPLE PERCENTAGES FOR WHICH $N = 5$, $k = 3$, AND $p_1 = p_2 = p_3 = \frac{1}{3}$

Type of sample			Probability
$\frac{N_1}{N}$	$\frac{N_2}{N}$	$\frac{N_3}{N}$	$P \left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N} \right)$
1.0	.0	.0	$\frac{1}{243}$
.0	1.0	.0	$\frac{1}{243}$
.0	.0	1.0	$\frac{1}{243}$
.8	.2	.0	$\frac{5}{243}$
.8	.0	.2	$\frac{5}{243}$
.2	.0	.8	$\frac{5}{243}$
.2	.8	.0	$\frac{5}{243}$
.0	.8	.2	$\frac{5}{243}$
.0	.2	.8	$\frac{5}{243}$
.6	.4	.0	$\frac{10}{243}$
.6	.0	.4	$\frac{10}{243}$
.0	.6	.4	$\frac{10}{243}$
.4	.6	.0	$\frac{10}{243}$
.4	.0	.6	$\frac{10}{243}$
.0	.4	.6	$\frac{10}{243}$
.6	.2	.2	$\frac{20}{243}$
.2	.6	.2	$\frac{20}{243}$
.2	.2	.6	$\frac{20}{243}$
.4	.4	.2	$\frac{30}{243}$
.4	.2	.4	$\frac{30}{243}$
.2	.4	.4	$\frac{30}{243}$

are marked by large dots. The probabilities of getting these various combinations, expressed in terms of $\frac{1}{2^4}$'s, are written slightly below and to the right or to the left of the dots.

First, it will be noticed that only a limited number of $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ combinations are possible. This is because $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ must always add up to 1. The various values of $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ are thus subject to the condition that $\frac{N_1}{N} + \frac{N_2}{N} + \frac{N_3}{N} = 1$. This condition, as the mathematicians put it, restricts the "degrees of freedom" in selecting $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ values. Without it there would be three degrees of freedom (three dimensions) for the selection of $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ values; with it the degrees of freedom are reduced to two (two dimensions). Geometrically this means that the $\frac{N_1}{N}$, $\frac{N_2}{N}$, and $\frac{N_3}{N}$ values must be confined to the slanting plane ABC , the equation of which is the condition $\frac{N_1}{N} + \frac{N_2}{N} + \frac{N_3}{N} = 1$. This particular aspect of the problem is emphasized here, because the concept of degrees of freedom will enter into most of the subsequent sampling analysis. If it is mastered now, the student should have little trouble with what is to come.

The symmetry of the given multinomial distribution will also be noted. Those combinations that have the most even distribution of cases among the three classes (.4,.4,.2; .4,.2,.4; and .2,.4,.4) have the highest probability of occurrence ($\frac{3}{2^4}$); the extremely uneven combinations (1.0,0,0; 0,1.0,0; and 0,0,1.0) have the lowest probability ($\frac{1}{2^4}$); and similarly for the intermediate combinations. This symmetry is due to the equality of p_1 , p_2 , and p_3 . A subsequent example will demonstrate what happens to the distribution when the p 's are unequal.

The Means and Standard Deviations of a Multinomial Distribution. The general similarity of the multinomial distribution to the binomial distribution (indeed the latter is merely a special case of the former) suggests that the equations for the means

and standard deviations of the multinomial distribution will be of the same type as those of the binomial distribution. This is in fact the case. The mean or expected percentage of cases falling in class 1 is p_1 , the mean percentage of cases falling in class 2 is p_2 , and in general the mean percentage of cases falling in class k is p_k . The standard deviation of the percentage of cases falling in class 1 is¹

$$\sigma_{\frac{N_1}{N}} = \sqrt{\frac{p_1(1 - p_1)}{N}}$$

the standard deviation of the percentage of cases falling in class 2 is

$$\sigma_{\frac{N_2}{N}} = \sqrt{\frac{p_2(1 - p_2)}{N}}$$

and, in general, the standard deviation of the percentage of cases falling in class k is

$$\sigma_{\frac{N_k}{N}} = \sqrt{\frac{p_k(1 - p_k)}{N}}$$

No general proof will be given of these equations. Their validity may be tested, however, by applying them to the data of Table 30 and by comparing the mean and standard deviation thus obtained with the mean and standard deviation obtained by the use of conventional methods of calculation.

By definition the mean of a variable X_i which has the relative frequencies p_i is

$$\bar{X} = \sum p_i X_i$$

The mean value of N_1/N (i.e., the mean percentage of cases falling in class 1) is accordingly

$$\bar{X} = \sum P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) \frac{N_1}{N}$$

Numerically, the value of $\bar{X}$ for class 1 can thus be obtained by multiplying the figures of the first column of Table 30 by the

¹ For the binomial case the corresponding equation is

$$\sigma_{\frac{N_1}{N}} = \sqrt{\frac{p_1(1 - p_1)}{N}} = \sqrt{\frac{p_1 p_2}{N}} \quad \text{since } p_1 + p_2 = 1$$

corresponding probabilities given in the last column. The results of these row-by-row calculations are given in the first column

TABLE 31.—CALCULATION OF THE MEAN AND STANDARD DEVIATION OF PERCENTAGE OF CASES FALLING IN CLASS 1
(Multinomial distribution representing the sample percentages for which $N = 5$, $k = 3$, and $p_1 = p_2 = p_3 = \frac{1}{3}$)

$\left(\frac{N_1}{N}\right) P \left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right)$	$\left(\frac{N_1}{N}\right)^2 P \left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right)$
1	1
$\frac{243}{243}$	$\frac{243}{243}$
4	3.2
$\frac{243}{243}$	$\frac{243}{243}$
4	3.2
$\frac{243}{243}$	$\frac{243}{243}$
1	.2
$\frac{243}{243}$	$\frac{243}{243}$
1	.2
$\frac{243}{243}$	$\frac{243}{243}$
6	3.6
$\frac{243}{243}$	$\frac{243}{243}$
6	3.6
$\frac{243}{243}$	$\frac{243}{243}$
4	1.6
$\frac{243}{243}$	$\frac{243}{243}$
4	1.6
$\frac{243}{243}$	$\frac{243}{243}$
12	7.2
$\frac{243}{243}$	$\frac{243}{243}$
4	.8
$\frac{243}{243}$	$\frac{243}{243}$
4	.8
$\frac{243}{243}$	$\frac{243}{243}$
12	4.8
$\frac{243}{243}$	$\frac{243}{243}$
12	4.8
$\frac{243}{243}$	$\frac{243}{243}$
6	1.2
$\frac{243}{243}$	$\frac{243}{243}$
81	37.8
$\frac{243}{243}$	$\frac{243}{243}$

of Table 31, and the sum of this column is therefore the mean value of N_1/N . Consequently, $\bar{X}$ for N_1/N is $\frac{81}{243} = \frac{1}{3}$.

By use of a short method the standard deviation of the sample values of N_1/N (i.e., the sample percentages of cases falling in class 1) can be calculated from Table 30 by the formula¹

$$\sigma_{\frac{N_1}{N}} = \sqrt{\sum P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) \left(\frac{N_1}{N}\right)^2 - p_1^2}$$

It will be recognized that the first term of the square root is merely the sum of the squares of the percentages of column (1), Table 30, multiplied by the corresponding probabilities. Each of these row-by-row products is given in column (2), Table 31, and their sum is the sum of this column, viz., 37.8/243. Hence,

$$\sigma_{\frac{N_1}{N}} = \sqrt{\frac{37.8}{243} - \left(\frac{81}{243}\right)^2} = .21$$

The same results are obtained by the use of the formulas; for $\bar{X}$ calculated above = $\frac{1}{3}$, which equals $p_1 = \frac{1}{3}$. Using the equation for standard deviation,

$$\sigma_{\frac{N_1}{N}} = \sqrt{\frac{p_1(1-p_1)}{N}} = \sqrt{\frac{(\frac{1}{3})(\frac{2}{3})}{5}} = \sqrt{\frac{2}{45}} = .21$$

An Illustration of a Skewed Multinomial Distribution. Another example will now be used to illustrate a skewed multinomial distribution.

Suppose that there are again three classes, that the probability of a single case falling in class 1 is $p_1 = \frac{1}{2}$, that the probability of a case falling in class 2 is $p_2 = \frac{1}{3}$, and that the probability of a case falling in class three is $p_3 = \frac{1}{6}$; again suppose that five cases are chosen at random. Under these conditions the probability of getting $\frac{N_1}{N}$ cases in class 1, $\frac{N_2}{N}$ cases in class 2, and $\frac{N_3}{N}$ cases in class 3 is as follows:

$$P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) = \frac{5!}{N_1!N_2!N_3!} \left(\frac{1}{2}\right)^{N_1} \left(\frac{1}{3}\right)^{N_2} \left(\frac{1}{6}\right)^{N_3}$$

¹ This is merely the short formula

$$\sigma = \sqrt{\sum \frac{F}{N} X^2 - \bar{X}^2}$$

where $P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) = \frac{F}{N}$, $\frac{N_1}{N} = X$, and $p_1^2 = \bar{X}^2$.

and $\frac{N_3}{N}$ are again constrained to lie in the slanting plane ABC , that is, the degrees of freedom are only two. The distribution of probabilities is no longer symmetrical, however. Since the probability of a single case falling in the first class is greater than the probability of a single case falling in either the second or the third class, the multinomial distribution shows a major bunching of cases in the direction of the $\frac{N_1}{N}$ -axis; and since the probability of a case falling in the second class is greater than the probability of a case falling in the third class, the distribution shows a minor bunching of cases in the direction of the $\frac{N_2}{N}$ -axis. All this is clearly revealed in Fig. 93. It is to be noted, however, that the mean and standard deviation formulas continue to hold true for this skewed form.

Accordingly, the mean $\frac{N_1}{N}$ equals $p_1 = \frac{1}{2}$, and the standard deviation of $\frac{N_1}{N}$ equals

$$\sqrt{\frac{p_1(1-p_1)}{N}} = \sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{5}\right)} = \sqrt{.05} = .2236$$

and similar calculations will give the means and standard deviations of $\frac{N_2}{N}$ and $\frac{N_3}{N}$.

Use of Multinomial Distribution in Testing Hypotheses. Like the binomial distribution, the multinomial distribution may be used to test hypotheses regarding the true division of cases in a population. Consider, for example, the following problem: Suppose there are three candidates for election to the same office. An inquiring reporter stops five persons at random and asks which candidate they favor. The results obtained are as follows:

Candidate	Number of Persons Favoring Specified Candidate
A	3
B	2
C	0

TABLE 32.—MULTINOMIAL DISTRIBUTION REPRESENTING THE DISTRIBUTION OF SAMPLE PERCENTAGES FOR WHICH $N = 5$, $k = 3$, AND $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, AND $p_3 = \frac{1}{6}$

Type of sample			Probability
$\frac{N_1}{N}$	$\frac{N_2}{N}$	$\frac{N_3}{N}$	$P \left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N} \right)$
1.0	.0	.0	$\frac{243}{7,776}$
.0	1.0	.0	$\frac{32}{7,776}$
.0	.0	1.0	$\frac{1}{7,776}$
.8	.2	.0	$\frac{810}{7,776}$
.8	.0	.2	$\frac{405}{7,776}$
.0	.8	.2	$\frac{80}{7,776}$
.2	.8	.0	$\frac{240}{7,776}$
.2	.0	.8	$\frac{15}{7,776}$
.0	.2	.8	$\frac{10}{7,776}$
.6	.4	.0	$\frac{1,080}{7,776}$
.6	.0	.4	$\frac{270}{7,776}$
.0	.6	.4	$\frac{80}{7,776}$
.4	.6	.0	$\frac{720}{7,776}$
.4	.0	.6	$\frac{90}{7,776}$
.0	.4	.6	$\frac{40}{7,776}$
.6	.2	.2	$\frac{1,080}{7,776}$
.2	.6	.2	$\frac{480}{7,776}$
.2	.2	.6	$\frac{120}{7,776}$
.4	.4	.2	$\frac{1,080}{7,776}$
.4	.2	.4	$\frac{810}{7,776}$
.2	.4	.4	$\frac{360}{7,776}$

On the assumption that the population from which this sample is taken is relatively large, do these results disprove the hypothesis that the population as a whole is equally divided with respect to the three candidates? That is, does the sample division of .6, .4, and .0 disprove the hypothesis of a true division of .33+, .33+, and .33+?

To answer this question it is necessary, as in the twofold case, first to decide upon a coefficient of risk of rejecting a hypothesis when it is really true. Let this be .10 in the present instance. In other words, the investigating organization is willing to run the risk of rejecting a true hypothesis 10 times out of 100.

The second step is to derive the distribution of sample percentages for samples of 5 from the assumed population. Since in the present instance the hypothesis is that $p_1 = p_2 = p_3 = \frac{1}{3}$, the desired sampling distribution is that described by the multinomial distribution of Table 30 and Fig. 92.

The third step is to select a group of unusual samples that may constitute a "region of rejection." In this particular instance the distribution of samples is discrete and the number of cases is small, so that approximation by a continuous distribution is likely to be very inaccurate. It may accordingly be impossible to find any region that has a probability exactly equal to the adopted coefficient of risk. In the present problem it seems reasonable to proceed as follows: In Fig. 94, where the plane *ABC* of Fig. 92 is laid out flat, all the more unusual samples (*i.e.*, those with the lowest probabilities) are seen to lie on or outside of a given circle that is ruled double in the diagram. The total probability of this group is .135. Although the region consisting of the circle and the area outside of it would thus constitute a region of rejection with a probability greater than the adopted coefficient of risk, this probability is not much greater.

Let it be assumed that the investigating organization is indifferent to the other values of the population percentages that might be true, *i.e.*, that it is willing to run the same chance of accepting the given hypothesis in whatever direction the true percentages might happen to lie relative to the assumed percentages. Under these circumstances, a symmetrical region of rejection, such as the circular area just described, should be the

type of region adopted. In view of this situation, therefore, it will be assumed that the investigating body raises its coefficient of risk to .135 and adopts the given region as its region of rejection.

The final step in the analysis is to note that the given sample (3,2,0) does not fall in the region of rejection. The investigating organization consequently concludes that the hypothesis of an

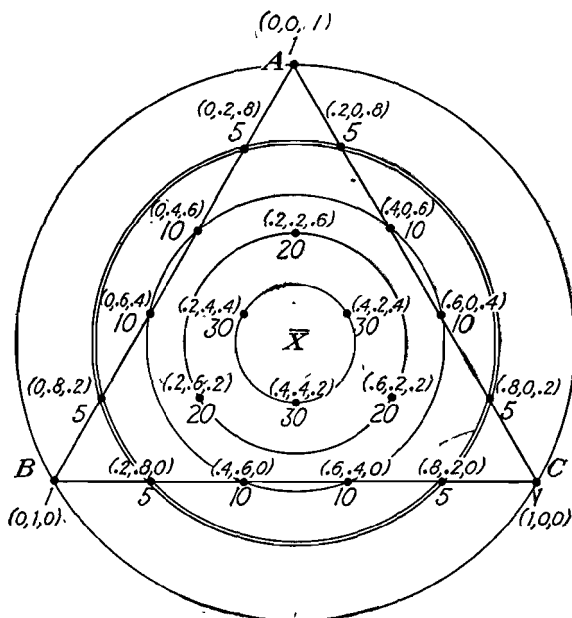


FIG. 94.—The plane of the multinomial distribution represented by the equation

$$P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) = \frac{5!}{N_1!N_2!N_3!} \left(\frac{1}{3}\right)^5.$$

(Probabilities are expressed in terms of $\frac{1}{2} \frac{1}{3}$ ds.)

equal division of sentiment regarding the three candidates cannot be rejected as a result of the knowledge obtained from the sample.

If the hypothesis had been other than an equal division of sentiment, the multinomial distribution describing the set of all possible samples would have been a skewed one such as that represented by Table 32 and Fig. 93. In these instances, the selection of a symmetrical region of rejection presents difficulties. This problem need not be discussed here, however, for in very few instances are such small samples taken as that assumed in the given problem. When larger samples are used, it is possible to

make a mathematical transformation of a skewed multinomial distribution that makes it more symmetrical and at the same time permits a redescription of the distribution in terms of a new statistic that has a simple one-dimensional sampling distribution. This is discussed in the ensuing section.

The Sampling Distribution of the Statistic $\sum \frac{(N_i - Np_i)^2}{Np_i}$.

One of the important contributions of mathematical statistics has been the demonstration that when the set of all possible samples from a manifold population is described in terms of the statistic $\sum \frac{(N_i - Np_i)^2}{Np_i}$, the sampling distribution is reasonably well described, if N is large, by the χ^2 distribution. Hence, instead of using the multinomial distribution for testing various hypotheses, it is much more practical in the case of large samples to calculate the sample statistic $\sum \frac{(N_i - Np_i)^2}{Np_i}$ and use the χ^2 distribution, for which tables are readily available.

The complete argument by which this important conclusion is reached is not given in this section. Nevertheless, multinomial problems arise so frequently and the χ^2 distribution is used as a substitute for the multinomial distribution in so many of these problems that it is well for the student to understand the basis upon which this use of the χ^2 distribution rests. The argument, therefore, by which it is established that the statistic $\frac{\sum (N_i - Np_i)^2}{Np_i}$ has a sampling distribution of the form of the χ^2 distribution is discussed in a simplified way in the following section.

A Simple Version of the Argument: the Symmetrical Case. In Fig. 94, the plane ABC from Fig. 92 is laid out horizontally; the sample points are all marked with heavy dots, and the coordinates of the point are written beside each. In addition, the probability of each is indicated. It will be noted that the point $\bar{X}$, which was not shown in the original figure (Fig. 92), represents the point whose coordinates are the means of the N_k/N values, i.e., for this case, where $p_1 = \frac{1}{3}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{3}$. Although the point $\bar{X}$ is not a sample point,¹ it is nevertheless important in that it marks the center of the distribution.

¹ If nine cases, for example, had been chosen instead of five, the mean point would have also been one of the sample points.

It will be noted from Fig. 94 that, for the particular multinomial distribution represented, sample points of equal probability all lie equally distant from the mean point $\bar{X}$. That is, points of equal probability lie in a circle with center at $\bar{X}$. This characteristic of the distribution suggests that the set of all possible samples might be described in a somewhat simpler way, *viz.*, by showing how the probability varies with the distance of sample points from the mean point $\bar{X}$ or preferably with the square of the distance, since the latter is easier to calculate.

In pursuance of this line of thought, the following table of probabilities may be calculated from Fig. 94 and used in all practical problems as a substitute for the original multinomial distribution. The point (.2, .4, .4), for example, is distant from

TABLE 33.—PROBABILITIES OF SPECIFIED VALUES OF D^2

D^2 = square of distance of sample point from mean point	$P(D^2)$ = probability of getting a sample point distant by D from mean point
.0266 +	$\frac{90}{243} = 0.370$
.1066 +	$\frac{60}{243} = 0.247$
.1866	$\frac{60}{243} = 0.247$
.2466	$\frac{30}{243} = 0.123$
.6666	$\frac{3}{243} = 0.012$

the mean point $\bar{X}$ by¹

$$D = \sqrt{\sum \left(\frac{N_i}{N} - p_i \right)^2}$$

$$= \sqrt{\left(.2 - \frac{1}{3} \right)^2 + \left(.4 - \frac{1}{3} \right)^2 + \left(.4 - \frac{1}{3} \right)^2} = \sqrt{.0266}$$

Points (.4, .2, .4) and (.4, .4, .2) are equally distant from $\bar{X}$. The probability of getting a particular one of these three sample

¹ The general equation for measuring the square of the distance between points (x_1, y_1, z_1) and (x_2, y_2, z_2) is

$$d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$$

results is $\frac{30}{243}$, and the probability of getting any one of them, i.e., the probability of getting a sample point distant from the point by $D = \sqrt{.0266}$, is $.3(30)/243 = \frac{30}{243} = .370$. Similar calculations give the other entries in Table 33. A more useful form of Table 33 is given in Table 34. This cumulates the probabilities of Table 33; it thus gives the probability of getting a sample point that is at least as far from the mean point as the given sample point.

 TABLE 34.—PROBABILITIES OF $D^2 \geq$ SPECIFIED QUANTITIES

$D^2 \setminus$	Probability of as great or greater D^2
.0266 +	$\frac{243}{243} = 1.000$
.1066 +	$\frac{153}{243} = .629$
.1866 +	$\frac{93}{243} = .382$
.2466 +	$\frac{33}{243} = .135$
.6666 +	$\frac{3}{243} = .012$

The advantage of Table 34 is the ease with which it can be used to test the hypothesis $p_1 = p_2 = p_3 = \frac{1}{3}$. If .135 is adopted as the coefficient of risk, then Table 34 indicates immediately that samples for which $D^2 \geq .2466$ would constitute a symmetrical region of rejection having the probability .135. Hence, given any sample, it is merely necessary to calculate

$$D^2 = \sum \left(\frac{N_i}{N} - p_i \right)^2$$

and to note whether the sample D^2 falls above or below .2466. There is no longer need to draw a diagram such as Fig. 92 in order to mark off a two-dimensional region of rejection, for the set of all possible samples has now been described in terms of a single statistic, D^2 , whose sampling distribution (Table 34) is one-dimensional. From the way Table 34 was derived, however, it is known that the upper .135 tail of the distribution of D^2 is the mathematical equivalent of the symmetrical two-dimensional region of Fig. 92. The simpler one-dimensional description

of the set of all possible samples thus accomplishes the same result as the multidimensional description.

The General Case. When the hypothesis gives to p_1 , p_2 , and p_3 values that are not equal, the result is a skewed multinomial distribution. Generally, therefore, it is not possible to draw neat circles through points of equal probability, such as was done in the special case just considered; nor is it possible to redescribe the set of all possible samples exactly in terms of the distance D^2 from the point given by the mean percentages p_1 , p_2 , and p_3 . Fortunately, however, a certain procedure can be adopted that will give approximate results similar to those obtained in the simpler case. This procedure is as follows:

In the more general case, considerable advantage is gained by changing the scale units. Instead of taking N_1/N , N_2/N , and N_3/N as the coordinates of a sample point, it is found convenient to take $N_1/\sqrt{Np_1}$, $N_2/\sqrt{Np_2}$, and $N_3/\sqrt{Np_3}$ as the sample coordinates.¹ This means that the scales are modified in proportion to square roots of the hypothetical probabilities, the net effect of which is to make the distribution more symmetrical. The point represented by the mean percentages becomes, in terms of the new coordinates, $p_1 \sqrt{N/p_1}$, $p_2 \sqrt{N/p_2}$, and $p_3 \sqrt{N/p_3}$, or, simply, $\sqrt{Np_1}$, $\sqrt{Np_2}$, and $\sqrt{Np_3}$; and the square of the distance between this point and any sample point becomes $\sum \left(\frac{N_i}{\sqrt{Np_i}} - \sqrt{Np_i} \right)^2$, or $\sum \frac{(N_i - Np_i)^2}{Np_i}$.

Owing to the approximate symmetry resulting from this change in the scale units, a symmetry that improves the larger the value of N , it is possible to give an approximate description of the set of all possible samples in terms of

$$D'^2 = \sum \frac{(N_i - Np_i)^2}{Np_i}$$

When this is done mathematically, it is found that the distribution of probabilities is approximately of the form of a χ^2 distribution where n in the χ^2 equation is taken equal to the number of classes, minus 1.

To illustrate this important conclusion, suppose that sampling is made from a population in which the various items fall into

¹ That is, the old coordinates are multiplied by $\sqrt{N/p_i}$.

three classes. Then on the assumption that the percentages of items in these three classes are p_1 , p_2 , and p_3 , the relative frequencies, or probabilities, with which various samples would have values of $D'^2 = \sum \frac{(N_i - Np_i)^2}{Np_i}$ equal to or exceeding certain specified values are as follows:¹

TABLE 35.—PROBABILITIES OF $D'^2 \geq$ SPECIFIED QUANTITIES

Values of D'^2	Probability of as Great or Greater Value ¹ (= Probability of as Great or Greater χ^2 for $n = 3 - 1 = 2$)
.0201	.99
.0404	.98
.103	.95
.211	.90
.446	.80
.713	.70
1.386	.50
2.408	.30
3.219	.20
4.605	.10
5.991	.05
7.824	.02
9.210	.01

That is, in the set of all possible samples from the given population, the quantity $D'^2 = \sum \frac{(N_i - Np_i)^2}{Np_i}$ would have a value equal to or greater than 5.991 in 5 per cent of the samples, or a value equal to or exceeding 4.605 in 10 per cent of the samples, and so forth.

In the original description of the χ^2 distribution, n was said to determine the nature of the curve². Here n may be identified with the degrees of freedom. The reason for this in the present problem should now be clear. When the number of classes into which a discrete population is divided is three, as has been assumed in the above discussion, then the quantity of $D'^2 = \sum \frac{(N_i - Np_i)^2}{Np_i}$ has a sampling distribution of the form

¹ This is merely the row " $n = 2$ " of a χ^2 table (Appendix, Table VIII).

² See p. 111.

of the χ^2 distribution with $n = 3 - 1 = 2$. It will be recalled, however, that when the set of all possible sample points is described in terms of the percentages falling into each class, the various possible sample percentages must conform to the equation $\frac{N_1}{N} + \frac{N_2}{N} + \frac{N_3}{N} = 1$. Geometrically this meant that all the sample points had to lie on a plane (the plane ABC of Fig. 92, for example). It was accordingly said that there were only two degrees of freedom for the selection of the various possible samples. Since for the case of three classes, n in the χ^2 formula is equal to 2, this n becomes the same as the degrees of freedom in the given problem.

This relationship also holds true for any number of classes. For the inevitable equation $\frac{N_1}{N} + \frac{N_2}{N} + \dots + \frac{N_k}{N} = 1$ —inevitable because the sum of all the class percentages must be 100 per cent—always reduces the degrees of freedom with which various possible samples may be selected from k to $k - 1$, while it is also always true that the quantity $D'^2 = \sum \frac{(N_i - Np_i)^2}{Np_i}$ has a distribution approximately of the form of the χ^2 distribution, with n in the χ^2 equation equal to $k - 1$. Hence in problems of this kind the n in the χ^2 equation is always the same as the degrees of freedom.

Estimation of Population Percentages. Zones of Confidence. When, as in the case of the multinomial distribution and the distribution of $\sum \frac{(N_i - Np_i)^2}{Np_i}$, there is more than one population parameter to be estimated, the determination of confidence intervals for these parameters presents difficulties. In abstract terms, the problem is simple enough. Thus, for any given case, the .05 point, say, of the distribution of $\sum \frac{(N_i - Np_i)^2}{Np_i}$ could be determined from a χ^2 table (all that need be known for this is the value of n in the χ^2 formula), and the $\sum \frac{(N_i - Np_i)^2}{Np_i}$ could be set equal to this value. The resulting equation would give the value of the population percentages p_i that would just be on the border of reasonableness (assuming a coefficient of confidence = .95),

Geometrically, the locus of the values of the p_i 's satisfying this equation would constitute a definite geometrical figure such as an ellipse or, when there are more than three parameters, some sort of an ellipsoid. All values lying outside this locus would be considered unreasonable values, and all inside would be considered reasonable values. The enclosed area would constitute a "zone of confidence," and in repeated sampling it might be said that this zone would include the true value 95 per cent of the time.

The difficulty that arises in trying to make use of the abstract analysis above is that the locus marking the zone of confidence is not a simple ellipse or ellipsoid but a much more complicated figure. In fact, practical determination of the values of p_i constituting that locus would involve so much trouble that it is almost never undertaken. In general, the investigator is content to test particular hypotheses that may be of importance rather than attempting to class all hypotheses into those that are reasonable and those that are unreasonable.

Maximum-likelihood. Although determination of zones of confidence for the values of the population percentages is thus not usually undertaken, single maximum-likelihood estimates may be readily made. For it is found that the values of p_1 , p_2 , p_3 , etc., that maximize the probability of a given sample result are but the sample percentages N_1/N , N_2/N , N_3/N , etc.¹

¹ This may be demonstrated as follows: If there are three class divisions, the probability of a given sample point is

$$P\left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}\right) = \frac{N!}{N_1!N_2!N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3}$$

By the method of maximum likelihood, the values of p_1 , p_2 , and p_3 are to be chosen so that the logarithm of this probability is a maximum. The percentages p_1 , p_2 , and p_3 are not independent, however, for they must add up to 100 per cent, i.e., $p_1 + p_2 + p_3 = 1$. In the process of estimation, therefore, let p_1 and p_2 be the independent estimates to be made, and let p_3 be determined from p_1 and p_2 . Under these conditions the maximizing values

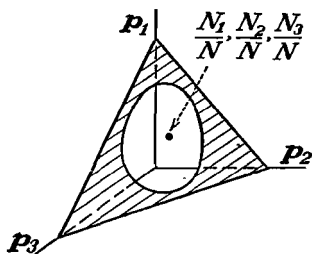


FIG. 95.—A confidence zone for p_1 , p_2 , and p_3 .

Maximum-likelihood estimates of the population percentages are thus a very simple matter.

APPLICATION OF THEORY TO SELECTED PROBLEMS

Sampling Public Opinion. In sampling public opinion, the division of opinion is often threefold or more. In all such instances the foregoing theory becomes immediately applicable. Consider, for example, the following problem:

In 1940 the election returns were as follows:

	Per Cent
F. D. Roosevelt, Democratic candidate.....	54.7
Wendell Willkie, Republican candidate.....	44.8
Other candidates.....	0.5

Suppose that in the fall of 1944 a given polling agency took a random sample of 1,000 voters¹ and found that 510 were for Roosevelt, again the Democratic candidate, 478 were for Dewey,

of p_1 , p_2 , and p_3 must be such that the partial derivatives of the logarithm of the probability with respect to p_1 and p_2 must both be equal to zero.

The logarithm of the above probability is

$$\log P \left(\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N} \right) = \log \left(\frac{N!}{N_1!N_2!N_3!} \right) + N_1 \log p_1 + N_2 \log p_2 + N_3 \log p_3$$

and the partial derivatives of this with respect to p_1 and p_2 , it being remembered that $\partial p_3 / \partial p_1 = \partial p_3 / \partial p_2 = -1$, are

$$\begin{aligned} \frac{N_1}{p_1} - \frac{N_3}{p_3} &= 0 \\ \frac{N_2}{p_2} - \frac{N_3}{p_3} &= 0 \end{aligned}$$

or $N_1 p_3 = N_3 p_1$ and $N_2 p_3 = N_3 p_2$.

Now add these two equations, and to each side of the total add $N_3 p_3$, giving

$$(N_1 + N_2 + N_3) p_3 = N_3 (p_1 + p_2 + p_3)$$

Since $N_1 + N_2 + N_3 = N$, and $p_1 + p_2 + p_3 = 1$, this gives (using a breve to denote an estimate)

$$\check{p}_3 = \frac{N_3}{N}$$

Then from the two previous equations, it is found that $\check{p}_1 = N_1/N$ and $\check{p}_2 = N_2/N$. Hence, as stated in the body of the text, the maximum likelihood estimates of the population percentages are the sample percentages.

¹ To avoid complications, a simple random sample is assumed to be taken. Actually, representative random sampling is employed by most polling agencies. See pp. 183-185.

the Republican candidate, and 12 were for other candidates. Could it have been reasonably inferred from this sample that popular sentiment regarding the various parties had changed? In other words, would the hypothesis of 54.7 per cent Democratic, 44.8 per cent Republican, and 0.5 per cent other affiliations have been a tenable hypothesis in view of the sample results?

To answer this question the investigation proceeds as follows: First the statistic $D'^2 = \sum \frac{(N_i - Np_i)^2}{Np_i}$ is calculated. This gives, in the present instance,

$$D'^2 = \frac{(510 - 547)^2}{547} + \frac{(478 - 448)^2}{448} + \frac{(12 - 5)^2}{5} = 11.5$$

Next it is noted that there are only three classes, so that D'^2 is approximately distributed as χ^2 , with the degrees of freedom n equal to $3 - 1 = 2$. Finally a region of rejection is chosen. If .05 is taken as the coefficient of risk, a χ^2 table shows that for $n = 2$ the region consisting of values of D'^2 equal to or greater than 5.991 is an appropriate region of rejection to select.¹ In the given instance the sample $D'^2 = 11.5$ clearly falls in the region of rejection. Consequently, the hypothesis that there has been no change in political sentiment cannot be accepted.

Since the maximum-likelihood estimates of population percentages are the sample percentages,² it follows in this case that the maximum-likelihood estimates to be made of the true political sentiment are 51 per cent Democratic, 47.8 per cent Republican, and 1.2 per cent other parties.

Goodness of Fit of a Frequency Curve. The use of the χ^2 distribution in testing the goodness of fit of a frequency curve has already been discussed in a practical way in Chap. VII.³ The theoretical basis of the method there outlined will now be discussed.

¹ It will be recalled that if $N_1/\sqrt{Np_1}$, $N_2/\sqrt{Np_2}$, etc., are taken as the coordinates of sample points, D' constitutes the distance from a sample point to the mean point $\sqrt{Np_1}$, $\sqrt{Np_2}$, etc. (see p. 326). Accordingly, $D' \geq \sqrt{5.991}$ constitutes a circular region. This is unbiased with respect to other values of p_1, p_2 , and p_3 than those chosen by hypothesis in the sense that the probability of samples falling in this region is least when p_1, p_2, p_3 have the given hypothetical values and increases in whatever direction p_1, p_2 , and p_3 deviate from these hypothetical values.

² See p. 329.

³ See pp. 137-152.

The procedure outlined previously called for the calculation of $\sum \frac{(F - f)^2}{f}$, where F was the frequency of cases in any class interval given by a sample histogram and f was the corresponding frequency for that interval given by some theoretical frequency curve. It should be recognized now that this is merely a special form of $\sum \frac{(N_i - Np_i)^2}{Np_i}$. For N_i is the sample number of cases falling in any class and hence is the same as F in the other equation, and Np_i is the number that would be expected to fall in that class if the sample total were divided in the same proportion as the population total and hence is the same as f . Since in the general discussion earlier in this chapter, $\sum \frac{(N_i - Np_i)^2}{Np_i}$ is approximately distributed in the form of a χ^2 distribution, so in the special case of a comparison of a sample histogram with a theoretical frequency curve the quantity $\sum \frac{(F - f)^2}{f}$ is distributed in the form of a χ^2 distribution.

There is one thing new in this problem, however, and this has to do with the determination of the degrees of freedom, n . When the theoretical curve that is fitted to a sample histogram is a normal curve, say, the theoretical probabilities pertaining to each class are not given directly by the mere hypothesis of normality. It is first necessary to estimate the mean and standard deviation of the curve from the sample. The effect of this is to impose additional conditions on the sample frequencies that reduce the degrees of freedom from k (the number of classes) minus 1 to k minus 3. For in all cases there is the condition

$$\frac{N_1}{N} + \frac{N_2}{N} + \frac{N_3}{N} + \cdots + \frac{N_k}{N} = 1.$$

Setting the mean and standard deviation of the curve equal to the mean and standard deviation of the sample histogram imposes the additional conditions¹

¹ It will be recalled that N_i is the same as F and that p_i is the same as $\frac{f}{N}$. It will also be recalled that when relative frequencies are used the mean of a distribution is $\sum \frac{f}{N} X$ and its standard deviation is $\sqrt{\sum \frac{f}{N} (X - \bar{X})^2}$.

$$\sum \frac{N_i}{N} X_i = \sum p_i X_i$$

$$\sum \frac{N_i}{N} (X_i - \bar{X})^2 = \sum p_i (X_i - \bar{X})^2$$

This means that in the set of all possible samples* (in this case, sample histograms) that might be drawn from a given population, only those samples will be considered for which the sample percentages N_i/N satisfy the above three equations.

Probabilities are thus calculated with respect to this special group of the set of all possible samples, and it is in this sense that the degrees of freedom are reduced. In general, when a frequency curve is fitted to a sample histogram and c parameters of the curve are estimated from the sample histogram, $c + 1$ conditions are imposed upon the sample percentages and the part of the set of all possible samples that is considered has $k - c - 1$

degrees of freedom. That means that $\sum \frac{(N_i - Np_i)^2}{Np_i}$, or its equivalent in terms of the frequency distribution analysis, *viz.*, $\sum \frac{(F - f)^2}{f}$, has a sampling distribution of the form of the χ^2 distribution with the degrees of freedom equal to $k - c - 1$. This is the explanation of the special value given to n in the procedure described in Chap. VII.

Testing Independence. Still another use of the multinomial distribution and the χ^2 distribution based upon it is to test whether one classification is independent of another classification or, conversely, whether the two are associated. Particular measures of association or correlation were discussed in Chap. XII, and tests of significance were developed for these measures. At this point, attention will be directed primarily to testing the existence of an association, whatever its form or degree.

Consider a simple case. In general, the percentage of males and females in a large population tends to be the same, no matter what the color of the people. That is, the sex ratio is independent of color. Within a limited area, however, this may not hold true. Economic and social forces in certain metropolitan districts, for example, may cause the percentage of black and other colored males to exceed considerably the percentage of white males, and vice versa for females. Suppose a

sample of 1,000 people is taken from a given city with a view to studying the relationship between sex and color. For this purpose the 1,000 people are "cross classified" as follows:

TABLE 36.—1,000 PEOPLE CLASSIFIED AS TO COLOR AND SEX

	Males	Females	Totals
Whites.....	380	320	700
Blacks.....	150	50	200
Other colors.....	70	30	100
Totals.....	600	400	1,000

The question is: Does this sample of 1,000 people demonstrate definitely that the population of the city as a whole has a higher percentage of black and other-color males than white males, or can this sample result reasonably be considered the effect of chance? Or, to put it another way, does this sample show conclusively that the sex ratio in this city is not independent of color?

In a problem of this kind, it is more convenient to test the negative, or "null," hypothesis. For if the classifications are assumed to be independent within the population as a whole, it is then possible to tell something about the expected distribution of the cases. Thus, if two classifications are independent, the distribution of cases in the various categories of one classification may be expected to be the same for each of the categories of the other classification. On the other hand, if the positive hypothesis of correlation is set up, nothing can be told about the expected distribution of cases unless the exact form of correlation is specified. Thus, in the absence of specific information regarding the nature of the correlation, it is customary to test the null hypothesis of independence.

With reference to the illustration, the assumption of independence means that the percentage of males (and hence the percentage of females) in the population as a whole is the same for all color groups and the percentages of whites and blacks (and hence the percentage of other colored people) are the same for both sexes. That is, the probability of picking a male at random (and hence the probability of picking a female) is independent of the probability of picking a white person at random, the probability of picking a black person at random, and

the probability of picking a person of another color at random. Likewise, both the probability of picking a white person at random and the probability of picking a person of another color are independent of the probability of picking a male at random and the probability of picking a female at random. Under these conditions, the probability of picking a *white male*, for example, is, according to the multiplication theorem for independent attributes, the product of the probability of picking a male and the probability of picking a white person. Or, again, the probability of picking a black female is simply the product of the probability of picking a female and the probability of picking a black person.

Symbolically independence may be described as follows:¹ If one classification is represented by the roman numerals I and II and the other by the arabic numbers 1, 2, and 3, then if the two classifications are independent, p_I (and hence p_{II} , which equals $1 - p_I$) is independent of p_1 , p_2 , and p_3 ; and p_I and p_2 (and hence p_3 , which equals $1 - p_1 - p_2$) are independent of p_I

TABLE 37.—SYMBOLIC ILLUSTRATION OF INDEPENDENCE

	I	II	
1	p_{I_1}	p_{II_1}	p_1
2	p_{I_2}	p_{II_2}	p_2
3	p_{I_3}	p_{II_3}	p_3
	p_I	p_{II}	1

and p_{II} . Hence, $p_{I_1} = p_I p_1$; $p_{I_2} = p_I p_2$; $p_{I_3} = p_I p_3$; $p_{II_1} = p_{II} p_1$; $p_{II_2} = p_{II} p_2$; and $p_{II_3} = p_{II} p_3$.

Now it is to be noted that the assumption of independence does not give the actual values of any of the above probabilities but merely states the existence of certain relationships among them.² Consequently, if the actual frequencies to be expected on the basis of the assumption of independence are to be found, it is necessary to estimate some of the underlying probabilities from the sample data (just as the mean and standard deviation of the population were estimated in the previous example).

¹ Reference to Table 37 may help the reader at this point.

² Just as knowledge that a population is normally distributed tells nothing about the actual distribution of probabilities but gives only its general form.

Thus, for the problem in hand, the probability of picking a male at random (p_1) can be estimated as equal to the percentage of males in the sample ($600/1,000 = 60$ per cent), the probability of picking a white person at random (p_1) can be estimated as equal to the percentage of white persons in the sample ($700/1,000 = 70$ per cent), and the probability of picking a black person at random (p_2), can be estimated as equal to the percentage of black persons in the sample ($200/1,000 = 20$ per cent). Symbolically, these "estimating equations" are

$$\frac{N_{I_1} + N_{I_2} + N_{I_3}}{N} = \check{p}_1$$

$$\frac{N_{I_1} + N_{II_1}}{N} = \check{p}_1$$

$$\frac{N_{I_2} + N_{II_2}}{N} = \check{p}_2$$

or

$$N_{I_1} + N_{I_2} + N_{I_3} = N\check{p}_1$$

$$N_{I_1} + N_{II_1} = N\check{p}_1$$

$$N_{I_2} + N_{II_2} = N\check{p}_2$$

where the N_i 's refer to the sample frequencies in the various classes, N is the total number of cases, and the $\check{p}$'s are the estimated probabilities.

Estimates of the other class probabilities¹ can be computed from these three probabilities with the help of the relationships given by the hypothesis of independence and the laws for addition and multiplication of probabilities.

Thus

$$\check{p}_{II} = 1 - \check{p}_1 = 100 \text{ per cent} - 60 \text{ per cent} = 40 \text{ per cent};$$

$$\check{p}_3 = 1 - \check{p}_1 - \check{p}_2 = 100 \text{ per cent} - 70 \text{ per cent} - 20 \text{ per cent} = 10 \text{ per cent};$$

$$\check{p}_{I_1} = \check{p}_1\check{p}_1 = (60 \text{ per cent})(70 \text{ per cent}) = 42 \text{ per cent};$$

$$\check{p}_{I_2} = \check{p}_1\check{p}_2 = (60 \text{ per cent})(20 \text{ per cent}) = 12 \text{ per cent};$$

$$\check{p}_{I_3} = \check{p}_1\check{p}_3 = (60 \text{ per cent})(10 \text{ per cent}) = 6 \text{ per cent};$$

$$\check{p}_{II_1} = \check{p}_{II}\check{p}_1 = (40 \text{ per cent})(70 \text{ per cent}) = 28 \text{ per cent};$$

$$\check{p}_{II_2} = \check{p}_{II}\check{p}_2 = (40 \text{ per cent})(20 \text{ per cent}) = 8 \text{ per cent};$$

and

$$\check{p}_{II_3} = \check{p}_{II}\check{p}_3 = (40 \text{ per cent})(10 \text{ per cent}) = 4 \text{ per cent}.$$

¹ It makes no difference what class probabilities are selected as the original three to be estimated directly from the sample data.

Applied to the number in the sample ($N = 1,000$), the foregoing estimated probabilities yield the following theoretical, or expected, frequencies for the various classes:

TABLE 38.—1,000 PEOPLE CLASSIFIED AS TO COLOR AND SEX

	Males	Females	Totals
Whites.....	420	280	700
Blacks.....	120	80	200
Other colors.....	60	40	100
Totals.....	600	400	1,000

The actual frequencies differ considerably from these expected frequencies, and the question arises: Are these differences sufficiently great to disprove the hypothesis of independence?

This is a problem involving estimated probabilities $\check{p}_{ij}$. Like the previous problems, however, it can be shown mathematically

that the quantity $\sum \frac{(N_{ij} - N\check{p}_{ij})^2}{N\check{p}_{ij}}$ has a sampling distribution

that is approximately¹ of the form of the χ^2 distribution with the degrees of freedom n equal to $k - c - 1$, k being the number of classes and c the additional restrictions imposed by the process of estimating the underlying probabilities. The number of these conditions is always equal to the number of the class probabilities originally estimated from the sample data (*i.e.*, the number of estimating equations). Another way of finding the degrees of freedom is to note the number of cells in the cross classification, or "contingency," table (such as Table 36) that can be filled in arbitrarily without changing the marginal totals. Since one cell must be left in each row and one in each column in order to make the frequencies in each row or column add up to the given marginal totals, it follows that a table of r rows and c columns will yield $(r - 1)(c - 1)$ degrees of freedom.

The problem in hand, then, may readily be solved as follows: From Tables 36 and 38, it is found that

¹Not only is the χ^2 distribution derived from the multinomial distribution by approximation, but it must be recalled that the derivation of the multinomial distribution itself is based on the assumption of sampling with replacements; when this is not true in practice, as is usually the case, the multinomial distribution gives merely approximate probabilities that are sufficiently accurate only for samples that are small relative to the population.

$$\sum \frac{(N_{ij} - N\check{p}_{ij})^2}{N\check{p}_{ij}} = \frac{(380 - 420)^2}{420} + \frac{(150 - 120)^2}{120} + \frac{(70 - 60)^2}{60} \\ + \frac{(320 - 280)^2}{280} + \frac{(50 - 80)^2}{80} + \frac{(30 - 40)^2}{40} = 32.44$$

Examining the χ^2 table for $n = 6 - 4 = 2$, it is seen that $\chi^2 = 32.44$ lies far beyond the .05 point. Hence, on the assumption that the upper .05 tail of the distribution is taken as the region of rejection, the hypothesis of independence cannot be accepted in this case; and it may be concluded that there is some association between color and sex in the given city. The general nature of that association is indicated by the sample. This suggests that the percentage of black and other-color males is greater than the percentage of white males, or, to put it another way, that the percentage of white females is greater than the percentage of black females or the percentage of other-color females.

Testing Homogeneity. Tests of independence that are used to determine the homogeneity of a sample are of sufficient importance to warrant special mention. Suppose, for example, that intelligence tests are given to 1,000 students in a coeducational university and also to 1,000 students in a university for men students only. Suppose that the distribution of grades in university *A* is quite different from that of university *B*, the latter showing in general a tendency toward higher grades. Because of this result, authorities in *B* infer that they are getting a more intelligent type of student than university *A*. Authorities in *A*, however, claim that the comparison is not fair, since, according to their contention, the women students in *A* are not generally as intelligent as the men students and the sample from *A* is consequently not a homogeneous one. Fortunately, it is possible to test the claim of the *A* authorities by applying the test of independence.

Suppose that the 1,000 grades from university *A* are cross classified with respect to intelligence grade and sex, with the results as shown in the table on page 339.

The question to be answered by these data is: Is the distribution of intelligence grades in university *A* independent of sex, or does sex make a difference? That is, can the distributions of men's and women's grades reasonably be taken to be two samples from a single homogeneous population? If they can, then authorities in *A* are probably wrong; at least the test of

TABLE 39.—1,000 STUDENTS CLASSIFIED AS TO SEX AND INTELLIGENCE GRADES

Grades	Men	Women	Totals
91-100	26	27	53
81-90	84	73	157
71-80	120	108	228
61-70	131	125	256
51-60	95	95	190
41-50	48	46	94
31-40	12	10	22
Totals	516	484	1,000

independence does not support their claim. If the distributions of men's and women's grades cannot reasonably be taken to be two samples from a single homogeneous population, then the authorities in *A* are right and the combination of the men's and women's grades produces a heterogeneous group that is not comparable with the presumably homogeneous group of grades from university *B*.

What do the data show? To answer this question would require a repetition of the sort of numerical work carried out in the previous example. The reader is therefore left to find out the answer for himself. He will find it a good exercise.¹

CAUTION: This chapter may well be concluded with a note of warning. It is always to be remembered that, when a hypothesis is *not disproved*, it is not necessarily *proved*. The χ^2 test of goodness of fit affords a good illustration of this. As already indicated, if a normal curve gives a good fit to a sample histogram, it is no proof that the sample came from a normal population. Any other curve that would give approximately the same theoretical frequencies as the normal curve would be an equally tenable hypothesis. Furthermore, it is to be noted that the χ^2 test takes no account of *sign*. A histogram may differ from a curve in a negative manner on one side of a central point and in a positive manner on the other. Still, if the differences are small, the χ^2 test may not reveal this obvious lack of conformity. If a hypothesis is not disproved by one test, it may thus be disproved by another. It is therefore well to examine a hypothesis from all angles before accepting it, even tentatively.

¹ The procedure is to set up the hypothesis of independence and from this and from the marginal totals to calculate a set of theoretical frequencies that may be compared with the actual frequencies by the χ^2 test.

CHAPTER XIV

JOINT SAMPLING FLUCTUATIONS IN MEAN AND STANDARD DEVIATION

Previous sampling analysis has concentrated attention on a single statistic and a single population parameter. Hypotheses regarding that parameter were tested and confidence limits established. Sometimes, however, problems arise in which two or more parameters are involved. For example, the question might be asked whether a given sample could have come from a normal population with a specified mean and a specified standard deviation.¹ A problem might also require joint confidence limits for the mean and standard deviation of a normal distribution. It is with these problems that the present chapter will be concerned. Similar questions could be asked with respect to normal bivariate or multivariate populations or to nonnormal populations in general, but these more difficult problems will not be discussed here.

DERIVATION OF JOINT SAMPLING DISTRIBUTION OF MEAN AND STANDARD DEVIATION

To answer questions about both the mean and the standard deviation of a normal population it is necessary to know the joint sampling distribution of the sample mean and standard deviation. This section will discuss the derivation of such a joint distribution.

¹ Previous analyses of sampling fluctuations in means and standard deviations were concerned with the following questions: (1) Given the standard deviation of the population as a known quantity, did the sample in hand come from a population in which the mean had some specified value? To answer this question the normal curve was used. (2) Regardless of what the standard deviation of the population might be, did the sample come from some population in which the mean has a specified value? Here the t distribution was used. (3) Regardless of what the mean of the population might be, did the given sample come from some population in which the standard deviation has some specified value? The answer to this question required the use of the χ^2 distribution.

In the present problem both the mean and the standard deviation are specified by hypothesis.

Numerical Illustration. For samples of 2 from a normal population, with $\bar{X} = 100$ and $\sigma = 10$, the joint distribution of the sample mean and sample standard deviation may readily be found from Fig. 93 (page 318). To find those samples, for example, whose means lie between 95 and 97 and whose standard

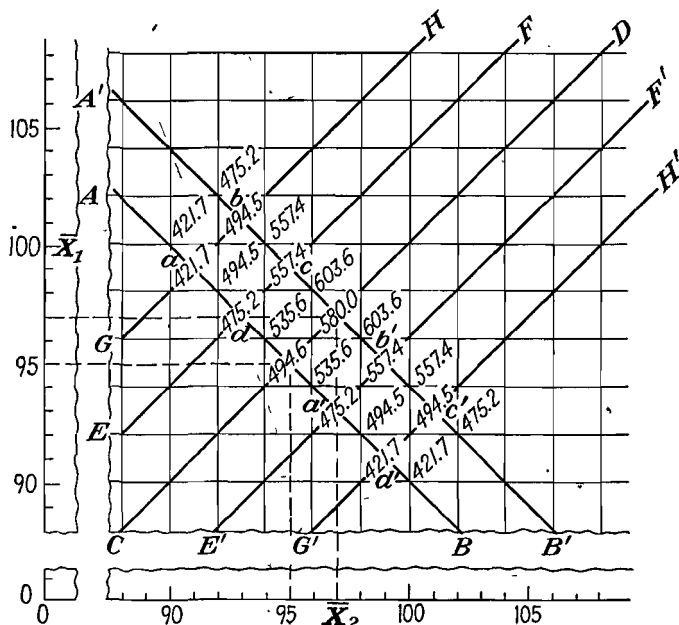


FIG. 96.—The sum of probabilities in cell $abcd$ is equal to

$$\frac{494.5}{100,000} + \frac{(\frac{1}{2})(494.5)}{100,000} + \frac{(\frac{1}{2})(421.7)}{100,000} + \frac{(\frac{1}{2})(475.2)}{100,000} + \frac{(\frac{1}{2})(557.4)}{100,000} = \frac{981.7}{100,000}$$

The sum of probabilities in cell $a'b'c'd'$ is the same. Hence the sum in both together is twice $\frac{981.7}{100,000}$ or $\frac{1963}{100,000}$. (The probabilities in the figure are expressed in $\frac{1}{100,000}$ ths.)

deviations lie between 2 and 4 (variances between 4 and 16), it is merely necessary to find the samples that lie between lines AB and $A'B'$ and also between the lines GH and EF and the lines $G'H'$ and $E'F'$. The procedure is illustrated in Fig. 96. In this figure the samples sought are those lying in squares $abcd$ and $a'b'c'd'$. The probability of these samples, as given by Fig. 93, is the probability of a sample with a mean between 95

and 97 and a standard deviation between 2 and 4. This probability can be entered in a joint distribution table such as is pictured in Fig. 97. When this has been done for all possible combinations of values for the mean and standard deviation, the result is that shown in Fig. 98. Figure 98 depicts the joint distribution of the mean and standard deviation of samples of 2. For larger samples the joint distribution of the mean and standard deviation looks more like Fig. 99.

A study of Fig. 98 will reveal that the distribution of sample standard deviations is the same in form, whatever the mean

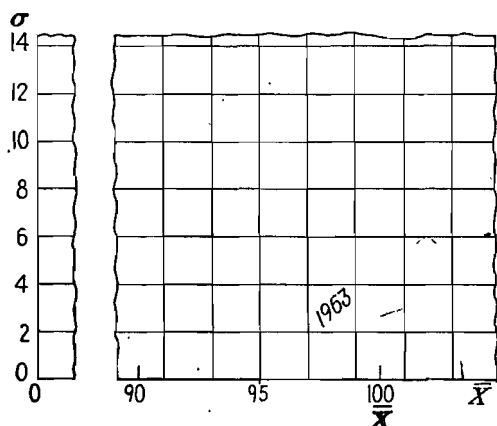


FIG. 97.—The probability of samples with a mean between 95 and 97 and a standard deviation between 2 and 4. (The probability in the figure is in $\frac{1}{100,000}$ ths.)

value of the sample. For samples with means lying between 115 and 117, for example, the probability of a sample having a standard deviation lying between 0 and 2 is equal to 198.6/100,000, the probability of a sample having a standard deviation lying between 2 and 4 is 183.6/100,000, the probability of a sample having a mean between 115 and 117 and a standard deviation lying between 4 and 6 is 156.8/100,000, etc. For samples with means lying between 117 and 119, the probabilities of various standard deviations are 101.2/100,000, 93.5/100,000, 79.9/100,000, etc. These are proportional to the former probabilities.

That is, $198.6/101.2 = 183.6/93.5 = 156.8/79.9 = \text{etc.}$, which indicates that the two distributions of sample standard devia-

tions are the same in form. This is true for all columns of Fig. 98, indicating that the distribution of sample standard deviations is independent of the value of the sample mean. This is true in general. For a normal population the sampling distribution of sample standard deviations is thus independent of the value of the sample mean, whatever the size of the sample.¹

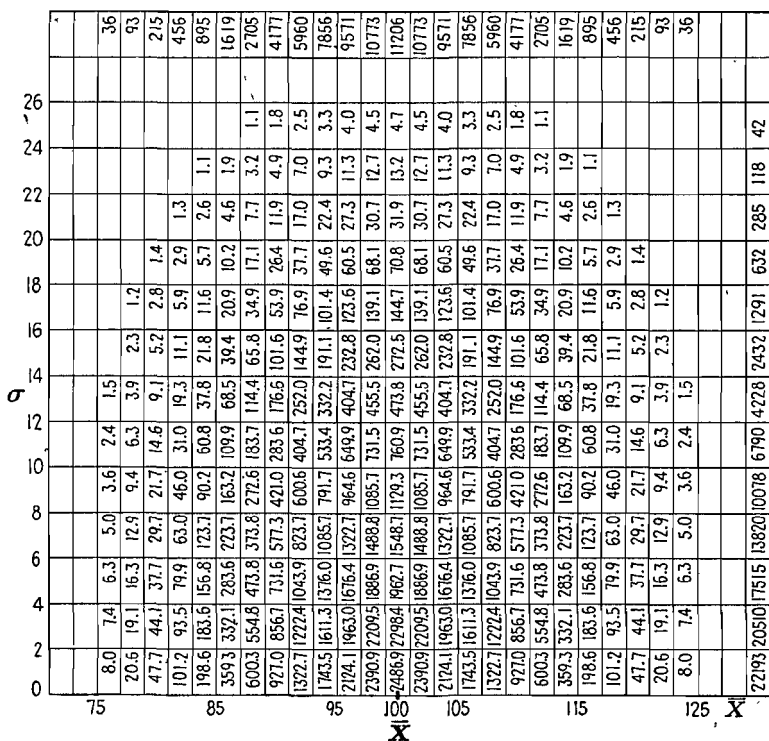


FIG. 98.—Probability set of sample means and sample standard deviations.
 $\Sigma(P) = 100,000$. $N = 2$, $n = 1$.

The sample standard deviation being independent of the sample mean, the joint probability of any given sample mean and any given sample standard deviation is just the product of their two individual probabilities. The equation for this joint distribution is therefore the product of Eqs. (1) and (3) of Chapter X. Accordingly, the joint distribution of the mean and standard deviation is given by

¹ See Appendix to Chap. X, p. 263.

$$dP(\bar{X}, \sigma) = \frac{1}{\delta \bar{x} \sqrt{2\pi}} \frac{N^{\frac{N-1}{2}} \sigma^{N-2}}{2^{\frac{N-3}{2}} \left(\frac{N-3}{2}\right)! \delta^{N-1}} \exp \left[-\frac{(\bar{X} - \bar{\bar{X}})^2}{2\delta^2 \bar{x}} \right] \exp \left[-\frac{N\sigma^2}{2\delta^2} \right] d\bar{X} d\sigma \quad (1)$$

USE OF JOINT SAMPLING DISTRIBUTION OF MEAN AND STANDARD DEVIATION

Testing Hypotheses. The use of the joint sampling distribution of mean and standard deviation for testing hypotheses may be illustrated with reference to Fig. 99, which gives the joint distribution of the mean and standard deviation for random samples of 11 from a normal population in which the mean is equal to 100 and the standard deviation is equal to 10. The question to be considered will be this: If a certain sample of 11 cases has a mean of 95 and a standard deviation of 13, is it reasonable to suppose that it came from a population whose mean is 100 and whose standard deviation is 10?

Choice of Region of Rejection. In answering this question the first step is to adopt a coefficient of risk that will determine the risk of rejecting the hypothesis when it is true. Let this be set at .05. The second step is to choose an appropriate region of rejection. This requires careful study.

Suggested Regions. Various regions of rejection suggest themselves. First are regions of rejection based entirely on the sample mean. Figure 100 shows a balanced region of this kind (call this region Ia), and Fig. 101 shows a region covering only the lower mean values (call this region Ib). A third region of this sort might cover only higher mean values but is not here depicted by a figure (call this region Ic).

A similar set of regions could be based entirely on the sample standard deviations. Figure 102 shows a balanced region of this kind (call this region IIa), and Fig. 103 shows a region covering only larger standard deviation values (call this region IIb).¹ A third region (not shown graphically) might cover only lower values of σ (call this region IIc).

¹ For the standard deviation the .02 and .98 points of the χ^2 distribution are used because the table of χ^2 is so set up that these values, and not the .025 points, can be obtained; the total probability for the regions of Fig. 102

A third set of regions could be based on the statistic

$$t = \frac{\sqrt{N} (\bar{X} - \bar{X})}{\bar{\sigma}}$$

which involves both the sample mean and the sample standard deviation.¹ To note how such regions are set up, consider the following relationships: In Fig. 99, $\bar{X} - \bar{X}$ is the distance of the sample from the vertical line through the population mean value $\bar{X} = \bar{X}$. Likewise, $\bar{\sigma}$ is equal to $\sigma \sqrt{\frac{N}{N-1}}$ and is thus equal to $\sqrt{\frac{N}{N-1}}$ multiplied by the distance of the sample from the horizontal axis, *i.e.*, the $\bar{X}$ -axis.

The statistic $\frac{\sqrt{N} (\bar{X} - \bar{X})}{\bar{\sigma}}$ is thus proportional to the cotangent of the angle that the line connecting the sample point with the point $(\bar{X}, 0)$ makes with the horizontal axis (angle β of Fig. 104).

A balanced region of rejection based on the sample value of $t = \frac{\sqrt{N} (\bar{X} - \bar{X})}{\bar{\sigma}}$ would thus look like that marked off in Fig. 104 (call this region IIIa). A similar region based solely on negative values of t is shown in Fig. 105 (call this region IIIb). A third region could be based solely on positive values of t but is not here depicted by a figure (call this region IIIc).

A fourth type of region that is considered (call it region IV) is a region based on a special set of contours known as " λ contours."²

The quantity λ is called the "likelihood ratio." The likelihood of a sample, it will be recalled, is the probability of that sample on the basis of certain assumed values of the population

is .04 and not .05. While the regions of rejection based on the standard deviation are thus not strictly comparable with those based on the mean, their position is good enough to illustrate the idea of regions of rejection.

¹ $\bar{\sigma}$ is the optimum estimate of σ , based on the sample standard deviation and is equal to $\sigma \sqrt{\frac{N}{N-1}}$.

² Cf NEYMAN, J., and E. S. PEARSON, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, Vol. 20A (1928), pp. 175-240.

0	93.1	93.7	94.3	94.9	95.5	96.1	96.7	97.3	97.9	98.5	99.1	99.7	100.3	100.9	101.5	102.1	102.7	103.3	103.9	104.5	105.1	105.7	106.3	106.9	$\bar{X}$
2																									
4																									
6																									
8																									
10																									
12																									
14																									
16																									
18																									
20																									
22																									
σ	714	1,085	1,585	2,224	2,999	3,886	4,839	5,790	6,657	7,355	7,808	7,966	7,808	7,355	6,657	5,790	4,839	3,886	2,999	2,224	1,585	1,085	714	100,000	

Fig. 99.—Probability set of sample means and sample standard deviations. $\bar{X} = 100$, $\sigma = 10$. $N = 11$, $n = 10$. $\Sigma(P) = 100,000$.

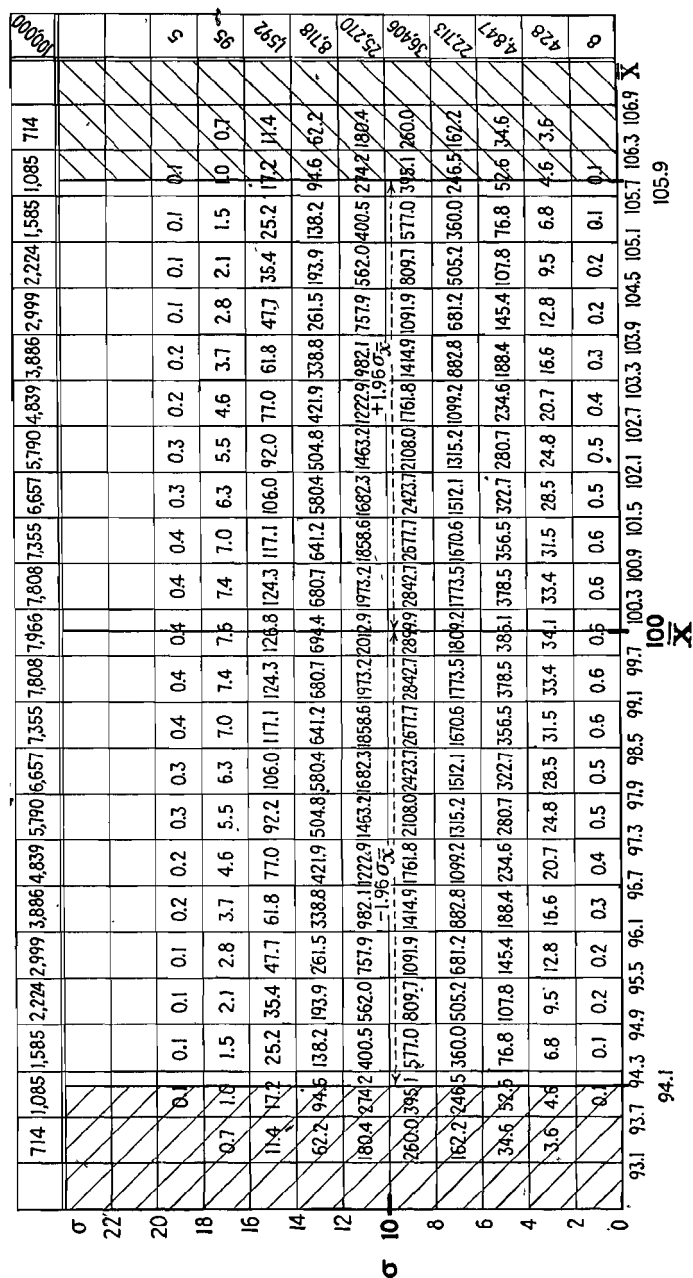


Fig. 100.—Region Ia.

σ	714	1085	1585	2,224	2,999	3,886	4,839	5,790	6,657	7,355	7,808	7,966	7,808	7,355	6,657	5,790	4,839	3,886	2,999	2,224	1,585	1,085	714	10000
22																								
20																								
18																								
16																								
14																								
12																								
10																								
8																								
6																								
4																								
2																								
0																								
	93.1	93.7	94.3	94.9	95.5	96.1	96.7	97.3	97.9	98.5	99.1	99.7	100.3	100.9	101.5	102.1	102.7	103.3	103.9	104.5	105.1	105.7	106.3	106.9
	$\bar{X}$																							
	95.04																							

Fig. 101.—Region Ib.

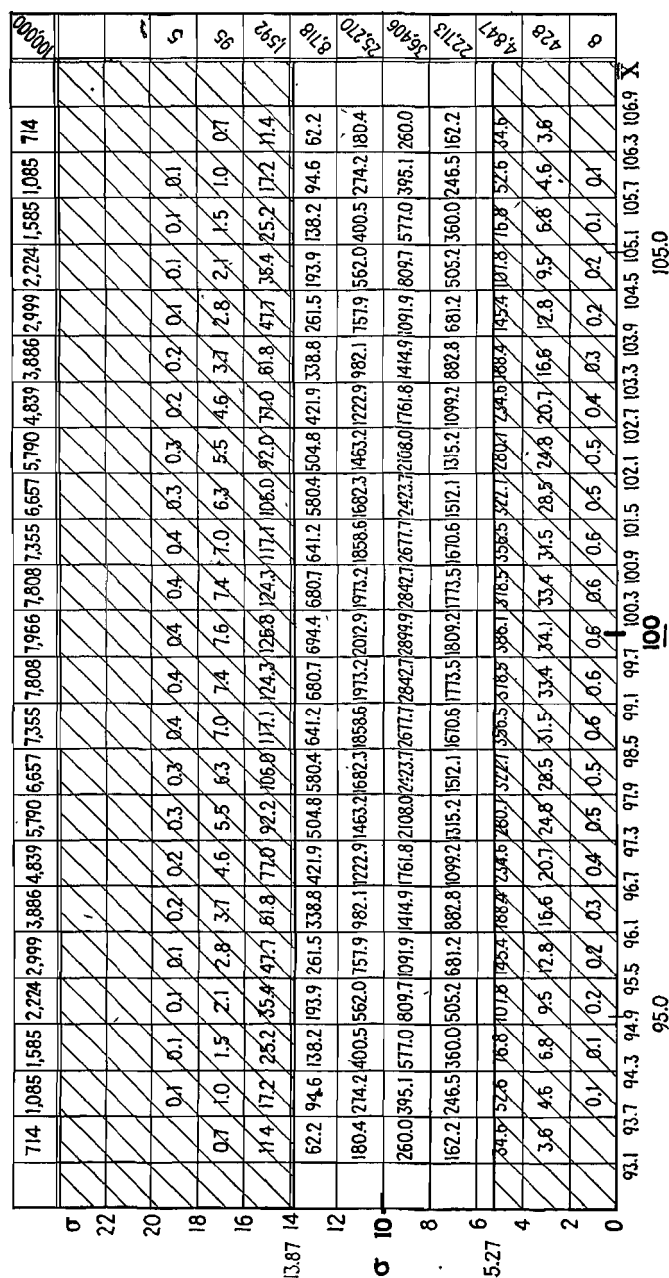
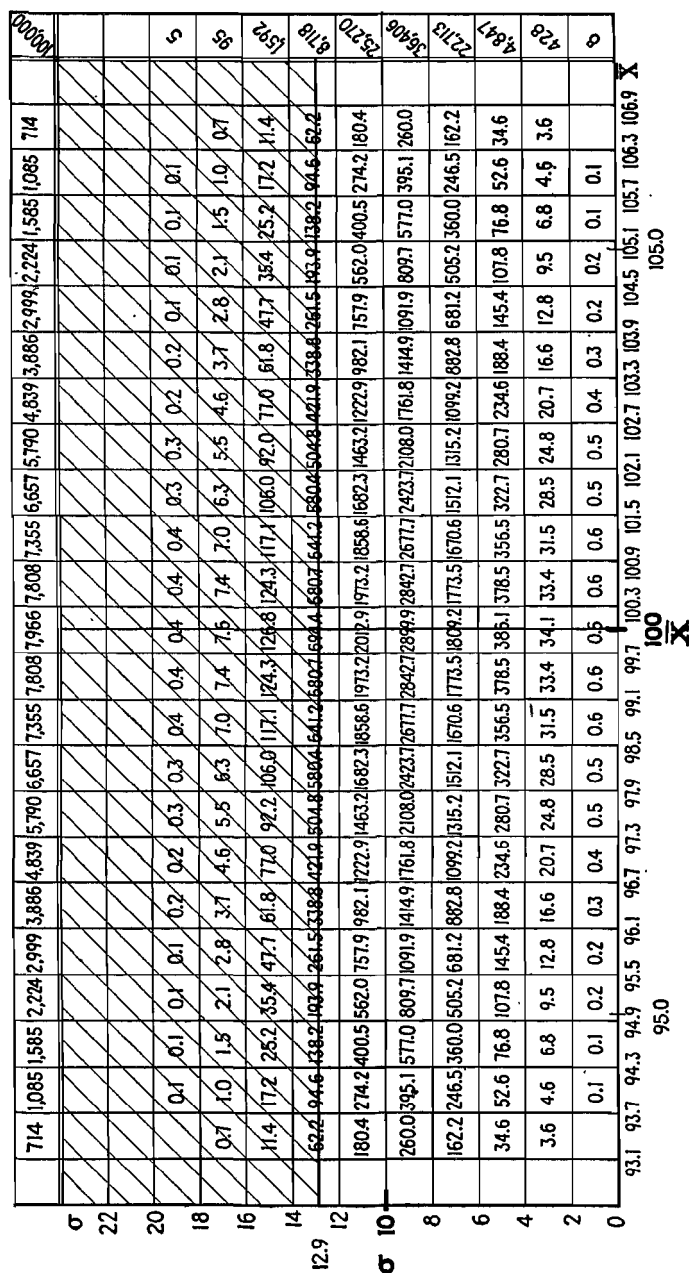


FIG. 102.—Region IIa.



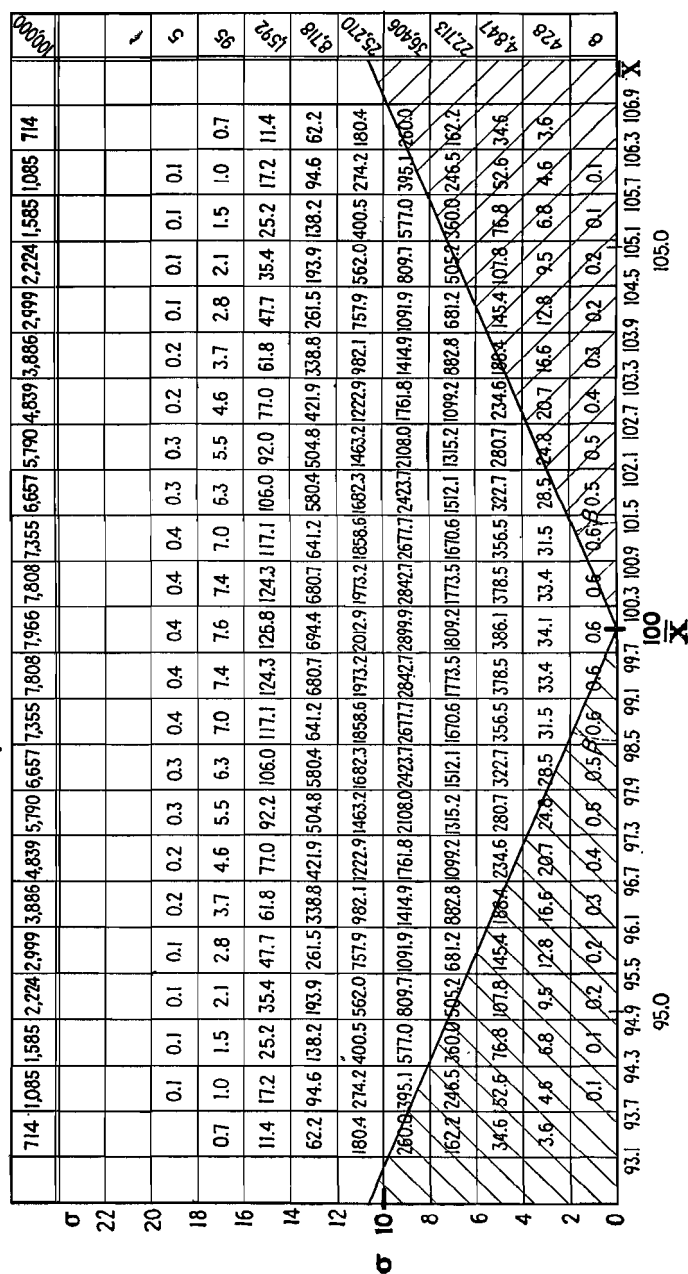


Fig. 104.—Region IIIa.

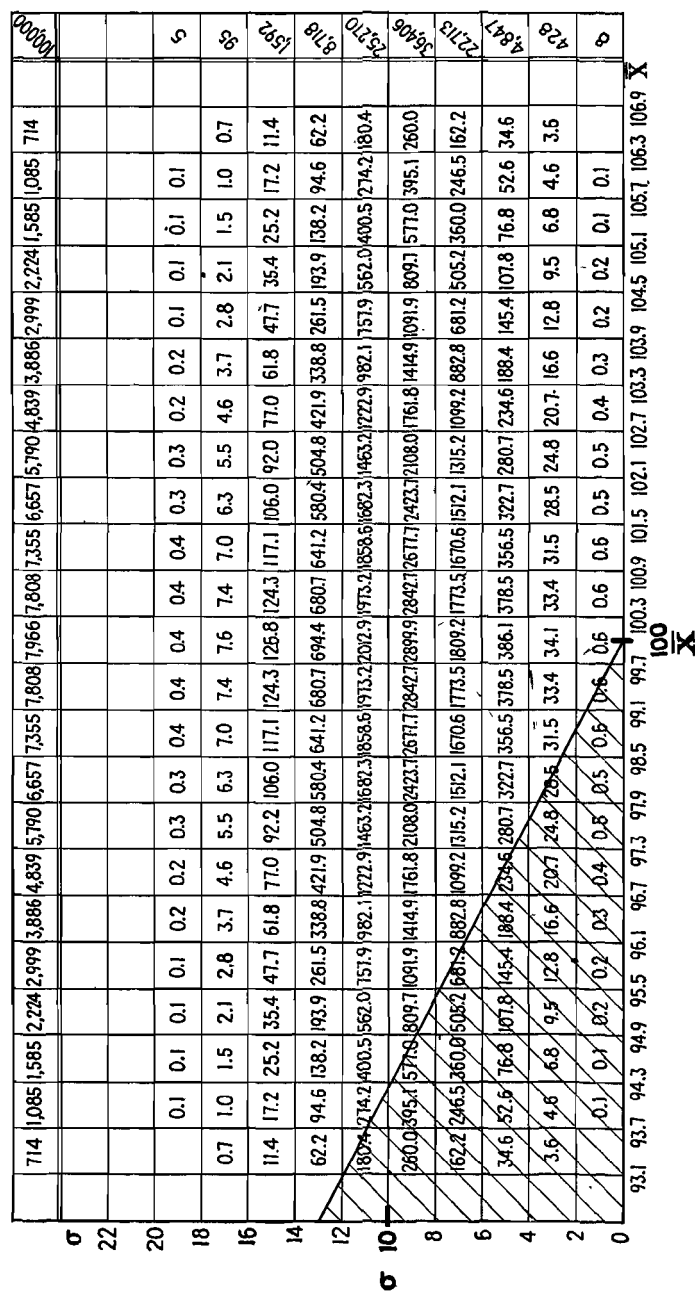


Fig. 105.—Region IIIb.

parameters. The likelihood ratio that Profs. Neyman and Pearson suggest is the ratio of the likelihood of the given sample on the assumption that the population has the mean and standard deviation given by the hypothesis in question to the likelihood of the given sample on the assumption that the population mean and standard deviation are equal to the joint maximum-likelihood estimates of these parameters. This ratio is given by

$$\lambda = S^N e^{-\frac{N}{2}(M^2 + S^2 - 1)} \quad (2)$$

or

$$\log_{10} \lambda = \frac{N}{2} [\log_{10} S^2 - (M^2 + S^2 - 1)] (.4343) \quad (3)$$

where $M = \frac{\bar{X} - \bar{X}}{\sigma}$, $S = \frac{\sigma}{\sigma}$, and $.4343 = \log_{10} e$.

Equation (3) may be written more succinctly as follows:

$$\log_{10} \lambda = \frac{N}{2} (.4343 - k) \quad (4)$$

$$\text{in which} \quad k = .4343(M^2 + S^2) - \log_{10} S^2 \quad (5)$$

This is the equation for the λ contours.

TABLE 40.—ILLUSTRATING THE CALCULATIONS NECESSARY FOR GRAPHING λ -CONTOUR REGION OF REJECTION (Region IV)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
σ	σ^2	$\log \sigma^2$	$\log \sigma^2$ - 1.3050	$\frac{(4)}{.004343}$	(5) - σ^2	$\sqrt{(6)}$	$\bar{X} = 100 \pm (7)$	
5.5	30.25	1.4807	.1757	40.46	10.21	± 3.3	103.2	96.8
6.0	36.00	1.5563	.2513	57.86	21.86	± 4.7	104.7	95.3
7.0	49.00	1.6902	.3852	88.69	39.69	± 6.3	106.3	93.7
8.0	64.00	1.8062	.5012	115.40	51.40	± 7.2	107.2	92.8
9.0	81.00	1.9085	.6035	138.96	57.96	± 7.6	107.6	92.4
10.0	100.00	2.0000	.6950	160.03	60.03	± 7.8	107.8	92.2
11.0	121.00	2.0828	.7778	179.09	58.09	± 7.6	107.6	92.4
12.0	144.00	2.1584	.8534	196.50	52.50	± 7.2	107.2	92.8
13.0	169.00	2.2279	.9229	212.50	43.50	± 6.6	106.6	93.4
14.0	196.00	2.2923	.9873	227.33	31.33	± 5.6	105.6	94.4
15.0	225.00	2.3522	1.0472	241.12	16.12	± 4.0	104.0	96.0
15.5	240.25	2.3807	1.0757	247.68	7.43	± 2.7	102.7	97.3

The calculations shown in Table 40 are solutions of an equation derived from Eq. (5). By interpolating in Table 41 (page 361) for the value of k , for which the P_λ is .05, k is found to be .695. Hence, by Eq. (5), if σ equals 10 and $\bar{X}$ equals 100,

$$.695 = .4343 \left[\frac{(\bar{X} - 100)^2}{100} + \frac{\sigma^2}{100} \right] - \log \sigma^2 + \log 100$$

and since $(\bar{X} - 100)^2 = \bar{x}^2$

$$.695 = .004343\bar{x}^2 + .004343\sigma^2 - \log \sigma^2 + 2$$

$$\bar{x}^2 = -\sigma^2 + \frac{1}{.004343} (\log \sigma^2 - 1.305)$$

$$\bar{X} = 100 \pm \sqrt{\frac{\log \sigma^2 - 1.305}{.004343} - \sigma^2}$$

This last is the form in which solutions are found for $\bar{X}$ corresponding to the specified values for σ shown in column (1) of Table 40. The corresponding values for $\bar{X}$ are shown in the two subcolumns of column (8). The graph of columns (1) and (8) gives the elliptical region depicted in Fig. 106. This is a picture of region IV, which is a λ -contour region.

A final set of regions that may be suggested is a type of region that cuts off a corner section of the distribution of sample means and standard deviations. Such regions are pictured in Figs. 107 to 109. These will be called, as a class, regions Va , Vb , and Vc .

Discussion of the Various Regions of Rejection. The reason why so many regions of rejection are suggested as possible alternatives is that the best region to choose in any particular instance depends on the circumstances of the problem. The regions suggested here include those likely, under varying circumstances, to be found most useful. Their relative advantages and disadvantages may now briefly be considered.

First note that all regions having a probability of .05 would lead to a rejection of the hypothesis 5 times out of 100 when it was true. Region I has the disadvantage that it would permit acceptance of hypotheses in cases in which the sample was very improbable because the σ^2 's were very large or very small. Similarly, region II would have the disadvantage of leading to the acceptance of hypotheses in cases in which the mean was far distant from the hypothetical mean of the population. Region III, based on both sample means and sample standard deviations

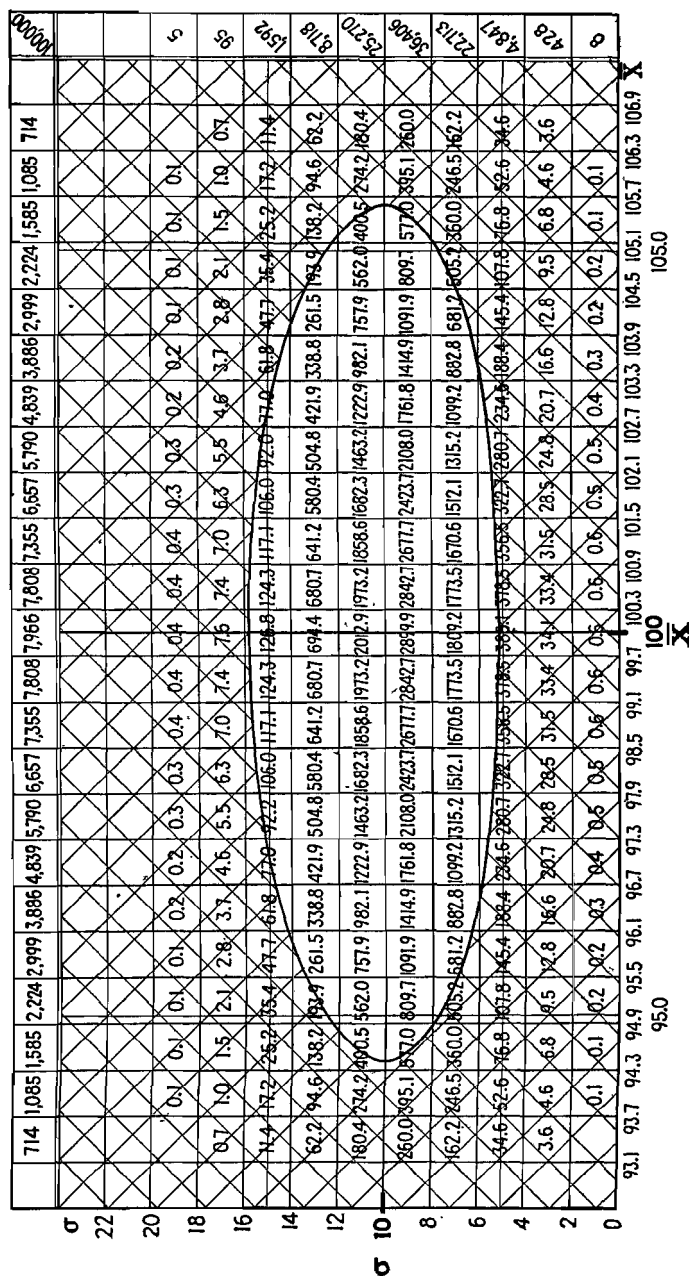
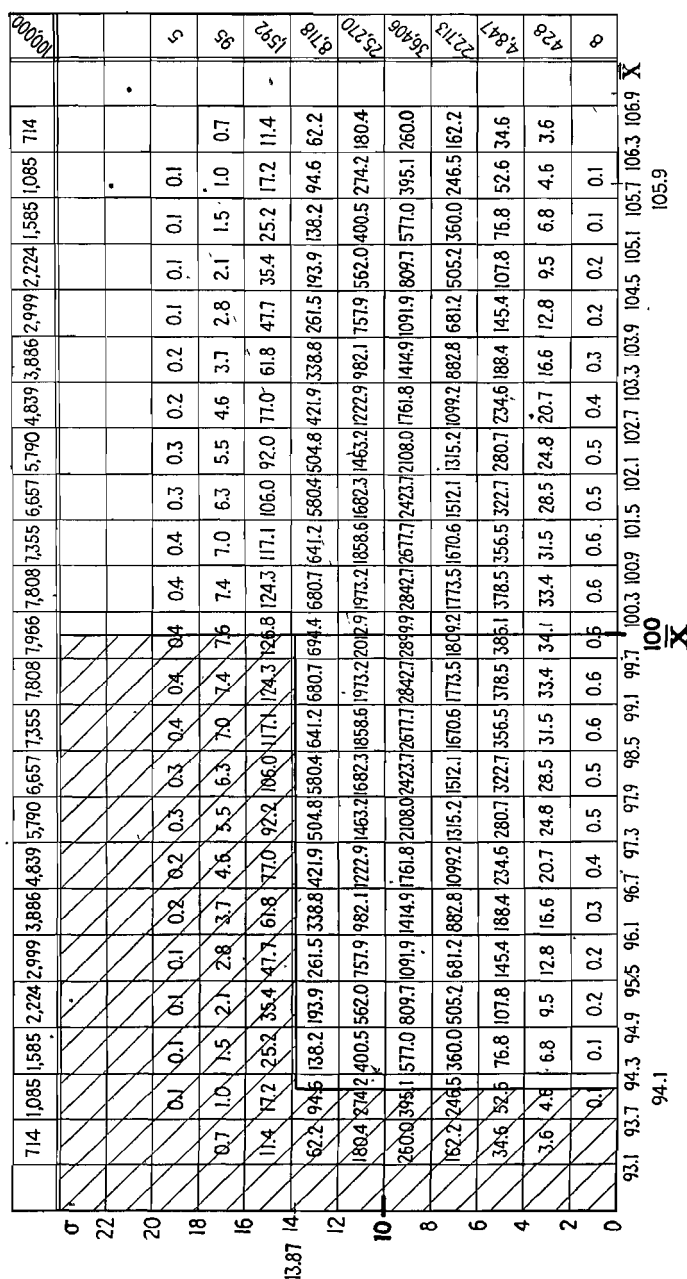


FIG. 106.—Region IV.



might presumably be thought to avoid this disadvantage of regions I and II of including improbable samples; but unfortunately this does not follow. As may be seen from Fig. 104, region III_a accepts the hypothesis in cases where the standard deviation is very large because the mean happens to be very small or very large at the same time, although both are very unlikely on the basis of a given hypothesis. Also, region III_a accepts the hypothesis when the sample standard deviation is very small because the mean is also very close to the hypothetical value.

Region IV avoids the disadvantage of accepting highly improbable hypotheses in that it has a certain elliptical symmetry about the center of the distribution of sample means and variances.

All these regions, it should be repeated, will lead to the rejection of the hypothesis when it is true 5 per cent of the trials; but only regions like IV will lead to rejection of the hypothesis in every case in which the sample itself is very improbable on the basis of that hypothesis.

More important, however, than the criterion that a region should include the more improbable samples is the criterion that a region should lead to rejection of the hypothesis most frequently when the hypothesis is not true. Which region fares best on this more important criterion will depend on the nature of the problem in hand. If the investigator desires to avoid acceptance of the hypothesis especially when the true value of the mean is less than the hypothetical mean being tested, region Ib should be used, for this will lead to rejection of the hypothesis more often when the true mean is less than the hypothetical mean.¹ Region III_b is also a good region to use in this case, but it does not appear to be as good as region Ib, based only on the means.

On the other hand, if the investigator wishes to avoid acceptance of the hypothesis especially when the true value of σ is greater than the hypothetical value, region II_b should be used; for when σ is increased, the distribution is stretched upward and its vertical mean moves higher. In this instance, region III_a, for certain values of σ , and region II_c, for all high values,

¹ For as the distribution is shifted to the left by decreasing $\bar{X}$, a larger percentage of the samples tends to fall in the established region of rejection.

are worse than useless, since if the true σ were greater than the hypothetical σ the probability of a sample falling in the region of rejection would be less than if the hypothetical σ were the true σ . Thus, if region IIc is used under these circumstances, there is more chance of accepting the hypothesis when it is not true (because σ is actually larger) than there would be if it were true. On the other hand, region IIIa, for certain values of σ , or region IIc, for all lower values, is not so bad in cases in which the true σ is less than the hypothetical σ .

Instances in which region IIIa or IIIb appear to be particularly useful are those in which the investigator wishes especially to reject the hypothesis when in fact the true mean is less or greater than the hypothetical mean and at the same time the true σ is less than the hypothetical σ . For in this instance the distribution is distorted so that more of the samples would fall in the established region of rejection.

Region IV is a compromise region that tends to give a high probability of rejection in whatever manner the true means and standard deviations differ from their hypothetical values. For such instances Region IV has been offered as an especially good region.

The region of rejection that would be best in most instances would probably be one that maximized the probability of rejecting the given hypothesis when in fact the true mean was less and the true σ greater than the hypothetical values. A manufacturer of tires, for example, would be concerned about getting tires that rendered less mileage on the average than desired and varied more from tire to tire. Of course, other cases would arise in which the investigator would wish to avoid acceptance of the hypothesis when in fact the mean was greater than the hypothetical value and the standard deviation was greater or less. It is believed, however, that the former instance is more common.

A region bounded by a smooth continuous curve would appear to be the best type of region to adopt for the reason explained in the preceding paragraph. Such a region is depicted in Fig. 107. But to find a curve that would permit the ready calculation of probabilities is not so easy. Probably the simplest region would be a rectangular one such as that pictured in Fig. 108. A method of defining such a region in practical cases will be discussed below. A region like that pictured in Fig. 109

would involve the same difficulties as one like that shown in Fig. 107.

TABLE 41.—PROBABILITY OF A VALUE OF λ AS GREAT AS, OR GREATER THAN SPECIFIED VALUE FOR VARIOUS VALUES OF $k = .4343 - \frac{N}{2} \log \lambda^*$

$N = 3$		$N = 4$		$N = 5$	
k	P_λ	k	P_λ	k	P_λ
1.50	.0830	1.25	.0578	1.05	.0568
1.80	.0415	1.30	.0486	1.10	.0450
2.40	.0106	1.50	.0243	1.40	.0113
2.70	.0053	1.80	.0086	1.45	.0090
$N = 6$		$N = 7$		$N = 8$	
k	P_λ	k	P_λ	k	P_λ
.90	.0666	.85	.0551	.80	.0512
.95	.0499	.90	.0389	.85	.0341
1.20	.0117	1.05	.0137	1.00	.0101
1.25	.0088	1.10	.0097	1.05	.0068
$N = 9$		$N = 10$		$N = 11$	
k	P_λ	k	P_λ	k	P_λ
.75	.0534	.70	.0625	.65	.0822
.80	.0336	.75	.0371	.70	.0461
.90	.0133	.85	.0131	.80	.0145
.95	.0084	.90	.0078	.85	.0081

* Abridged from more elaborate tables in J. Neyman and E. S. Pearson, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, Vol. 20A (1928), pp. 175-240, 238-240.

Illustrations. The two regions whose use will be illustrated here will be region IV, based on the λ contours, and region Vb. These regions are the ones that would probably be used in most instances when a joint hypothesis was being tested. Region IV would be used when the investigator is indifferent to what the actual values of $\bar{X}$ and σ might be and region Vb would be used when the investigator is especially troubled about these values deviating from the hypothetical values in a particular direction.

Regions based entirely on either the mean or the standard deviation alone would probably not be used in cases in which a joint hypothesis is being tested. It is unlikely that the investigator would be at all concerned about the actual value of the standard deviation, say, when the hypothesis does involve this quantity.

TABLE 42.—VALUES OF THE RATIO P_λ/λ FOR VARIOUS VALUES OF k^*

k	$\frac{P_\lambda}{\lambda}$	k	$\frac{P_\lambda}{\lambda}$	k	$\frac{P_\lambda}{\lambda}$
.435	1.0008	.520	1.0965	.605	1.2024
.440	1.0062	.525	1.1024	.610	1.2089
.445	1.0116	.530	1.1084	.615	1.2155
.450	1.0170	.535	1.1144	.620	1.2221
.455	1.0224	.540	1.1205	.625	1.2287
.460	1.0279	.545	1.1266	.630	1.2353
.465	1.0334	.550	1.1328	.635	1.2420
.470	1.0390	.555	1.1390	.640	1.2488
.475	1.0446	.560	1.1452	.645	1.2555
.480	1.0502	.565	1.1514	.650	1.2623
.485	1.0559	.570	1.1576	.700	1.332
.490	1.0616	.575	1.1639	.750	1.407
.495	1.0673	.580	1.1702	.800	1.485
.500	1.0731	.585	1.1766	.850	1.569
.505	1.0789	.590	1.1830	.900	1.658
.510	1.0847	.595	1.1894	.950	1.753
.515	1.0906	.600	1.1959	1.000	1.855

* Reproduced from Table XII in J. Neyman and E. S. Pearson, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, Vol. 20A (1928), pp. 175-240, 235.

The Probability Distribution of λ . To make use of λ contours for regions of rejection, the probability distribution of λ must be determined. This has been done by Neyman and Pearson and the results published in a table of probabilities. These are condensed in Tables 41 and 42. For $N = 3$ to $N = 11$ and for various values of k , the probability P_λ of getting as great or greater values of λ by chance is given in Table 41. These have been taken from a more elaborate table worked out by Neyman and Pearson. For larger values of N , Table 42 is found useful. This gives, for various values of k , the ratio P_λ/λ , affording a set of relationships that is practically independent of N .¹

¹ It will be noted that P_λ means the probability of as great or greater value

Testing a Hypothesis with a λ Contour. To make use of these fundamental relationships in an actual problem proceed as follows: Calculate $M = \frac{(\bar{X} - \bar{X})}{s}$ and $S = \frac{\sigma}{s}$, and

$$k = .4343(M^2 + S^2) - \log_{10} S^2.$$

If N is equal to or less than 10, use Table 41 to note whether the probability of as great or greater λ $\left[\text{given by } \log \lambda = \frac{N}{2}(.4343 - k) \right]$ is less than or equal to .05 (or .01 if this is taken as the coefficient of risk). If it is less than or equal to the coefficient of risk, then the given sample must fall in the region of rejection marked off by the .05 (or .01) λ contour. If N is greater than 10, find from Table 42 the value of P_λ/λ corresponding to the value of k computed for the sample. Then compute $\log \lambda = \frac{N}{2}(.4343 - k)$, and multiply P_λ/λ by this value of λ . The result is P_λ , the probability of as great or greater value of λ . If this is less than the adopted coefficient of risk, then the sample falls in the region of rejection marked off by the λ contour.

For illustration this procedure may be applied to the problem stated above on page 344. In the sample, $N = 11$, $\bar{X} = 95$, and $\sigma = 13$. It is desired to test the hypothesis that the mean of the population from which this sample was drawn is 100 and its standard deviation 10. The investigator is indifferent as to whether the actual values of the population mean and population standard deviation are greater or less than 100 and 10, respectively, and he decides upon a coefficient of risk of .05.

For these data, $M = \frac{95 - 100}{10} = -.5$, and $S = \frac{13}{10} = 1.3$. Hence, $M^2 = .25$, $S^2 = 1.69$, and

$$k = (.4343)(.25 + 1.69) - .2279 = .6146.$$

Looking up this value of k in Table 41 it is found that, for $N = 11$ and $k = .65$, $P_\lambda = .0822$. Hence, for $k = .6146$, P_λ must be greater than .08 and certainly greater than .05. Accordingly,

of λ . If it had been written $P(\lambda)$ it would have meant simply the probability of λ . Thus P_λ is the probability of a range of values beginning at λ and running to infinity.

the sample in question does not fall in the region of rejection, and the hypothesis is not rejected.

Since $N = 11$, Table 42 could have been used instead of Table 41. Interpolation in Table 42 shows that, for $k = .6145$, $P_\lambda/\lambda = 1.1969$. For this given value of k , the value of $\log \lambda$ is $(\frac{1}{2})(.4343 - .6146) = -.9922$, or $9.0078 - 10$. This gives $\lambda = .1018$. Multiplying the value of P_λ/λ by this value of λ gives $P_\lambda = (1.1969)(.1018) = .1218$, which corroborates the previous conclusion that P_λ was greater than .08.

The calculations in Table 40 have illustrated the construction of a λ -contour region of rejection for the testing of a hypothesis as to the mean and standard deviation of the population. The resulting λ -contour region of rejection (region IV) is graphically depicted in Fig. 106. It will be noted that the λ contour is a symmetrical ellipse with the intersection of $\bar{\sigma}$ and $\bar{X}$ as its center.¹ Table 40 and Fig. 106 are presented as an aid in visualizing the principle involved in these problems; such a table and figure need not, of course, be constructed for an actual problem. The problem illustrated above was solved without the use of such a figure.

Use of a Corner Region Illustrated. The foregoing has been based on the assumption that the investigator is indifferent as to whether the true values of the mean and standard deviation are greater or less than the hypothetical values being tested. If an investigator is more concerned with the possibility that the true mean is less, say, than the hypothetical mean being tested and that the true standard deviation is greater than the hypothetical standard deviation, he will have the least probability of accepting the hypothesis when the true values actually bear this relationship to the hypothetical values if he locates his region of rejection entirely in the upper left-hand quadrant of the distribution of samples. Logically, it would seem appropriate to find the λ contour that marked off a .05 region in this upper quadrant. Unfortunately, this presents such great mathematical difficulties that the procedure is impracticable. The following discussion is therefore offered as a crude substitute:

¹ The ellipse obtained in this way is different from that of Fig. 111. The latter gives joint values for the population mean and standard deviation for given values of the sample mean and sample standard deviation (see p. 369).

The determination of a corner region of rejection may be explained with reference to the following problem: Suppose it is claimed that a new method of cultivating potatoes in a given locality will average a yield of 100 bushels per acre and a standard deviation of 10 bushels. Suppose that this offers a gain over the old method of cultivation in that the mean is greater and the standard deviation is less. To test this claim a farmer uses the new method of cultivation on 10 plots of ground of the same size and finds that the mean yield is 97 bushels and the standard deviation is 13.5 bushels. Can he reasonably reject the hypothesis that the average yield will in the long run be 100 bushels and the standard deviation 10 bushels?

If the true mean of the population is 100 bushels and the true standard deviation is 10 bushels, 50 per cent of samples of 10 will have an average yield equal to or less than 100. Likewise, if the true standard deviation is 10 bushels, 50 per cent of samples of 10 will have standard deviations equal to or greater than 9.15.* Accordingly, the probability of a sample of 10 having a mean less than 100 and a standard deviation greater than 9.15 is .25.

If two other values for the mean and standard deviation can be found such that 20 per cent of the samples have means lying between 100 and this second mean value and standard deviations lying between 9.15 and this second standard deviation value, these second values for the mean and standard deviation will mark off a .05 region in the upper quadrant that would appear to be a good region of rejection for the present problem. The square root of .20 is .4472. Therefore, if a second mean value can be found such that .4472 of the samples have means lying between this value of 100 and a second standard deviation value can be found such that .4472 of the samples have standard deviations lying between this second standard deviation value and 9.15, the two values for the mean and standard deviation so determined will mark off the desired region of rejection.

A table of normal probabilities shows that .4472 of the samples would fall between the mean and the mean less 1.618σ . Since the standard deviation of the mean is $\sigma/\sqrt{N}$, it would have the value $10/\sqrt{10} = 3.16$. Hence, .4772 of the samples would fall between 100 and $100 - 1.618(3.16) = 94.9$.

* For $n = 9$, probability is 50 per cent that a χ^2 will exceed 8.343; therefore, probability is 50 per cent that $N\sigma^2/\sigma^2 = 10\sigma^2/10^2$ will exceed 8.343 or that σ will exceed $\sqrt{83.43} = 9.15$.

Similarly, a table of the χ^2 distribution¹ shows that, for $n = 9$, .4772 of the cases lie between the median value 8.343 and the value $\chi^2 = 16.774$. Since $N\sigma^2/s^2$ has a sampling distribution of the form of the χ^2 distribution, this means that .4472 of the samples will have a σ lying between 9.15 and 12.95; for $10\sigma^2/100 = 16.774$ gives $\sigma = 12.95$.

Accordingly, the desired .05 region of rejection will be such as the shaded area of Fig. 110. It will include all samples whose

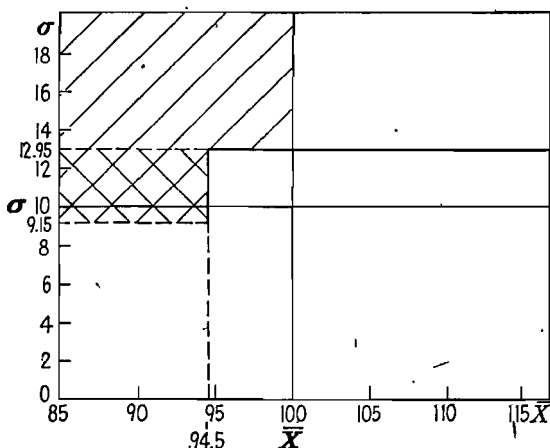


FIG. 110.—An example of a corner region of rejection.

means are less than 94.9 and whose standard deviations are greater than 9.15 and also all samples whose means lie between 94.9 and 100 and whose standard deviations are greater than 12.95. Inasmuch as the given sample has a mean of 97 and a standard deviation of 13.5, it falls in this region of rejection and the hypothesis must be rejected.

Determination of Joint Confidence Zone for the Mean and Standard Deviation. In the estimation of a single population parameter it was found possible to mark off a range of values in the neighborhood of the sample result that could be said to include the population value with a given degree of probability. Such a range of values was called a "confidence interval" for the given population parameter, and the probability of its including

¹ Pearson's *Tables for Biometricians and Statisticians* was used for this purpose since it permits a finer interpolation than the table included in the Appendix of this book.

the true value of this parameter was called the "confidence coefficient."

In a similar fashion, when two parameters are being estimated jointly, it is possible to lay off joint ranges of values that may be said to include the true population values of these parameters with a given degree of probability. This joint range of values is called a "confidence zone," and the probability associated with it is its confidence coefficient.

It will be recalled that the upper limit of the confidence interval for a single parameter was taken as the value of the parameter that would make the probability of the given sample result or a lower value just equal to a predetermined figure, say .025. Likewise, the lower confidence limit was the value of the parameter that would make the probability of the sample result or a higher value just equal to another predetermined quantity, again say .025. The sum of these two probabilities is equal to the selected coefficient of risk, and the complement of the sum is the confidence coefficient, say .95.

Another way of looking at it is that the upper confidence limit is so chosen that the sample would fall exactly on the upper boundary of the lower region of rejection, and the lower confidence limit is so chosen that the sample would fall exactly on the lower boundary of the upper region of rejection.

In order to determine a joint confidence zone for two parameters a similar procedure is possible. The limits of the zone are determined such that if the population parameters have any of these limiting values then the sample will fall on the boundary of the region of rejection. In the case of the mean and standard deviation of the population a confidence zone with a confidence coefficient of .95 would be determined by finding the values of these parameters that would make the given sample fall on the λ contour marking the .05 region of rejection. The equation for the λ contours, it will be recalled,¹ is as follows:

$$\log_{10} \lambda = \frac{N}{2} [\log_{10} S^2 - (M^2 + S^2 - 1)] (.4343) \quad (3)$$

or

$$k = .4343(M^2 + S^2) - \log_{10} S^2 \quad (5)$$

in which $k = .4343 - \frac{2}{N} \log_{10} \lambda$, $M = \frac{\bar{X} - \bar{X}^*}{\sigma}$, and $S = \frac{\sigma}{\hat{\sigma}}$.

¹ See p. 353.

If hypothetical values are assigned to $\bar{X}$ and σ and if λ (or k) is given the value that makes the probability of as great or greater λ just equal to .05, the equation above will represent the boundary line of the .05 region of rejection to be used in testing the given hypothesis. If, however, a particular sample value is substituted for $\bar{X}$ and σ and if λ (and hence k) is given its .05 value, then this equation will show what hypothetical values of $\bar{X}$ and σ would cause the given sample to fall exactly on the boundary line of the .05 region of rejection. When so used the equation becomes the formula for the boundary of the joint confidence zone for the mean and standard deviation.

Suppose, for example, that a random sample of 11 cases is found to have a mean of 95 and a standard deviation of 13. What is the joint confidence zone that may be said to cover the true values of the population mean and standard deviation with a probability of .95? For samples of 11, the value of k for which the probability of as great or greater λ is just .05 is found from interpolation in Table 41 to be approximately .695. On substituting this value of k and the given sample value of $\bar{X}$ and σ in Eq. (5), the boundary of the joint confidence zone for the mean and standard deviation of the population is given by the following,

$$.695 = .4343 \left[\frac{(95 - \bar{X})^2}{\sigma^2} + \frac{169}{\sigma^2} \right] - \log_{10} \left(\frac{169}{\sigma^2} \right)$$

which, on solving for $\bar{X}$ gives

$$\bar{X} = 95 \pm \sqrt{-169 + 2.3026\sigma^2(2.9229 - \log_{10} \sigma^2)}$$

By substituting various values for σ and obtaining the corresponding values of $\bar{X}$ a graph may be drawn of the boundary line of the desired confidence zone. This has been done in Table 43 and depicted in a graph in Fig. 111.

The result so obtained may be interpreted as follows: The set of $\bar{X}$, σ values included within the oval-shaped curve of Fig. 111 constitutes the .95 zone of confidence for the mean and standard deviation of the population. In other words, there is a chance of .95 that one of the pairs of values included within the curve is the pair whose values are those of the actual mean and standard deviation of the population. Accordingly, any pair of values included within this curve is a reasonable hypothesis as to the

true values of the mean and standard deviation of the population—it would be a hypothesis that would not cause the given sample to fall in the .05 region of rejection of Fig. 106.

Any pair of values outside the oval-shaped curve of Fig. 111 would be an unreasonable hypothesis as to the true value of the mean and standard deviation of the population. The curve shows that the maximum tenable value for the mean of the population is approximately 107 and the minimum tenable value is approximately 83. It also shows that the maximum tenable value for the standard deviation of the population is

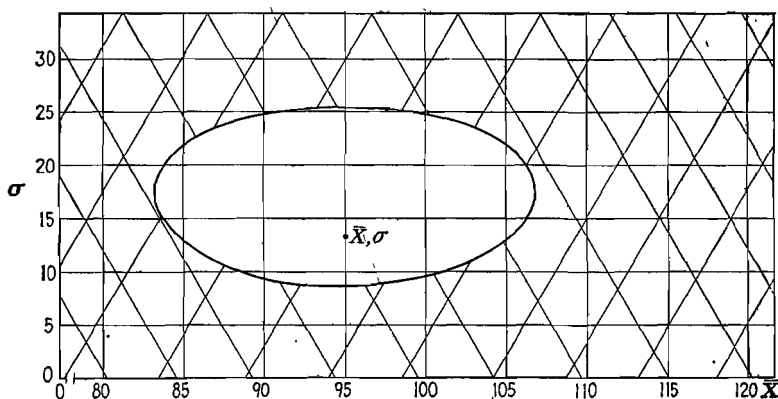


FIG. 111.—The .95 joint confidence zone for $\bar{X}$ and σ , given $\bar{X} = 95$, $\sigma = 13$, and $N = 13$.

approximately 25.3 and the minimum tenable value is approximately 8.8.

It is to be noted, however, that these extreme values for the mean and standard deviation of the population are not jointly tenable. If the mean should be assumed to have the value 107, the only tenable value for the standard deviation of the population would be 17.5. If the standard deviation should be assumed to have the value 25.3, the only tenable value for the mean of the population would be 95. Joint tenability, of course, is the very essence of the diagram. It shows what range of values for one parameter is jointly tenable with a given value of the other variable.

If the mean is assumed to be 85, for example, then the values of the standard deviation that are jointly tenable with this value of the mean are the values from 13 to 22. Likewise, if the

standard deviation is assumed to have the value 20, say, the values of the mean that are then jointly tenable are the values from 84 to 106.

TABLE 43.—ILLUSTRATING THE CALCULATIONS NECESSARY FOR THE GRAPHING OF A JOINT CONFIDENCE ZONE
For the mean and standard deviation of the population

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
d	d^2	$\log d^2$	$2.9229 - \log d^2$	$2.3026d^2$	$(5) \cdot (4)$	$(6) - 169$	$\sqrt{(7)}$	$95 \pm (8)$	
9	81	1.9085	1.0144	186.5106	189.20	20.20	± 4.5	99.5	90.5
10	100	2.0000	.9229	230.2600	212.51	43.51	± 6.6	101.6	88.4
11	121	2.0828	.8401	278.6146	234.06	65.06	± 8.1	103.1	86.9
12	144	2.1584	.7645	331.5744	253.49	84.49	± 9.2	104.2	85.8
13	169	2.2279	.6950	389.1394	270.45	101.45	± 10.1	105.1	84.9
14	196	2.2923	.6306	451.3096	284.60	115.60	± 10.8	105.8	84.2
15	225	2.3522	.5707	518.0850	295.67	126.67	± 11.3	106.3	83.7
16	256	2.4082	.5147	589.4656	303.40	134.40	± 11.6	106.6	83.4
17	289	2.4609	.4620	665.4514	307.44	138.44	± 11.8	106.8	83.2
18	324	2.5105	.4124	746.0424	307.67	138.67	± 11.8	106.8	83.2
19	361	2.5575	.3654	813.2386	303.73	134.73	± 11.6	106.6	83.4
20	400	2.6021	.3208	921.0400	295.47	126.47	± 11.2	106.2	83.8
21	441	2.6444	.2785	1015.4466	282.80	113.80	± 10.7	105.7	84.3
22	484	2.6848	.2381	1114.4584	265.35	96.35	± 9.8	104.8	85.2
23	529	2.7235	.1994	1218.0754	242.88	73.88	± 8.6	103.6	86.4
24	576	2.7604	.1625	1326.2976	215.52	46.52	± 6.8	101.8	88.2
25	625	2.7959	.1270	1439.1250	182.77	13.77	± 3.7	98.7	91.3

The equation to which the calculations above give solutions is as follows:

$$.695 = .4343 \left[\frac{(95 - \bar{X})^2}{d^2} + \frac{169}{d^2} \right] - \log 169 + \log \sigma^2$$

or

$$\bar{X} = 95 \pm \sqrt{-169 + 2.3026d^2(2.9229 - \log d^2)}$$

This equation is Eq. (5) with $k = .695$, $N = 11$, $\bar{X} = 95$, and $\sigma = 13$.

Joint Maximum-likelihood Estimates of the Mean and Standard Deviation. A final use of the joint distribution of the mean and standard deviation is to make joint maximum-likelihood estimates of the mean and standard deviation of the population. The equation for this joint distribution is Eq. (1). This equa-

tion may be viewed from two aspects. For a given value for each of the population parameters $\bar{X}$ and σ , it may be viewed as giving the probability of getting various sample values for statistics $\bar{X}$ and σ ; for given sample values of $\bar{X}$ and σ , it may be viewed as giving the probability of getting such a sample result for various hypothetical values of the parameters $\bar{X}$ and σ .

In testing hypotheses and determining confidence limits the first view of the equation has been adopted and illustrated; in determining maximum-likelihood estimates it is the second aspect of the equation that is significant. For, by definition, the maximum-likelihood estimates of the population mean and standard deviation are the values of these parameters that will make the probability of the given sample a maximum. Thus a study is made of how the probability of the sample varies with different hypothetical values for $\bar{X}$ and σ , and the values that make this probability the greatest are the maximum-likelihood estimates. Algebraically the procedure is as follows:

Since the probability of the given sample will be a maximum when its logarithm is a maximum, the first step is to simplify Eq. (1) by taking logarithms. This gives the following result

$$\log P(\bar{X}, \sigma) = -\log \sigma_{\bar{X}} - \frac{(\bar{X} - \bar{X})^2}{2\sigma_{\bar{X}}^2} - \frac{N\sigma^2}{2\sigma^2} - (N-1) \log \sigma \\ + \text{other terms not involving } \bar{X} \text{ or } \sigma \quad (6)$$

or, since $\sigma_{\bar{X}}^2 = \sigma^2/N$,

$$\log P(\bar{X}, \sigma) = \frac{N(\bar{X} - \bar{X})^2 + N\sigma^2}{2\sigma^2} - N \log \sigma \\ + \text{other terms not involving } \bar{X} \text{ or } \sigma$$

The values of $\bar{X}$ and σ that make this a maximum will be those for which the partial derivatives with respect to these parameters are equal to zero. Taking these derivatives and setting them equal to zero give the following results:

$$\left. \begin{aligned} \frac{2N(\bar{X} - \bar{X})}{2\sigma^2} &= 0 \\ -\frac{N}{\sigma} + \frac{N(\bar{X} - \bar{X}) + N\sigma^2}{\sigma^2} &= 0 \end{aligned} \right\} \quad (7)$$

The common solutions of these two equations are $\bar{X} = \bar{X}$ and $\sigma = \sigma$. These, then, are the joint maximum-likelihood estimates of the mean and standard deviation of the population.

CHAPTER XV

SAMPLING FLUCTUATIONS IN REGRESSION STATISTICS

Lines and planes of regression are commonly used devices to estimate one variable from one or more other variables. It is therefore desirable to study sampling fluctuations in regression statistics, higher order variances, and lines and planes of regression in order to determine the accuracy of the estimates made from them. Such a study is the purpose of this chapter.

MAXIMUM-LIKELIHOOD ESTIMATES OF REGRESSION STATISTICS AND HIGHER-ORDER VARIANCES

Two Variables. The problem to be considered here is this: If a random sample has been taken from a given bivariate population, what are the best estimates that may be made of the regression parameters of the population and how will these estimates fluctuate from sample to sample? In what follows it will be assumed that the population is a normal population in which the lines of regression are the loci of mean values and the distributions of cases around these lines are normal distributions.

Consider first the line of regression of X_1 on X_2 , and let the equation of this line in the population be represented by the equation

$$X'_1 = a_{1.2} + b_{12}X_2 \quad (1)$$

The problem is to estimate $a_{1.2}$ and b_{12} and to determine the sampling fluctuations of these estimates. This section will be devoted to the first of these problems. Although the analysis relates to only one of the lines of regression, it is equally applicable to the other.

As in other cases the method of solving this problem will be the method of maximum likelihood. In other words, the estimates of $a_{1.2}$ and b_{12} will be those that will make the logarithm of the probability of the given sample a maximum. The procedure is as follows: First note that any pair of X_1, X_2 values may be represented by a point in a plane such as point P in

Fig. 112. If values are assigned to $a_{1.2}$ and b_{12} , the line of regression represented by Eq. (1) will be a straight line in this plane. The vertical deviation of the point P from the line of regression will be equal to

$$v = X_1 - X'_1 = X_1 - a_{1.2} - b_{12}X_2$$

For each pair of X_1, X_2 values a vertical deviation from the line of regression can thus be calculated. A sample set of X_1, X_2 pairs of values can thus be translated into a sample set of deviations from the line of regression. If the values assigned to $a_{1.2}$ and b_{12} are such as to make the logarithm of the probability of the corresponding set of deviations a maximum, they will also be such as to make the logarithm of the probability of the set of sample X_1, X_2 pairs of values a maximum, and vice versa.

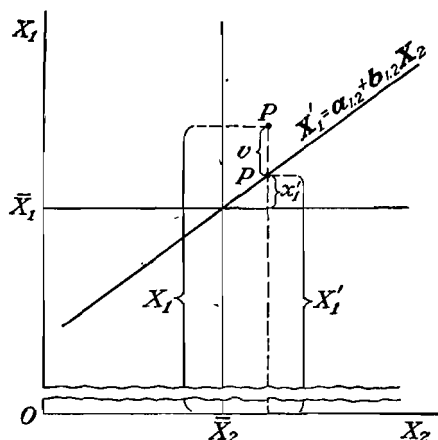


FIG. 112.—Graph of a vertical deviation of a point from the line of regression of X_1 on X_2 .

Since the population is assumed to be a normal bivariate population, it follows that the deviations from the line of regression are normally distributed with a mean of zero and a standard deviation equal to the population first-order standard deviation $\sigma_{1.2}$. The probability of each deviation will thus be of the form

$$dP(v) = \frac{1}{\sigma_{1.2} \sqrt{2\pi}} \exp \left[-\frac{v^2}{2\sigma_{1.2}^2} \right] dv \quad (2)$$

Since the sample is assumed to be a random one, the various deviations will be independent of each other and the probability

of the whole sample of deviations will be the product of their individual probabilities. Inasmuch as exponents are added in multiplication, the probability of the sample set of deviations will be given by the equation

$$dP(v_1, v_2, v_3, \dots, v_N) = \frac{1}{(\sigma_{1.2} \sqrt{2\pi})^N} \exp \left[-\frac{v_1^2 + v_2^2 + v_3^2 + \dots + v_N^2}{2\sigma_{1.2}^2} \right] dv_1 dv_2 dv_3 \dots dv_N$$

which may be written

$$dP(v_1, v_2, v_3, \dots, v_N) = \frac{1}{(\sigma_{1.2} \sqrt{2\pi})^N} \exp \left[-\frac{\Sigma v^2}{2\sigma_{1.2}^2} \right] dv_1 dv_2 \dots dv_N \quad (3)$$

But each v is of the form $X_1 - a_{1.2} - b_{12}X_2$, so that the exponent of e can be written $\Sigma(X_1 - a_{1.2} - b_{12}X_2)^2$. According to the method of maximum likelihood, the values of $a_{1.2}$ and b_{12} are to be so chosen that the logarithm of Eq. (3) is a maximum.

Since the logarithm of $\frac{1}{(\sigma_{1.2} \sqrt{2\pi})^N}$ is independent of $a_{1.2}$ and b_{12} , it may be ignored in determining their maximum-likelihood estimates. The logarithm of the second part of Eq. (3) is merely the exponent of e , since $\log_e e^x$ is x by definition. Hence the logarithm of Eq. (3) will be a maximum when

$$\exp [-\Sigma(X_1 - a_{1.2} - b_{12}X_2)^2]$$

is a minimum, or $\Sigma(X_1 - a_{1.2} - b_{12}X_2)^2$ is a minimum.

Accordingly, the mathematical problem is so to choose $a_{1.2}$ and b_{12} that $\Sigma(X_1 - a_{1.2} - b_{12}X_2)^2$ is a minimum. It will be recognized that this is the criterion of least squares. That is, if a line of regression is fitted to a sample set of data so that the sum of the squares of the deviations from the line is a minimum, then the values of the regression statistics so obtained are the maximum-likelihood estimates of the population regression parameters. But if $\Sigma(X_1 - a_{1.2} - b_{12}X_2)^2$ is to be a minimum, its partial derivatives with respect to $a_{1.2}$ and b_{12} must be zero. These two conditions give the equations:

$$\left. \begin{aligned} \Sigma(X_1 - a_{1.2} - b_{12}X_2) &= 0 \\ \Sigma(X_1 - a_{1.2} - b_{12}X_2)X_2 &= 0 \end{aligned} \right\} \quad (4)$$

If $\check{a}_{1.2}$ and $\check{b}_{12}$ represent the maximum-likelihood estimates of $a_{1.2}$ and b_{12} , respectively, as given by Eq. (4), then

$$\left. \begin{aligned} \check{a}_{1.2} &= \bar{X}_1 - \check{b}_{12}\bar{X}_2 \\ \check{b}_{12} &= \frac{\Sigma X_1 X_2 - N \bar{X}_1 \bar{X}_2}{\Sigma X_2^2 - N \bar{X}_2^2} \end{aligned} \right\} \quad (5)$$

The sample least-squares values of $a_{1.2}$ and b_{12} are thus the maximum-likelihood estimates of the population parameters. Similarly, the sample least-squares values of $a_{2.1}$ and b_{21} are the maximum-likelihood estimates of $a_{2.1}$ and b_{21} .

More Than Two Variables. In like manner, it can be shown that the sample least-squares values of the regression coefficients of planes and hyperplanes of regression are the maximum-likelihood estimates of the corresponding population regression parameters. Thus the maximum-likelihood estimates of $a_{1.23}\dots$, $b_{12.3}\dots$, $b_{13.2}\dots$ are given by the solutions of the least-squares equations.

$$\left. \begin{aligned} \Sigma(X_1 - a_{1.23}\dots - b_{12.3}\dots X_2 - b_{13.2}\dots X_3 - \dots) &= 0 \\ \Sigma(X_1 - a_{1.23}\dots - b_{12.3}\dots X_2 - b_{13.2}\dots X_3 - \dots)X_2 &= 0 \\ \Sigma(X_1 - a_{1.23}\dots - b_{12.3}\dots X_2 - b_{13.2}\dots X_3 - \dots)X_3 &= 0 \\ \dots &\dots \end{aligned} \right\} \quad (6)$$

Similar equations hold for maximum-likelihood estimates of $a_{2.13}\dots$, $b_{21.3}\dots$, $b_{31.2}\dots$, etc.

SAMPLING DISTRIBUTIONS OF REGRESSION STATISTICS

Two Variables. *The Sampling Distributions of $\check{a}_{1.2}$, $\check{b}_{12}$, $\check{a}_{2.1}$, and $\check{b}_{21}$.* The sampling distributions of regression statistics calculated by the method of least squares (*i.e.*, the sampling distributions of the maximum-likelihood estimates of the regression parameters) are similar to the sampling distributions of the mean. If the population is normal, these distributions are normal. The means of the sampling distributions are the population values and the standard deviations are the population higher-order standard deviations $\sigma_{i,jk}\dots$ divided by $\sqrt{N}$ times some function of the independent variable. These general formulas will first be explained and illustrated with reference to two variables.

For a line of regression the maximum-likelihood estimates of $a_{1.2}$ and b_{12} are given by Eq. (5). The sampling distributions

of these estimates ($\check{a}_{1.2}$ and $\check{b}_{12}$) are normal with means equal to the population values of $a_{1.2}$ and b_{12} and standard deviations equal to $\frac{\sigma_{1.2}}{\sqrt{N}}$ and $\frac{\sigma_{1.2}}{\sigma_2 \sqrt{N}}$, respectively. Similarly, the sampling distributions of $\check{a}_{2.1}$ and $\check{b}_{21}$ are normal with means equal to $a_{2.1}$ and b_{21} and standard deviations equal to $\frac{\sigma_{2.1}}{\sqrt{N}}$ and $\frac{\sigma_{2.1}}{\sigma_1 \sqrt{N}}$, respectively.

If the first-order standard deviation of the population is not known, it must be estimated from the sample and the t distribution must be used in place of the normal curve, the appropriate value of n being $N - 2$. If the sample is large, the t distribution is so close to the normal distribution that the latter may be used instead. Hence, it makes little difference in the case of large samples whether or not the population first higher-order variance is known. In general, all the discussion of Chap. XI concerning the sampling fluctuations in the mean are applicable to the sampling fluctuations in $\check{a}_{1.2}$, $\check{b}_{12}$, $\check{a}_{2.1}$, and $\check{b}_{21}$.

Use of Distributions in Testing Hypotheses. To illustrate the testing of a hypothesis regarding a regression coefficient of a normal bivariate population, consider again the data on grades of Mount Holyoke students in first-semester English, X_2 , and second-semester English, X_1 . It was seen in Chap. XIV of Smith and Duncan's *Elementary Statistics and Applications* that the value of $\check{b}_{12}$ is .8322. Since grading in the two courses was apparently on a similar basis, it might be expected that a student whose grade exceeded the mean grade in first-semester English by 20 points, say, would tend to exceed the mean grade in second-semester English by an equal amount, so that the value of b_{12} might be expected to be 1.00. Let the hypothesis, therefore, that $b_{12} = 1$ be tested in the light of the sample result.

Since the population first-order standard deviation is not known, it is necessary to estimate it from the sample. As indicated below,¹ the maximum-likelihood estimate of the population first-order standard deviation is derived from the sample first-order standard deviation by multiplying it by

$\frac{\sqrt{N}}{\sqrt{N-2}}$. Since the sample first-order standard deviation is

¹ See p. 383.

19.53,* the maximum-likelihood estimate of the first-order standard deviation of the population is thus

$$\hat{\sigma}_{1.2} = 19.53 \sqrt{\frac{81}{70}} = 19.78.$$

The estimate of the standard error of b_{12} is consequently

$$\hat{\sigma}_{b_{12}} = \frac{\hat{\sigma}_{1.2}}{\sigma_2 \sqrt{N}} = \frac{19.78}{(47.29)(9)} = .04647,$$

since $\sigma_2 = 47.29$ and $N = 81$.

The difference between the sample value, .8322, and the hypothetical value, 1.00, is 0.168, which is more than three times the estimated standard error. Since the sample is large, the normal curve may be used to test the hypothesis. The lower .05 point on a normal curve comes at 1.645σ . Hence the sample value obviously falls far below the .05 point, and therefore the hypothesis must be rejected. The value of b_{12} is almost certainly not 1.00.

Use of Sampling Distribution in Determining Confidence Limits. Confidence limits can be obtained for b_{12} as follows: Suppose first that the confidence coefficient is set at .95, in other words, that the confidence limits are to be so chosen that the chances of their including the population value are 95 out of 100. Further, let the confidence interval be such that the chance of failing to include the population value because the interval is set too high is just equal to the chance of failing to include the population value because the interval is set too low. In short, let the desired confidence interval be an unbiased interval with a confidence coefficient of .95.

For the Mount Holyoke data the value of $\hat{b}_{12}$ was .8322, and the maximum-likelihood estimate of the standard error of $\hat{b}_{12}$ has just been seen to equal .04647. Since the sample was large ($N = 81$) and the sampling distribution can be taken as normal, an unbiased confidence interval for b_{12} with a confidence coefficient of .95 will be given by $\hat{b}_{12} \pm 1.96\hat{\sigma}_{b_{12}}$, which for the given data yields $.8322 \pm 1.96(.04647) = 0.9233$ and 0.7411 . The desired confidence interval is thus $0.7411-0.9233$. If the sample had been small, the t distribution would perforce have been used. For example, if $N = 20$, then the desired confidence limits would have been $0.8322 \pm 2.10(.04647)$, where 2.10 is the

* See SMITH, J. G., and A. S. DUNCAN, *Elementary Statistics and Applications*, p. 363.

.05 value for t with $n = 20 - 2 = 18$. Had there been but 20 cases in the sample, therefore, the confidence limits would have been 0.7326 and 0.9318.

More than Two Variables. The argument for two variables can be extended without difficulty to samples involving more than two variables. Thus, as already indicated, the sampling distribution of any maximum-likelihood estimate of a regression parameter of a normal population is normally distributed with a mean equal to the population b_{12} and a standard deviation equal to $\sigma_{i,jk \dots n}$ divided by $\sqrt{N}$ times some function of the independent variables. In the case of two variables the function of the independent variable that was used was its standard deviation,

Thus σ_{i1} equaled $\frac{\sigma_{1,2}}{\sigma_2 \sqrt{N}}$. For more than two variables the standard-error formulas are equally simple, the form of their equations being symmetrical, as follows:

$$\sigma_{d_{1,23}} = \frac{\sigma_{1,23}}{\sqrt{N}}$$

$$\sigma_{b_{12,3}} = \frac{\sigma_{1,23}}{\sigma_{2,3} \sqrt{N}}$$

$$\sigma_{b_{13,2}} = \frac{\sigma_{1,23}}{\sigma_{3,2} \sqrt{N}}$$

$$\sigma_{d_{1,234}} = \frac{\sigma_{1,234}}{\sqrt{N}}$$

$$\sigma_{b_{12,34}} = \frac{\sigma_{1,234}}{\sigma_{2,34} \sqrt{N}}$$

$$\sigma_{b_{13,24}} = \frac{\sigma_{1,234}}{\sigma_{3,24} \sqrt{N}}$$

$$\sigma_{b_{14,23}} = \frac{\sigma_{1,234}}{\sigma_{4,23} \sqrt{N}}$$

and, in general

$$\left. \begin{aligned} \sigma_{d_{i,jk \dots n}} &= \frac{\sigma_{i,jk \dots n}}{\sqrt{N}} \\ \sigma_{b_{i1, k \dots n}} &= \frac{\sigma_{i,jk \dots n}}{\sigma_{j,k \dots n} \sqrt{N}} \end{aligned} \right\} \quad (7)$$

The standard deviations $\sigma_{j,k \dots n}$ can be calculated by the equation¹

$$\sigma_{i,jkt \dots pn}^2 = \sigma_i^2(1 - r_{ij}^2)(1 - r_{ik,j}^2) \dots (1 - r_{in,jkt \dots p}^2)$$

¹Cf. Smith and Duncan, *op. cit.*, p. 436.

If the value of $\sigma_{i,jk\dots n}$ is not known, it must be estimated from the sample value $\sigma_{i,jk\dots n}$. When this is done, the t distribution must be used in place of the normal distribution with $n = N - m$ where m equals the number of regression statistics in the regression equation. If the sample is relatively large, however, say $N - m > 30$, the normal curve can be used with sufficient accuracy, even if $\sigma_{i,jk\dots n}$ is estimated from the sample value.

Testing a Hypothesis. In Chap. XVII of *Elementary Statistics and Applications*, it was found that for Mount Holyoke students $b_{12.34} = .7211$ where X_1 represents a student's grade in second-semester English, X_2 her grade in first-semester English, X_3 her verbal scholastic-aptitude test score, and X_4 her grade on the College Board Entrance Examination in English. The value of $\sigma_{1.234}$ was 18.63, and the maximum-likelihood estimate of $\sigma_{1.234}$ is¹ found by the equation

$$\hat{\sigma}_{1.234} = 18.63 \sqrt{\frac{81}{81 - 4}} = 19.55.$$

The maximum-likelihood estimate of the standard error of $\hat{b}_{12.34}$ is $\hat{\sigma}_{\hat{b}_{12.34}} = \frac{\hat{\sigma}_{1.234}}{\sigma_{2.34} \sqrt{N}}$. The value of $\sigma_{2.34}$ may be obtained from the equation²

$$\sigma_{j.kl} = \sigma_j \sqrt{1 - r_{jk}^2} \sqrt{1 - r_{jl.k}^2} \quad (8)$$

In the present instance this gives

$$\begin{aligned} \sigma_{2.34} &= \sigma_2 \sqrt{1 - r_{23}^2} \sqrt{1 - r_{24.3}^2} \\ &= 47.29(.8046)(.9192) = 34.98 \end{aligned}$$

The values of $1 - r_{23}^2$ and $1 - r_{24.3}^2$ are derived from the data for r_{23} and $r_{24.3}$ given in *Elementary Statistics and Applications*.³ Hence the maximum-likelihood estimate of the standard error of $\hat{b}_{12.34}$ is

$$\hat{\sigma}_{\hat{b}_{12.34}} = \frac{19.55}{(34.98) \sqrt{81}} = .0621$$

¹ Cf. p. 376.

² Cf. p. 378.

³ SMITH and DUNCAN, *op. cit.*, pp. 445, 456-458. To obtain values of $1 - r^2$ for given values of r , the sine and cosine tables may be used; for $\sin x = \sqrt{1 - \cos^2 x}$.

So much for the preliminary calculations. Suppose the hypothesis is set up that grades in second-semester English tend to show less variation than those in first-semester English owing to training received in first-semester English. Suppose, further, that it is commonly believed that after allowance is made for differences in verbal aptitudes and in secondary-school training as represented by the College Board Entrance Examination in English, a deviation of 10 points from the mean of first-semester English grades will on the average be accompanied by a deviation of only $7\frac{1}{2}$ points in second-semester English. This is a hypothesis that the value of $b_{12.34} = .7500$. The value of $b_{12.34}$ calculated from the sample is .7211. How does the hypothesis fare in the light of this sample result?

To test the given hypothesis it is necessary merely to compare the deviation of the value of $\hat{b}_{12.34}$ from the hypothetical value for $b_{12.34}$ with the standard deviation of $b_{12.34}$. For the given data this yields the following result:

$$\frac{.7211 - .7500}{.0621} = \frac{-.0289}{.0621} = -.464$$

Since the sample is large and the population may be taken as normal, the sampling distribution of $\hat{b}_{12.34}$ may also be taken as normal. Let the coefficient of risk be taken as .05 and let the region of rejection be taken as values of $x/\delta = -1.645$. The sample value of $x/\delta = -.464$ obviously does not fall in the region of rejection, and the hypothesis of $b_{12.34} = .7500$ is therefore not rejected. If the sample had been small, then the t distribution, with $n = N - 4$, would have been used.

Confidence Limits. If the confidence coefficient is set at .95, unbiased confidence limits for the $b_{12.34}$ of the Mount Holyoke data are given by $\hat{b}_{12.34} \pm 1.96\delta_{\hat{b}_{12.34}}$, which for the given data yields $.7211 \pm 1.96(.0621) = .843$ and

$$.7211 - 1.96(.0621) = .599.$$

Hence the chances are .95 that the range from .599 to .843 covers the population value of $b_{12.34}$. For a small sample the .05 value of t for $n = N - 4$ would have had to be used in place of 1.96.

SAMPLING FLUCTUATIONS IN LINE OR PLANE OF REGRESSION

Lines of Regression.¹ For any given value of X_2 the sample regression value of X'_1 is a function of $\tilde{a}_{1.2}$ and $\tilde{b}_{12}$ given by the sample regression equation $X'_1 = \tilde{a}_{1.2} + \tilde{b}_{12}X_2$. Hence sampling fluctuations² in $\tilde{a}_{1.2}$ and $\tilde{b}_{12}$ will cause sampling fluctuations in X'_1 . Let X_2 be measured from its mean value so that the regression equation can be written $X'_1 = \tilde{a}_{1.2} + \tilde{b}_{12}x_2$. In this case, $\tilde{a}_{1.2} = \bar{X}_1$, but to avoid a shift in notation it will continue to be called $\tilde{a}_{1.2}$. Under these conditions, sampling fluctuations in X'_1 for an arbitrarily selected value³ of x_2 will be normally distributed with a mean equal to the population value of X'_1 for the selected value of x_2 , that is, $X'_1 = a_{1.2} + b_{12}x_2$, and a variance equal to the variance of $\tilde{a}_{1.2}$ plus the variance of $\tilde{b}_{12}$ multiplied by x_2^2 . These relationships may be written succinctly as follows:⁴

$$\left. \begin{aligned} \bar{X}'_1 &= a_{1.2} + b_{12}x_2 \\ \sigma_{X'_1} &= \sqrt{\sigma_{\tilde{a}_{1.2}}^2 + \sigma_{\tilde{b}_{12}}^2 x_2^2} \end{aligned} \right\} \quad (9)$$

where $\bar{X}'_1$ refers to the mean of sample values of X'_1 for the selected x_2 and $\sigma_{X'_1}$ refers to the standard error.

With the help of these formulas the sampling fluctuations in X'_1 can be analyzed in the same manner as the sampling fluctuations in a mean or a regression coefficient were analyzed. For example, a particular hypothesis regarding a certain value of X'_1 can be tested by taking the ratio of the difference between the hypothetical value and the sample value to the estimated stand-

¹ In this section and the rest of this chapter, X_1 is taken as the dependent variable. The argument is valid for any dependent variable, however, and formulas in which X_2 or X_3 are taken as the dependent variable may be derived by interchanging the subscript 2 and 1 or 3 and 1.

² The assumption underlying the argument of this section is that the values of X_2 are the same from sample to sample but the values of X_1 vary at random. The sampling is therefore of those values of X_1 associated with given values of X_2 .

³ See footnote 2, page 380.

⁴ It will be noted that the "variables" in this case are $\tilde{a}_{1.2}$ and $\tilde{b}_{12}x_2$ (the selected value of x_2 being a constant, a sort of "coefficient" for $\tilde{b}_{12}$). Equations (9) are thus merely applications of the theorem that the mean of the sum of two normally distributed variables is the sum of their means and the variance of their sum is the sum of their variances if the variables are independent (*cf.* pp. 419-421). It may be shown that $\tilde{a}_{1.2}$ and $\tilde{b}_{12}x_2$ are independent in their sampling fluctuations.

and error of this difference. If the sample is small and the standard error has to be estimated from the data, this ratio is looked up in a t table with $n = N - 2$. If the sample is large or if the standard error is based upon population values, then the normal curve may be used.

Or, again, unbiased confidence limits for X'_1 can be determined by laying off $\pm t_{.05}\sigma_{X'_1}$ from the sample X'_1 , if the sample is small, or $\pm 1.96\sigma_{X'_1}$, if the sample is large. The symbol $t_{.05}$ refers to the .05 point¹ in a t table for the given value of n . The analysis is essentially the same as before and will not be repeated here.

One aspect of the sampling fluctuations in X'_1 , however, deserves further consideration. Since the standard error of X'_1 is a function of the values of the independent variables, confidence limits for X'_1 will vary with its location on the line or plane of regression. Furthermore, it is to be noted from Eqs. (9) that, the closer x_2 is to its mean, the closer the confidence intervals are to the sample values of X'_1 . The loci of the confidence limits for all points on the line of regression yield confidence limits for the line itself. If the confidence coefficient is .95, it may be said that the chances are .95 that these limiting loci include the population line of regression.

If the sample is small and the standard errors have to be estimated from the data, the equations of the limiting loci with a .95 confidence coefficient are

$$X'_1 = \bar{a}_{1.2} + \bar{b}_{12}x_2 \pm t_{.05} \sqrt{\bar{\sigma}^2_{\bar{a}_{1.2}} + \bar{\sigma}^2_{\bar{b}_{12}}x_2^2} \quad (10)$$

where $t_{.05}$ is the .05 point of a t table for $n = N - 2$. If the sample is large, 1.96 can be used in place of $t_{.05}$. A graph of the limiting loci for the line of regression of X_1 on X_2 for the Mount Holyoke data is shown in Fig. 113, and the numerical values for the graph are given in Table 44. These were computed from the equations

$$X'_1 = 217.3 + 0.8322x_2 \pm 1.96 \sqrt{4.83 + .002159x_2^2}$$

which are the equations for the limiting loci. Here $4.83 = \bar{\sigma}^2_{\bar{a}_{1.2}}$ and is calculated from the equation

$$\bar{\sigma}^2_{\bar{a}_{1.2}} = \frac{\bar{\sigma}^2_{1.2}}{N},$$

¹ This is the point on each tail that gives an individual tail area of .025 and a combined tail area of .05 (see Table VII in the Appendix).

where $\sigma_{1.2}^2$ is the maximum-likelihood estimate of the population first-order variance and is equal to the sample first-order variance multiplied by $\frac{N}{N-2}$. For the given data,

$$\sigma_{1.2}^2 = \frac{(81)(19.53)^2}{79} = 391.07$$

and therefore $\sigma_{\hat{\alpha}_{1.2}}^2 = 391.07/81 = 4.83$. The quantity 0.002159 is $\sigma_{\hat{\alpha}_{1.2}}^2$ and is equal¹ to the square of .04647, which was found

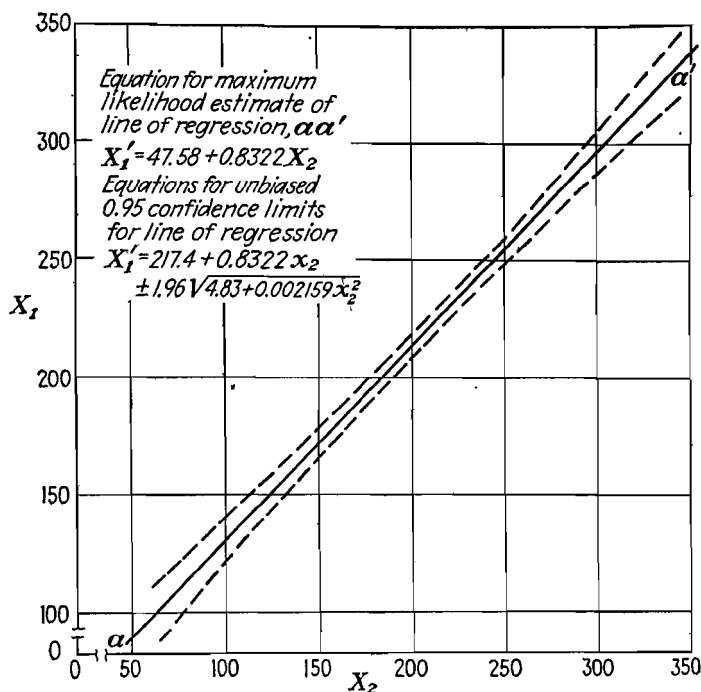


FIG. 113.—Maximum likelihood estimate and confidence limits for line of regression. The chances are .95 that these limiting loci include the population line of regression.

above to be the standard error of $\hat{b}_{12}$. The numerical values from which Fig. 113 was constructed are given in Table 44. It may be noted again that the limiting loci are closest to the sample regression line at the mean of X_2 (i.e., where $x_2 = 0$)

¹ Actually, the value 0.002159 was calculated from the squares of the components and differs slightly from the square of 0.04647 itself.

and tend to swing away from this line as X_2 departs further and further from its mean value.

TABLE 44.—SOLUTIONS FOR THE EQUATION SHOWING CONFIDENCE LIMITS FOR A LINE OF REGRESSION

Equation of confidence limits:

$$X'_1 = 217.4 + 0.8322x_2 \pm 1.96 \sqrt{4.83 + 0.002159x_2^2}$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		(9)
x_2	x_2^2	$0.8322x_2$	$0.002159x_2^2$	$(4) + 4.83$	$\sqrt{(5)}$	$1.96(6)$	$217.4 \pm (3)$ $\pm (7)$		X_1^*
-140	19,600	-116.5	42.32	47.15	6.866	13.4	114.3	87.5	64
-120	14,400	-99.9	31.09	35.92	5.993	11.7	129.2	105.8	84
-100	10,000	-83.2	21.59	26.42	5.140	10.1	144.3	124.1	104
-80	6,400	-66.6	13.82	18.65	4.319	8.5	159.3	142.3	124
-60	3,600	-49.9	7.77	12.60	3.550	7.0	174.5	160.5	144
-40	1,600	-33.3	3.45	8.28	2.878	5.6	189.7	178.5	164
-20	400	-16.6	0.86	5.69	2.385	4.7	205.5	196.1	184
0	0	0	0	4.83	2.198	4.3	221.7	213.1	204
20	400	16.6	.86	5.69	2.385	4.7	238.7	229.3	224
40	1,600	33.3	3.45	8.28	2.878	5.6	256.3	245.1	244
60	3,600	49.9	7.77	12.60	3.550	7.0	274.3	260.3	264
80	6,400	66.6	13.82	18.65	4.319	8.5	292.5	275.5	284
100	10,000	83.2	21.59	26.42	5.140	10.1	310.7	290.5	304
120	14,400	99.9	31.09	35.92	5.993	11.7	329.0	305.6	324
140	19,600	116.5	42.32	47.15	6.866	13.4	347.3	320.5	344

* $X_2 = x_2 + \bar{X}_2$ and $\bar{X}_2 = 204$. Cf. SMITH, J. G., and A. J. DUNCAN, *Elementary Statistics and Applications*, p. 360. The line of regression plotted in Fig. 113 is based upon the equation in *ibid.*, p. 362.

Planes of Regression. The argument concerning sampling fluctuations of a line of regression can be extended easily to a plane of regression. Thus, if samples are selected so that the values of the independent variables $X_2, X_3, \dots$ are the same for each set of samples¹ and only the values of X_1 are allowed

¹ Suppose, for example, that $N = 3$ and the first set of sample values is

Dependent Variable	Independent Variables		
X_1	X_2	X_3	X_4
10	2	3	5
12	3	1	6
14	4	0	7

Then all other samples of 3 would have to be selected so that X_2, X_3 , and

to vary at random from sample to sample, then the mean value of X'_1 for an arbitrarily selected set of values of $X_2, X_3, \dots$ will be

$$\bar{X}'_1 = a_{1.23} \dots + b_{1.23} \dots x_2 + b_{13.2} \dots x_3 + \dots \quad (11)$$

where $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$, $\dots$ and $\bar{X}'_1$ refers to the mean of the sample values of X'_1 for the selected values of $X_2, X_3, \dots$. The standard error of these sample values of X'_1 would be

$$\sigma_{X'_1} = \sigma^2_{a_{1.23} \dots} + \sigma^2_{b_{12.3} \dots} x_2^2 + \sigma^2_{b_{13.2} \dots} x_3^2 + \dots \quad (12)$$

The distribution of X'_1 will be normal, and this may be used to test hypotheses and determine confidence limits if $\sigma^2_{1.23} \dots$ is known or if it is estimated and the sample is large. If the sample is small, the t table must be used with $n = N - m$ where m is the number of regression statistics in the regression equation.

The limiting loci for a plane of regression are given by the sample plane plus the .05 value of t for $n = N - m$ times the standard error of X'_1 . The equations are

$$X'_1 = \check{a}_{1.23} \dots + \check{b}_{12.3} \dots x_2 + \check{b}_{13.2} \dots x_3 + \dots \pm t_{.05} \sqrt{\check{\sigma}^2_{a_{1.23} \dots} + \check{\sigma}^2_{b_{12.3} \dots} x_2^2 + \check{\sigma}^2_{b_{13.2} \dots} x_3^2 + \dots} \quad (13)$$

where $t_{.05}$ is the .05 point of a t table for $n = N - m$. For large samples (that is, $N - m > 30$), ± 1.96 may be substituted for $t_{.05}$.

SAMPLING DISTRIBUTION OF HIGHER-ORDER VARIANCES

The sampling distribution of a higher-order variance is essentially the same as that of a zero-order variance, the form of the distribution being that of the χ^2 distribution. More precisely,

it may be said that the statistic $\frac{N\sigma^2_{1.23} \dots}{\sigma^2_{1.23} \dots}$ will vary from sample

to sample in the manner of the χ^2 distribution whose n is equal to $N - m$, m being the number of regression statistics in the regression equation. Thus the only difference between the sampling distribution of a zero-order variance and that of a

X_4 had the values 2, 3, 5; 3, 1, 6; and 4, 0, 7, respectively, in each set of sample values. Only the values of the dependent variable X_1 would be allowed to vary at random from one sample set to another.

higher-order variance is in the value of n in the χ^2 equation. In the first case, $\frac{N\sigma^2}{\sigma^2}$ has a χ^2 distribution with $n = N - 1$, and in the second case $\frac{N\sigma_{1,23\dots}^2}{\sigma_{1,23\dots}^2}$ has a χ^2 distribution with $n = N - m$.

All the sampling analysis regarding zero-order variance may be applied directly to higher-order variances. For example, if the sample variance $\sigma_{2,1}^2$ is $(21.02)^2 = 441.84$, and if there are 10 cases in the sample, the hypothesis that the population variance $\sigma_{2,1}^2$ is 500, say, could be tested by looking up the value of $\frac{N\sigma_{2,1}^2}{\sigma_{2,1}^2}$ in a χ^2 table with $n = N - 2$. For the given figures,

this would give the result $\frac{10(441.84)}{500} = 8.83$. On the assumption that a coefficient of risk of .05 is adopted and that the region of rejection is taken all at the lower end of the distribution, it is found that for $n = 10 - 2 = 8$ the .95 point of the χ^2 distribution is 2.733. Since the sample value is larger than this, the hypothesis would have to be accepted.

If the sample is larger than 30, say, the normal curve may be used in place of the χ^2 distribution. In this case, the value of $\frac{\sigma^2 - \sigma^2}{\sigma^2 \sqrt{2/N}}$ would be determined and the result looked up in a normal table. For example, for the Mount Holyoke data the size of the sample is 81 and $\sigma_{2,1}^2 = 441.84$. To test the hypothesis that the population first-order variance is 490, say, calculate $\frac{\sigma_{2,1}^2 - \sigma_{2,1}^2}{\sigma_{2,1}^2 \sqrt{2/N}}$. For the given data this has the value

$$\frac{441.84 - 490}{490 \sqrt{\frac{2}{81}}} = -.63.$$

Since this is numerically less than -1.645 , the hypothesis must again be accepted.

Confidence limits with a confidence coefficient of .96 can also be established in the same manner as before. For small samples, set $\frac{N\sigma^2}{\sigma^2}$ equal first to the upper and then to the lower .02 points of the χ^2 distribution for which $n = N - m$, and solve for σ^2 . This will give unbiased confidence limits. For example, if

$N = 10$ and $\sigma_{2,1}^2 = 441.84$, the upper and lower limits are given¹ by $\sigma_{1,2}^2 = \frac{10(441.84)}{2.032} = 2,175$ and $\sigma_{1,2}^2 = \frac{10(441.84)}{18.168} = 243$. Limits with only an upper or lower bound can also be derived in a manner similar to that described for the zero-order case.² The work will not be duplicated here. For large samples, unbiased confidence limits are given by $\frac{\sigma^2 - \sigma^2}{\sigma^2 \sqrt{2/N}} = \pm 1.96$. For

$$\sigma_{1,2}^2 = 441.84$$

and $N = 81$ as in the Mount Holyoke problem, these become $\frac{441.84 - \sigma_{1,2}^2}{\sigma_{1,2}^2 \sqrt{\frac{2}{81}}} = \pm 1.96$, which gives $\sigma_{1,2}^2 = 639$ and $\sigma_{1,2}^2 = 338$.

Since the sampling distribution of a higher-order variance is the same as that of a zero-order variance except that in its χ^2 form $n = N - m$ instead of $N - 1$, it follows that the maximum-likelihood estimate of the corresponding population higher-order variance is equal to $\frac{N}{N - m}$ times the sample variance instead of $\frac{N}{N - 1}$. The proof of this is the same as that given in Chap. XI (pages 290 to 294). In the present instance, $\sigma_{1,2}^2$ for the sample equals 441.84. Hence the maximum-likelihood estimate of the value of $\sigma_{1,2}^2$ is $\sigma_{1,2}^2 = \frac{81}{79}(441.84) = 453.03$.

USE OF LINE OF REGRESSION AND HIGHER-ORDER STANDARD DEVIATION IN ESTIMATING THE DEPENDENT VARIABLE

The principal use of a line or plane of regression and the corresponding higher-order standard deviation is to estimate the dependent variable from the independent variables. For example, the sample data on Mount Holyoke grades show that a student's grade in second-semester English (X_1) is on the average related to her first-semester English grade (X_2) by the equation (line of regression of X_1 on X_2)

$$X_1' = 217.4 + .8322(X_2 - \bar{X}_2)$$

This equation may therefore be used to forecast a student's

¹ Note that $n = 10 - 2 = 8$.

² See p. 289.

second-semester grade from her first-semester grade. For example, if a student gets a first-semester grade of 180, the best estimate that can be made of her second-semester grade will be

$$\begin{aligned} X' &= 217.4 + .8322(180 - 204.1) \\ &= 197.4 \end{aligned}$$

But how reliable is this estimate? It is in answering this question that use is made of the first-order standard deviation around the foregoing line of regression. This first-order standard deviation was found to be $\sigma_{1.2} = 19.53$. If the foregoing equation represented the population line of regression and the foregoing standard deviation was the population first-order standard deviation, and if the data were normally distributed, then it could be said that the chances are 95 out of 100 that the actual second-semester grade will fall within the limits $197.4 \pm 1.96\sigma_{1.2}$, that is, between $197.4 + (1.96)(19.53) = 235.7$ and

$$197.4 - (1.96)(19.53) = 159.3.$$

The foregoing line of regression and standard deviation are not those of the population, however, but were obtained from a previous sample of 81 grades. Hence additional allowance must be made for the sampling fluctuation in the line of regression and in the first-order standard deviation. The procedure may be outlined as follows:

It should be noted that the difference between the estimate of the second-semester English grade and the actual value that occurs will consist of the sum of two independent deviations. First, there is the deviation of the sample regression line from the population regression line. This is the error that arises from taking the regression of the 81 sample cases as the population regression. Second, there is the deviation of the actual second-semester grade from the population regression value. This is the error that arises from taking the mean of a distribution as representative of any individual case. The latter will, of course, differ from the mean in accordance with the laws of sampling. There are thus two sampling errors involved in estimating a future grade, one relating to the past sample of 81 grades, the other to the sampling of the future grade. The total error is the sum of these two. Symbolically

$$X_{1\text{actual}} - X'_{1\text{estimated}} = (X'_{1\text{actual}} - X'_{1\text{estimated}}) + (X_{1\text{actual}} - X'_{1\text{actual}}) \quad (14)$$

Previous analysis indicated that sample regression values X'_1 tend to be normally distributed around the population X'_1 as a mean with a standard deviation equal to

$$\sigma_{X'_1} = \sqrt{\sigma^2_{d_{1.2}} + \sigma^2_{b_{12}}x_2^2}$$

Also, on the assumption that the grades form a normal bivariate population, it can be said that the actual second-semester grades tend to be normally distributed about the population regression line with a standard deviation equal to the (population) first-order standard deviation. Since sample grades in the future are independent of sample grades in the past, with respect to deviations from the population regression line, the sampling variance of the first of the right-hand members of Eq. (14) is independent of the sampling variance of the second right-hand member. Hence the sampling variance of the left-hand member is the sum of these two independent sampling variances. Thus

$$\begin{aligned} \sigma^2_{X_{1\text{actual}} - X'_{1\text{estimated}}} &= \sigma^2_{X'_1} + \sigma^2_{1.2} \\ \sigma_{X_{1\text{actual}} - X'_{1\text{estimated}}} &= \sqrt{\sigma^2_{X'_1} + \sigma^2_{1.2}} \end{aligned} \quad (15)$$

If the population first-order variance $\sigma^2_{1.2}$ is not known and must be estimated from the sample value, then the ratio of $X_{1\text{actual}} - X'_{1\text{estimated}}$ to its estimated standard error will have a sampling distribution that is of the form of the t distribution, with $n = N - 2$. Hence to determine "confidence limits"¹ for the actual value of the second-semester English grade, set

$$X_{1\text{actual}} = X_{1\text{estimated}} \pm t_{.05} \sqrt{\sigma^2_{d_{1.2}} + \sigma^2_{b_{12}}x_2^2 + \sigma^2_{1.2}}$$

where $t_{.05}$ is the .05 value of a t table for $n = N - 2$. For large samples, $t_{.05}$ can be replaced by 1.96, and the normal curve can be used.

As already noted, $\sigma^2_{d_{1.2}} = 4.83$ and $\sigma^2_{b_{12}} = .002159$. Furthermore, $x_2^2 = (180 - 204.1)^2 = 580.81$. Thus the confidence limits

¹ Confidence limits in the sense that in repeated sampling these limits will include the actual value 95 per cent of the time. They are not confidence limits in the usual use of the words since they do not refer to a population parameter.

for the actual value of the second-semester English grade are

$$\begin{aligned} X_{1\text{upper}} &= 197.4 + 1.96 \sqrt{4.83 + (.002159)(580.81) + 391.07} \\ &= 197.4 + 1.96(19.93) = 236.5 \end{aligned}$$

and

$$X_{1\text{lower}} = 197.4 - 1.96(19.93) = 158.3$$

The chances are 95 out of 100 that the range 158.3–236.5 will include the actual second-semester grade for the student whose grade is being predicted.

Since N is large, these limits for the actual value of X_1 do not differ materially from those previously obtained on the assumption that the sample regression and first-order variance are the population regression and first-order variance.¹ For small samples, however, the allowance for sampling errors in the regression line and the first-order variance will make a much greater difference.

Similar analyses will determine confidence limits for forecasts from a plane of regression. For small samples, the limits are given by

$$X_1 = X'_1 \pm t_{.05} \sqrt{\sigma^2_{a_{1.23}\dots} + \sigma^2_{b_{1.23}\dots} x_2^2 + \sigma^2_{b_{12.3}\dots} x_3^2 + \dots + \sigma^2_{1.23}\dots} \quad (16)$$

where $t_{.05}$ is the .05 point of a t table for $n = N - m$. If the samples are large, then 1.96 may be used in place of $t_{.05}$.

¹ See p. 388.

CHAPTER XVI

PROBLEMS INVOLVING TWO SAMPLES

Some of the more interesting problems in statistical analysis involve two samples. A sample poll, for example, is taken 1 month before election and another 2 days before election. The former shows a Democratic vote of 54 per cent and a Republican vote of 46 per cent; the latter a Democratic vote of 49 per cent and a Republican vote of 51 per cent. If the samples both number 100, can the difference in the two percentages be taken to represent a real change in sentiment or can such a difference be reasonably attributed to the chance fluctuations of sampling? Again, 200 automobile tires of make *A* show an average mileage of 16,400 miles; 300 tires of make *B*, subjected to the same test, show an average mileage of 15,900 miles. Does the difference in mileage indicate that make *A* is really better than make *B*, or can the difference be reasonably attributed to chance? It is with such problems as these that the present chapter will be concerned.

OUTLINE OF THE GENERAL ARGUMENT

Proceeds from a Null Hypothesis. The statistical analysis by which it is determined whether the difference between two samples can reasonably be attributed to chance starts with a null hypothesis. The hypothesis is first set up that the populations from which the two samples are taken do not differ with respect to the characteristic in question. This is called the "null hypothesis." Second, some statistic is computed from the two samples that is based upon the difference in characteristics being studied. This may be the difference between their mean values, the ratio of their two variances, or the like. Third, the sampling distribution of this statistic is derived on the basis of the null hypothesis. That is, the set of all possible pairs of samples from the assumed populations is derived, and the percentages of these pairs of samples having various values of the selected statistic

are determined. Fourth, a special subset of the set of all possible pairs of samples is marked as an appropriate region of rejection. For example, this might be the upper 5 per cent of a normal curve. Finally, it is noted whether the statistic for the given pair of samples falls in the chosen region of rejection. If it does, the null hypothesis is rejected and the difference between the two samples is not attributed to chance. This general procedure will be explained more fully below with reference to several selected problems.

Sampling Distribution of the Difference between Two Independent Sample Statistics. Before turning to particular problems, however, two general relationships are worth noting. Often the statistic that is taken to measure the difference between two samples is the arithmetic difference between two individual sample statistics. Thus the statistic may be $\bar{X}_1 - \bar{X}_2$, $\sigma_1 - \sigma_2$, or the like.

In general, let such a statistic be designated as θ , and let the two individual sample statistics be designated as θ_1 and θ_2 , so that $\theta = \theta_1 - \theta_2$. Now it is shown in the Appendix to this chapter that if the two samples are independent of each other, whatever the nature of the populations from which the two samples have been drawn, the mean of the sampling distribution of θ is equal to the mean of the sampling distribution of θ_1 , minus the mean of the sampling distribution of θ_2 , and the variance of the sampling distribution of θ is equal to the variance of the sampling distribution of θ_1 plus the variance of the sampling distribution of θ_2 . In summary, if the two samples are independent,

$$\left. \begin{aligned} \bar{\theta} &= \bar{\theta}_1 - \bar{\theta}_2 \\ \sigma^2_{\theta} &= \sigma^2_{\theta_1} + \sigma^2_{\theta_2} \end{aligned} \right\} \quad (1)$$

These general relationships are worth remembering.

DIFFERENCE BETWEEN TWO SAMPLE PERCENTAGES

A problem that often arises is whether the percentage of favorable cases in one sample is significantly different from the percentage of favorable cases in another sample. Consider, for example, a recent senatorial campaign in Kentucky in which the two candidates for the Democratic nomination were Barkley and Chandler. On Apr. 10 a sample poll was taken which gave Barkley 67 per cent of the total vote and Chandler 33 per

cent. On July 8 another sample poll was taken which gave Barkley 64 per cent of the vote and Chandler 36 per cent.¹

If the first sample numbered 100 votes and the second 200 votes, was it to be inferred that, during the 3 months from April to July, public sentiment shifted from Barkley to Chandler, or could the difference between the two returns have been reasonably attributed to chance? This is the question that the following argument will seek to answer.

The Argument. *The Null Hypothesis.* In accordance with the general procedure outlined above, the first step in the analysis is to set up a null hypothesis. In the present instance this would state that the sentiment of the voters as a whole was the same on July 8 as it was on Apr. 10 and that the difference in sample returns for these two dates was merely the chance result of random sampling. The purpose of the following analysis is to see whether this hypothesis should be accepted or rejected.

The Statistic. The second step in the analysis is to choose an appropriate statistic for measuring the difference between the two sample returns. The statistic usually selected is the difference between the two sample percentages. Call this statistic p_d , and let it be defined by $p_d = p'_1 - p''_1$, where p'_1 is the sample percentage in favor of Barkley on Apr. 10 and p''_1 is the sample percentage in favor of Barkley on July 8.

The Sampling Distribution of p_d . It was shown in Chap. IX that the percentage of favorable cases in a sample would vary from sample to sample in accordance with the binomial distribution. The mean of the distribution was found to be equal to Np_1 and the variance p_1p_2/N , where p_1 is the percentage of favorable cases in the population and p_2 the percentage of unfavorable cases. It was also pointed out that, if the sample was large, the binomial distribution could be approximated by a normal curve whose mean and variance were the same as those of the binomial distribution.

The same sort of result can be demonstrated in the present instance.² Thus it can be shown that the difference between the percentages of two large samples independently derived from the same population (the null hypothesis) is distributed approxi-

¹ GALLUP, G., and S. F. RAE, "Is There a Bandwagon Vote?" *The Public Opinion Quarterly*, Vol. 4, (1940), p. 245.

² The mathematical analysis is beyond the scope of this book.

mately in the form of a normal frequency distribution with a mean of zero and a variance equal to the sum of the variances of the two individual sample percentages. That is, p_d can be taken to be normally distributed with a mean of zero and a variance equal to $\frac{p_1 p_2}{N'} + \frac{p_1 p_2}{N''}$, where p_1 and p_2 are the percentages of favorable and unfavorable cases in the population from which the two samples are assumed to have been taken and N' and N'' the number in each of the samples.

In the present instance, p_1 and p_2 are not known and must therefore be estimated from the samples. This can be done by taking p_1 as the weighted average of p'_1 and p''_1 and by taking p_2 as 1 minus the estimated value of p_1 . Thus p_1 and p_2 can be estimated from the equations

$$\check{p}_1 = \frac{N'p'_1 + N''p''_1}{N' + N''} \quad \text{and} \quad \check{p}_2 = 1 - \check{p}_1 \quad (2)$$

For the data given, this gives

$$\check{p}_1 = \frac{100(.67) + 200(.64)}{100 + 200} = .65 \quad \text{and} \quad \check{p}_2 = 1 - \check{p}_1 = .35$$

Hence, for the problem in hand, p_d can be taken as normally distributed with a mean of zero and a variance equal to

$$\sigma_{p_d}^2 = (.65)(.35)\left(\frac{1}{100} + \frac{1}{200}\right) = .003363$$

The Region of Rejection. Let it be assumed that in instances of this kind the polling agency does not wish to reject the null hypothesis when it is true more than 5 times out of 100. This means that the region of rejection should be of such a size that the probability of a sample falling within it should be just equal to .05.

Since a real shift in public sentiment away from Barkley and toward Chandler might ultimately lead to the election of the latter, whereas a shift toward Barkley would merely strengthen his existing lead, it may be presumed that the polling agency would be more concerned about accepting the null hypothesis when the former shift in public sentiment had occurred than when the latter had taken place. On the assumption of such an attitude, it would appear that the upper .05 of the sampling distribution of p_d would be the best region of rejection to adopt.

For the probability of a sample p_d falling in this region would then be greatest if public sentiment had actually shifted toward Chandler. The region of rejection in the present instance will therefore be the upper .05 tail of a normal curve whose mean is zero and whose variance is .003363.

Testing the Null Hypothesis. The final step is to test the given hypothesis by noting whether the given sample falls in the selected region of rejection. Since this region comprises the upper .05 tail of the normal curve, it will include all values of x/σ equal to or greater than 1.645. In the present instance,

$$p_d = 67 \text{ per cent} - 64 \text{ per cent} = 3 \text{ per cent}, \quad \text{and}$$

$$\sigma_{p_d} = \sqrt{.003363} = 5.8 \text{ per cent}$$

Hence,

$$\frac{p_d}{\sigma_{p_d}} = \frac{3 \text{ per cent}}{5.8 \text{ per cent}} = .517$$

The sample obviously does not fall in the region of rejection. The null hypothesis is thus accepted, and the difference in the sample polls is attributed to chance.

In conclusion, it should be noted once again that the foregoing analysis relates only to independent samples. If the same people were questioned on July 8 as those questioned in April, the sample results would not be independent of each other and the above analysis could not be applied.

Alternative Argument. Instead of using the foregoing argument it would have been possible to solve the given problem by a test of independence such as that described in Chap. XIII. The steps in this alternative argument are as follows:

First the results of the two polls are set up in a contingency table in which the votes are classified according to candidates on the one hand and dates of polls on the other. This has been done in Table 45. In this form the problem can be stated as

TABLE 45.—CROSS CLASSIFICATION OF POLL DATA

Date of poll	Votes favoring Barkley	Votes favoring Chandler	Total votes
April.....	67	33	100
July.....	128	72	200
Totals.....	195	105	300

follows: Is the classification according to candidates independent of the classification according to dates, or does the division of votes vary significantly from April to July?

The null hypothesis assumes that the true division of votes between the two candidates is the same in April and July and that the apparent differences are due to the chance effects of sampling. On the basis of this hypothesis, estimates are made of the true division of sentiment by taking a weighted average of the sample results for the 2 months. Thus the percentage in favor of Barkley may be estimated as equal to

$$\frac{100(.67) + 200(.64)}{300} = \frac{195}{300} = .65$$

and the percentage in favor of Chandler may be estimated as equal to $\frac{100(.33) + 200(.36)}{300} = \frac{105}{300} = .35$. The total vote in each month is then distributed in the same proportion as these estimates of the hypothetical division of sentiment. The results are shown in Table 46. The question then is: Do the actual results, shown in Table 45, differ significantly from the expected results shown in Table 46?

TABLE 46.—POLL RESULTS THAT WOULD BE EXPECTED ON THE ASSUMPTION OF INDEPENDENCE

Date of poll	Expected votes favoring Barkley	Expected votes favoring Chandler	Total votes
April.....	65	35	100
July.....	130	70	200
Totals....	195	105	300

As pointed out in Chap. XIII, questions of this kind can be answered by calculation of the statistic

$$\sum_i \frac{(\text{actual number in each cell minus expected number})^2}{\text{expected number}} \quad (3)$$

and making use of the fact that this statistic has a sampling distribution of the form of a χ^2 distribution. The degrees of freedom, it is noted, will be equal to $(r - 1)(c - 1)$, where r is the number of rows in the contingency table and c the number of

columns. Since r and c both equal 2 in the present problem it follows that statistic (3) for this problem will be distributed like χ^2 with $n = (1)(1) = 1$.

It remains now to determine what part of the sampling distribution will serve as the best region of rejection. In the previous argument the upper .05 tail of the normal curve (the sampling distribution there used) was considered as the best region to employ. For it was assumed that the polling agency would be especially anxious not to accept the null hypothesis if the actual trend in public sentiment was away from Barkley and toward Chandler. The same assumption will be made here.

In view of this assumption it might appear at first glance that the upper .05 tail of the sampling distribution would again be the most appropriate region of rejection to employ, but this is not so. Statistic (3) does not take account of signs, and large values might arise from sample differences in favor of Barkley just as readily as from sample differences in favor of Chandler. If the region of rejection in this case is to be comparable with the upper .05 tail of the normal curve used in the previous argument, it should include only those values of the statistic that arise from differences in favor of Chandler. Such differences, on the basis of the null hypothesis, will occur only 50 per cent of the time. Hence, a .05 region of rejection comparable to that of the previous argument may be taken to constitute all values of statistic (3) that arise from differences in favor of Chandler and that at the same time are equal to or greater than the upper .10 point of the χ^2 distribution ($n = 1$), that is, equal to or greater than 2.706.

The problem can now be solved. For the given data the value of statistic (3) is as follows:

$$\frac{(67 - 65)^2}{65} + \frac{(33 - 35)^2}{35} + \frac{(128 - 130)^2}{130} + \frac{(72 - 70)^2}{70} = .264$$

Since this is less than 2.706, the sample does not fall in the selected region of rejection and the null hypothesis is again accepted.¹ As before, the difference between the two samples is attributed to the chance effects of sampling.

¹ It is interesting to compare probabilities in this problem. By the first method $x/\sigma = .517$, and the probability of as great or greater value in either direction (+ or -) is approximately .60. By the second method the value of χ^2 is .264, and the probability of as great or greater value is again roughly

DIFFERENCE BETWEEN TWO INDEPENDENT SAMPLE MEANS

When the Populations Have a Common Known Variance.

The Problem. The foregoing section was concerned with the difference between two samples from a discrete population. Similar problems arise in the case of continuous data. A problem that often occurs is whether the means of two independent samples are significantly different. This problem is attacked in different ways, depending on the conditions involved. In this section it will be assumed that the variances of the populations from which the two samples have been drawn are the same and that this common value is known. The question to be tested will be: Are the means of the populations also the same? Only normal populations will be considered here; nonnormal cases will be discussed in Chap. XVIII.

Some years ago the *New York Sun* published employment data for a large number of industrial companies. Data were given for each company for the years 1929 and 1935. A random sample of 10* of the smaller companies was taken from both these 2 years; the mean of the first was found to be 89.1 men, and the mean of the second 123.5 men. A different set of companies was taken for each year so as to make the samples independent of each other. If the standard deviations of the populations are known in this instance to be both 100 and if the populations are taken to be normal, can it be inferred from these two samples that industrial employment among small companies was in general larger in 1935 than in 1929? This is the problem that the following argument will seek to solve.

The Null Hypothesis. To determine whether the means of the two samples are significantly different, it will first be assumed that they are not different, and then the consequences of this assumption will be compared with the actual results. The null hypothesis will thus be that the two samples are from identical normal populations with standard deviations of 100.

.60. In fact $\sqrt{\chi^2}$ is, for $n = 1$, distributed like double the upper half of a normal curve, and in the given problem $\sqrt{.264}$ equals approximately .514, which is the same as .517, except for errors arising from the use of decimals.

* Usually a larger sample than this would be taken. A small sample is taken here to show that the analysis is applicable to both large and small samples.

The Statistic. The statistic that is generally selected in problems of this sort is the difference between the sample means. If the mean employment of the 10 companies in 1929 is indicated as $\bar{X}_1$ and the mean employment of the 10 companies in 1935 is indicated as $\bar{X}_2$, then the selected statistic is defined as

$$\bar{X}_{1-2} = \bar{X}_1 - \bar{X}_2$$

The Sampling Distribution of $\bar{X}_{1-2}$. It can be shown by mathematical analysis¹ that, when the population from which the two samples are independently drawn are normal populations, then the sampling distribution of $\bar{X}_{1-2}$ is also normal, the mean of the distribution being zero and its variance being equal to the identical variance of the two populations multiplied by $\frac{1}{N_1} + \frac{1}{N_2}$. That is, if all possible pairs of samples of 10 each were selected from the assumed populations and the differences between the means of these samples were computed, it would be found that the differences would form a normal frequency distribution with a mean of zero and a variance equal to $\sigma^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$. Since in the given problem σ is 100 and $N_1 = N_2 = 10$, it follows that $\bar{X}_{1-2}$ has a sampling distribution whose mean is zero and whose variance is $(100)^2 \left(\frac{1}{10} + \frac{1}{10} \right) = 2,000$.

The Region of Rejection. Let the risk of rejecting the null hypothesis when it is true be placed at .05. This means that the total region of rejection should be of this size. Furthermore, there appears to be no special reason in this instance to put all the region of rejection at one end of the sampling distribution. For, in the absence of any special motive in making the statistical investigation, it would appear to be as bad an error to accept the null hypothesis when in fact employment was greater in 1929 than in 1935 as it would be to accept the null hypothesis when in fact employment was greater in 1935 than in 1929. On this basis the total region of rejection will be split up equally so as to include the .025 tails at each end of the sampling distribution. The points marking these two tails, it will be recalled, are $x/\sigma = \pm 1.96$.

¹ The analysis is essentially the same as that showing that the mean of a sample itself is normally distributed (cf. Chap. X).

The Test of the Null Hypothesis. The difference between the sample means is $89.1 - 123.5 = -34.4$. The variance of the sampling distribution of $\bar{X}_{1-2}$ was computed to be 2,000. The standard deviation is the square root of this, or 44.7. The ratio of the difference between the two means to its standard deviation is thus $-34.4/44.7 = -.77$, which is well within the 1.96 point marking the region of rejection. Accordingly, there is no basis for rejecting the null hypothesis, and the difference between the two sample means can presumably be attributed to chance.

When the Populations Have a Common but Unknown Variance. In some problems it may be assumed that the variances of the populations from which the two samples have been taken are identical, but the value of this common variance may not be known. In this instance the population variance must be estimated from the samples. For small samples, this requires some modification in the foregoing analysis.

When two samples are taken from normal populations with identical variances, the maximum-likelihood estimate of this common variance based on the sampling variation in the variance that is independent of the difference between the sample means is¹

$$\hat{\sigma}^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2 - 2} \quad (4)$$

where σ_1^2 and σ_2^2 are the two sample variances and N_1 and N_2 are the number in the samples. Except for the -2 , this is merely a weighted mean of the two sample variances. The reason for the division by $N_1 + N_2 - 2$ instead of $N_1 + N_2$ lies in the fact that each of the sample variances is measured from its own mean instead of the true population mean. Since the sample mean itself varies from sample to sample, this reduces somewhat the average value of the sample variances as compared with the true population variance. To correct for this bias in both the sample variances, the factor $N_1 + N_2 - 2$ is substituted for $N_1 + N_2$. The value of $\hat{\sigma}^2$ is said in this instance to be an estimate of σ^2 based on $N_1 + N_2 - 2$ degrees of freedom.

¹ The probability of getting the two samples can be broken up into two parts, one depending only on the two sample means, the other only on the two sample variances. When the common population variance is chosen so as to maximize this second part of the total probability, its value is found to be that given in the text.

Let the foregoing estimate of the population variance be multiplied by $\frac{1}{N_1} + \frac{1}{N_2}$, and take the ratio of the difference between the two sample means to the square root of this quantity. The resulting ratio, which may be written

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad (5)$$

is similar to the $\frac{\bar{x}}{\sigma}$ that was found in the previous problem to be distributed in accordance with the standard normal curve. The only difference is that the estimated rather than the actual population variance is now used. This difference, however, causes the statistic $\frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{(1/N_1) + (1/N_2)}}$ to have a sampling distribution that is nonnormal for small samples. Actually, the sampling distribution of this statistic is of the form of the t distribution, with n in the t formula, *i.e.*, the degrees of freedom, equal to $N_1 + N_2 - 2$.

Since the t distribution approaches the normal curve when n is large (say greater than 30), the use of the estimated value of the population variance leads to no change in the analysis if the samples are large.¹ When the samples are small, however, *i.e.*, when $N_1 + N_2 - 2$ is less than 30, it is better to use the t distribution in place of the normal distribution for testing the given hypothesis. This is the only fundamental change required in the previous analysis.

To illustrate the estimation of the common population variation, consider once again the data on industrial employment in 1929 and 1935. Let it be assumed that the variance in employment of small industrial companies is the same in both years but

¹ The t distribution is leptokurtic, and its variance equals $\frac{n}{n-2}$. Hence a better approximation can be obtained for values of n between 30 and 100 by multiplying the statistic $(\bar{X}_1 - \bar{X}_2)/\sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$ by $\sqrt{\frac{n-2}{n}}$ before looking up the value in the normal table. It will be noted that here

$$n = N_1 + N_2 - 2.$$

that its value is not known and must therefore be estimated from the samples. Suppose that the variance of the 1929 sample is 9,834.5 and the variance of the 1935 sample is 18,832.1. As stated above, the maximum-likelihood estimate of the common variance is found as follows,

$$\hat{\sigma}^2 = \frac{10(9,834.5) + 10(18,832.1)}{10 + 10 - 2} = 15,926.$$

The square root of this is 126.2, and the statistic

$$\frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

has the value

$$\frac{89.1 - 123.5}{126.2 \sqrt{\frac{1}{10} + \frac{1}{10}}} = .64$$

For $n = 10 + 10 - 2 = 18$, the .025 points of the t distribution are ± 2.101 . Values of t numerically greater than 2.101 will thus constitute a symmetrical .05 region of rejection. It is clear that the sample value of t does not in this instance fall in the region of rejection, and again the null hypothesis is not rejected. As before, the difference between the two sample means is apparently due to chance.

If the samples had numbered 50 cases each instead of 10, the value of

$$\frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

would have been

$$\frac{89.1 - 123.5}{118 \sqrt{\frac{1}{50} + \frac{1}{50}}} = 1.46$$

Since the samples are large, the normal curve can in this case be used instead of the t curve, even though the population variance has been estimated. The region of rejection will therefore constitute values of $x/\hat{\sigma}$ numerically greater than 1.96. The sample value of $x/\hat{\sigma}$ is 1.46, and this again fails to fall in the region of rejection. The null hypothesis continues to be accepted even though the samples are now larger.

When the Populations Do Not Have the Same Variance. If it cannot be assumed that the populations from which the two samples have been drawn have identical variances, only a rough test can be employed to determine whether the populations might have the same means. If the samples are fairly large (say 100 or more), it can be assumed with a fair degree of accuracy that the sampling distribution of the difference between two sample means is a normal distribution with a mean of zero and a variance that is approximately equal to the variance of sample 1 divided by N_1 plus the variance of sample 2 divided by N_2 .

That is, for large samples, $\sigma^2_{\bar{x}_1 - \bar{x}_2}$ can be taken as equal to $\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$, and the ratio

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad (6)$$

can be treated as if it were normally distributed. Except for the different method of estimating $\sigma_{\bar{x}_1 - \bar{x}_2}$, the procedure is essentially the same as that outlined in the foregoing sections, and illustrations will be omitted.

CORRELATED SAMPLES

The Problem. If in the foregoing problem the same companies had been taken in 1935 as were taken in 1929, the two samples would not have been independent and the foregoing analysis could not have been validly applied. Data often occur in this form. A group of students, for example, may be tutored for one examination and not tutored for another. Again, a set of hogs may be fed one diet for one month and the same set of hogs fed another diet for another month. How in these cases is a statistical test to be made of the effect of tutoring on students' grades or the effect of diet on the rate of growth of hogs? It is these questions that the following analysis seeks to answer.

Testing Individual Differences. When two samples relate to the same set of individuals, the simplest method of analysis is to take the difference between the two results for each individual. If each sample numbers N cases, this process will give a set of N individual differences. If the two sets of data are not really different, the whole population of individual differences will have a mean of zero. Sample sets of differences will, of course, have

means that are not zero, but these sample means will tend to be distributed around this population mean of zero.

If σ^2 is the variance of the whole population of differences, the distribution of sample mean differences will have a variance of $\frac{\sigma^2}{N}$. This distribution will be normal in form, so that, if the variance of the population of differences is known, the mean of any particular sample of differences can be tested by taking the ratio of this mean value to $\frac{\sigma}{\sqrt{N}}$ and looking up the result in a normal frequency table.

If the variance of the population of differences is not known, it must be estimated from the given sample. The maximum-likelihood estimate of the population variance that can be made independently of the sample mean is $\hat{\sigma}^2 = \frac{N}{N-1} \sigma_s^2$, where σ_s^2 is the variance of the sample of differences. The ratio of the sample mean difference to $\frac{\hat{\sigma}}{\sqrt{N}}$ gives a statistic the sampling distribution of which is of the form of the t distribution with n , the degrees of freedom, equal to $N - 1$. When the variance of the population of differences is not known, therefore, but must be estimated from the sample, the t distribution is used to test the significance of the sample mean difference. Of course, if N is large, the normal curve can be used as an approximation to the t curve. Illustrations of these procedures follow.

Illustrations. Employment figures for 10 small industrial companies in 1929 and 1935 are given in Table 47. This also shows the individual differences for the 2 years. The mean of these differences is 10.5, and the question is: Does this sample mean difference differ significantly from zero, or can its positive value be reasonably attributed to chance?

First suppose that the standard deviation of the whole population of differences is known to be 50. Then the standard deviation of the distribution of sample means of differences is equal to $50/\sqrt{10} = 15.8$, and the ratio of the given sample mean to this standard error is $10.5/15.8 = .66$. This ratio, as pointed out above, is distributed in accordance with the standard normal curve. If the region of rejection is taken as the values of x/σ that are numerically greater than 1.96 (a symmetrical region

appears to be justified here), then this sample x/δ does not fall in the region of rejection and the null hypothesis is not rejected. That is, the sample mean difference is not significantly different from zero, and the apparent difference in employment may reasonably be attributed to the chance effects of sampling.

TABLE 47.—INDIVIDUAL DIFFERENCE IN EMPLOYMENT OF 10 INDUSTRIAL COMPANIES, 1929 AND 1935

Number employed		Difference in number employed, 1929-1935	Differences squared
1929	1935		
110	48	62	3,844
166	93	73	5,329
130	120	10	100
95	85	10	100
36	52	-16	256
188	240	-52	2,704
85	135	-50	2,500
185	130	55	3,025
201	217	-16	256
144	115	29	841
		Σ +105	Σ 18,955

If the population standard deviation is not known, as is generally the case, then it must be estimated from the sample. In the present instance, the standard deviation of the sample differences is equal to 42.25.* The best estimate, therefore, that can be made of the standard deviation of population differences is

$$\sqrt{\frac{10}{9}} (42.25); \text{ and this gives } \frac{\bar{\sigma}}{\sqrt{N}} = \frac{\sqrt{\frac{10}{9}} (42.25)}{\sqrt{10}} = 14.08, \text{ and}$$

$\frac{\bar{X}_d}{\bar{\sigma}/\sqrt{N}} = \frac{10.5}{14.08} = .75$. This last quantity, as noted above, is distributed like t . Since for $n = 9$ the .025 points of the t dis-

* This is calculated by use of the short formula

$$\sigma^2 = \frac{\Sigma X^2}{N} - \bar{X}^2$$

which gives

$$\sigma^2 = \frac{18,955}{10} - (10.5)^2 = 1,785.25$$

and

$$\sigma = 42.25$$

tribution are ± 2.262 , values of t numerically greater than 2.262 may be taken as a symmetrical region of rejection with the coefficient of risk equal to .05. The sample $t = .75$ and obviously does not fall in this region of rejection. The null hypothesis that the difference between the two samples is due to chance cannot therefore be rejected, and the assumption that there is no real difference in employment in the 2 years continues to be a reasonable one.

If there had been 50 companies instead of 10, the ratio $\frac{\bar{X}}{(\bar{\sigma}/\sqrt{N})}$ could have been treated as if it were normally distributed. For example, if the mean and standard deviation of a sample of 50 equaled 10.5 and 42.25, then $\bar{\sigma}$ would have the value $\sqrt{\frac{5.0}{49}} (42.25)$, and

$$\frac{\bar{\sigma}}{\sqrt{N}} = \frac{\sqrt{\frac{5.0}{49}} (42.25)}{\sqrt{50}} = 6.03. \quad \text{The ratio } \frac{\bar{X}_d}{\bar{\sigma}/\sqrt{N}}$$

would then equal $\frac{10.5}{6.03} = 1.74$. This is less than 1.96, so that, if the two .025 tails of the normal curve are taken as the approximate region of rejection, the null hypothesis would be accepted. It would be concluded once again that employment in 1929, among small firms, was not really greater than employment among small firms in 1935.

It is interesting to note at this point that, when X_1 and X_2 are correlated, the variance of the differences $X_1 - X_2$ is equal to the variance of X_1 plus the variance of X_2 minus twice the product of the standard deviations by the correlation between X_1 and X_2 . Symbolically,

$$\sigma_{X_1-X_2}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 - 2\sigma_{X_1}\sigma_{X_2}r_{X_1X_2} \quad (7)$$

Hence it follows that, in problems of correlated samples, the variance of the differences is less if the correlation between X_1 and X_2 is greater. The practical importance of this conclusion is that differences in central tendencies can be more readily detected if the correlation between individual members of the samples is increased.

DIFFERENCE BETWEEN TWO SAMPLE VARIANCES

Another problem that often arises in statistical analysis is to determine whether two samples have come from populations with

different variances. To refer to a previous example, it might be claimed that a new process for the manufacture of electric-light bulbs will reduce the variability in length of life of the bulbs. To check this claim two lots of bulbs could be produced, one by the old process, the other by the new, and the difference in variability could be subjected to a statistical test. Again, two different brands of razor blades might be tested as to their sharpness, and the variability in the two brands might be compared. It is with such problems that the following sections are concerned. As before, it will always be assumed that the populations are normal; for nonnormal cases the reader is referred to Chap. XVIII.

Testing a Difference in One Direction. The method of testing the difference between two sample variances varies to some extent with the particular problem involved. If the problem is concerned with testing a difference in one direction only, one method is applicable; if it is concerned with a difference in either direction, another method is required. The present section will deal with the testing of a difference in one direction only.

The Specific Problem. To keep the discussion concrete, consider once again the data on industrial employment among small companies. Suppose it is claimed that, owing to the instability introduced by the depression, there was a greater variability in the size of companies in 1935 than in 1929, size of companies being measured by amount of employment. The variance in employment among the ten 1929 companies, it will be recalled, was 9,834.5, and the variance among the ten 1935 companies was 18,832.1, which would seem offhand to substantiate this claim. The immediate statistical problem is to determine whether this apparent difference in variability is great enough to be attributed to some specific causal factor such as the depression or whether such a difference could reasonably be attributed to the chance effects of sampling.

The Null Hypothesis. In carrying out this statistical test the first step is to set up the null hypothesis that the difference in variance is due, not to some specific causal factor, but only to chance. In other words, the hypothesis states that the 1929 and 1935 populations have the same variance. It will be noted that it does not state that the populations are the same, for nothing is said about the means of the populations. The following test is

therefore independent of whether the means of the populations are the same or not. It thus differs from tests of the difference between two means, which are based upon the assumption that the variances are the same.

It should be noted also that the hypothesis assumes that the samples are independent of each other. If the data on employment pertain to the same companies in both years, then the samples would not be independent of each other and the following analysis could not be validly used.

The Statistic. The statistic that is most convenient to use in the present instance is the ratio of the maximum-likelihood estimates of the population variances in the 2 years. The maximum-likelihood estimate of the population variance made independently from the first sample¹ is $\frac{\sigma_1^2 N_1}{N_1 - 1}$, and the maximum-likelihood estimate of the population variance made independently from the second sample is $\frac{\sigma_2^2 N_2}{N_2 - 1}$, where σ_1^2 and σ_2^2 are the two sample variances and N_1 and N_2 are the number of cases in each sample.

The statistic that is used for this problem is the ratio of these two maximum-likelihood estimates, *i.e.*,

$$\frac{\sigma_1^2 N_1 / (N_1 - 1)}{\sigma_2^2 N_2 / (N_2 - 1)} \quad (8)$$

If the two populations have identical variances, these two estimates will be approximately equal and the above statistic will be close to 1. The statistical problem is to determine whether it differs from unity by an amount greater than can reasonably be attributed to chance.

The Sampling Distribution of the Ratio of Two Maximum-likelihood Estimates of Variance. The reason why the ratio of the two estimates of variance is used rather than their arithmetic difference is that the sampling distribution of the former can more readily be determined. Mathematical analysis shows that this sampling distribution is of the form of the *F* distribution with n_1 and n_2 of the *F* equation equal to $N_1 - 1$ and $N_2 - 1$, respectively. As in other cases, n_1 is the degrees of freedom involved in estimating σ_1 and n_2 the degrees of freedom involved in estimating σ_2 .

¹ That is, independently of the mean.

That is, if pairs of samples are drawn independently from populations with identical variances and if the ratio of the maximum-likelihood estimates of variance is calculated for each pair, these ratios will have a frequency distribution that is of the form of the F distribution, with $n_1 = N_1 - 1$ and $n_2 = N_2 - 1$. It might be well for the reader to turn back at this point to Chap. VI and reread the section there on the F distribution. It is also well to note again that this sampling distribution is valid only for ratios that are computed from independent samples.

The Region of Rejection. The present problem is concerned with whether the 1935 variance is greater than the 1929 variance. If the statistic selected is the ratio of the 1935 variance to the 1929 variance, then the best region of rejection to employ is the upper tail of the F distribution. For it has been shown that, if the presumably larger variance is actually the larger, there will be more chance of rejecting the null hypothesis if the upper tail of the F distribution is used than if any other region is adopted.¹ Of course, it would be possible to take the ratio of the presumably smaller variance to the presumably larger one, and in this case the lower tail of the F distribution would be the more appropriate one to employ. This alternative course is not followed, however, for the tables of the F distribution are computed only for the upper tail and, as noted, the distribution is not symmetrical. In the present instance the upper tail of the F distribution will be employed as the region of rejection, and, as usual, the size of this region will be taken as .05.

For the given problem, the maximum-likelihood estimate of the population variance based upon the 1935 sample is $\frac{(10)(18,832.1)}{9}$, and the maximum-likelihood estimate of the population variance based upon the 1929 sample is $\frac{(10)(9,834.5)}{9}$. The ratio of the first to the second is 1.915. In this problem, this is the same as the ratio of the two sample variances themselves, since the samples are of the same size. In another problem in which the samples are of different sizes, this equality would not exist.

¹ Cf. NEYMAN, J., and E. PEARSON, "On the Problem of the Most Efficient Test of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, Series A, Vol. 231 (1933), pp. 289-337.

Testing the Hypothesis. To test the given hypothesis it is necessary merely to note whether the above ratio is greater or less than the .05 point of the F distribution for which $n_1 = 9$ and $n_2 = 9$.* The tables, unfortunately, do not give the .05 point of this particular F distribution, so that its value must be interpolated. For $n_1 = 8$, $n_2 = 9$, the .05 point is $F = 5.467$; for $n_1 = 12$, $n_2 = 9$, the .05 point is 5.111. Therefore, for the F curve for which $n_1 = 9$ and $n_2 = 9$, the .05 point must lie between these two values. Whatever its exact value, it is clear that the sample ratio of 1.915 is well within the area of acceptance, so that the sample value does not fall in the region of rejection, which has been taken to consist of all values of F equal to or greater than the .05 value. The null hypothesis is thus not rejected in the present problem, and the difference in sample variance is to be attributed to chance.

If neither n_1 nor n_2 have values that are given directly in the F table, it would be necessary to interpolate in both directions. For example, if $n_1 = 9$ and $n_2 = 40$ for the given samples, then it would be necessary to find the following .05 values:

For	The .05 point is
$n_1 = 8$ and $n_2 = 30$	3.173
$n_1 = 12$ and $n_2 = 30$	2.843
$n_1 = 8$ and $n_2 = 60$	2.823
$n_1 = 12$ and $n_2 = 60$	2.496

It is obvious that the .05 point for $n_1 = 9$, $n_2 = 40$ lies somewhere between 3.173 and 2.496. If the given sample ratio lies beyond 3.173, it will certainly lie beyond the .05 value for $n_1 = 9$ and $n_2 = 40$. Similarly, if the sample ratio lies inside 2.496, it certainly lies within the .05 point for $n_1 = 9$ and $n_2 = 40$.

In such cases there is no need for exact interpolation; but if the sample lies between 3.173 and 2.496, then the .05 point for $n_1 = 9$ and $n_2 = 40$, may be roughly obtained by straight-line interpola-

* It will be noted that $n_1 = N_1 - 1$ and that N_1 refers to the size of the sample whose variance is put on top of the fraction expressing the ratio, in this case to the 1935 sample. Thus N_1 refers to the sample that has the presumably larger of the two estimates of variance. Likewise, $n_2 = N_2 - 1$, where N_2 refers to the size of the sample whose variance is put in the denominator of the fraction.

tion.¹ First interpolate for $n_2 = 40$. Since the .05 point for $n_1 = 8$, $n_2 = 30$ is 3.173 and the .05 point for $n_1 = 8$, $n_2 = 60$ is 2.843 and since 40 is distant from 30 by $\frac{10}{30}$ of the distance between 30 and 60, the .05 point for $n_1 = 8$, $n_2 = 40$ will be approximately equal to $3.173 - (\frac{10}{30})(3.173 - 2.843) = 3.063$. Likewise, the .05 point for $n_1 = 12$, $n_2 = 40$ will be approximately equal to $2.823 - (\frac{10}{30})(2.823 - 2.496) = 2.714$. Values for $n_1 = 8$, $n_2 = 40$ and $n_1 = 12$, $n_2 = 40$ having been obtained, it remains to interpolate for $n_1 = 9$. The .05 point for $n_1 = 9$, $n_2 = 40$ will thus be approximately equal to

$$3.063 - (\frac{1}{4})(3.063 - 2.714) = 2.976.$$

This is the .05 point desired. As noted above, straight-line interpolation is not perfectly accurate. Consequently, if a sample value falls close to the interpolated value, it is well to forego any definite conclusion. In such an event it is better to judge the result a borderline case.

Testing a Difference in Either Direction. *The Problem.* The foregoing test of difference between two sample variances was based upon the assumption that the investigator was interested in a significant difference in one direction only. It was on this basis, it will be recalled, that the upper tail of the F distribution was selected as the region of rejection. There may be cases, however, in which the investigator is indifferent as to whether any difference that might exist is in one direction or the other. This is the problem that will now be considered.

The Test. It might be thought offhand that, when the investigator is indifferent as to the way in which the two variances might differ, a satisfactory test could be devised by distributing the region of rejection equally between the two tails of the F distribution. This is not true; for such a test is biased in that the probability of accepting the null hypothesis in such an instance may be greater in some cases in which the null hypothesis is not true than when it is true. The test described below avoids this

¹ Better results may possibly be obtained by taking the variable

$$u = \frac{n_1 F}{n_1 F + n_2}$$

and using tables of the incomplete beta function. Cf. RIDER, PAUL R., *Introduction to Modern Statistical Methods* (1939), p. 119.

difficulty. It is unbiased in that the probability of accepting the null hypothesis when it is true is greater than for any instance when it is not true and in that the probability of rejecting the null hypothesis when it is true is less than for any instance when it is not true.

Let two samples be taken from normal populations. Let the variance of one sample be σ_1^2 and the variance of the other sample be σ_2^2 . The investigator, it will be presumed, is interested in testing the null hypothesis that the two population variances are the same, and he is indifferent as to whether any possible difference between them is in one direction or the other. On these assumptions, the best statistic to employ is

$$L = \frac{(n_1 + n_2) \log_e \hat{\sigma}^2 - n_1 \log_e \hat{\sigma}_1^2 - n_2 \log_e \hat{\sigma}_2^2}{1 + \alpha} \quad (9)$$

in which $n_1 = N_1 - 1$, $n_2 = N_2 - 1$, $\hat{\sigma}^2 = \frac{N_1 \sigma_1^2 + N_2 \sigma_2^2}{N_1 + N_2 - 2}$ (that is, $\hat{\sigma}^2$ is the maximum-likelihood estimate of the population variance based upon the two sample variances taken together), $\hat{\sigma}_1^2 = \frac{N_1 \sigma_1^2}{N_1 - 1}$ and $\hat{\sigma}_2^2 = \frac{N_2 \sigma_2^2}{N_2 - 1}$ (the maximum-likelihood estimates of the population variance based upon the two sample variances separately), and

$$\alpha = \frac{1}{3} \left(\frac{1}{N_1 - 1} + \frac{1}{N_2 - 1} - \frac{1}{N_1 + N_2 - 2} \right).$$

The statistic shown in Eq. (9) has been found to have a sampling distribution that is approximately of the form of the χ^2 distribution with the degrees of freedom n equal to 1. It leads to an unbiased test if the upper tail of the distribution is taken as the region of rejection.¹

An Example. To illustrate this test, consider again the variance in employment among small industrial companies in 1929 and 1935. The variance of a sample of 10 companies from the first year, it will be recalled, was 9,834.5 and the variance of a sample of 10 companies from the second year was 18,832.1. The question is: Do these two sample variances differ sufficiently to

¹ Cf. PITMAN, E. J. G., "Test of Hypotheses concerning Location and Scale Parameters," *Biometrika*, Vol. 31 (1939), pp. 200-215. The L used above is equal to Pitman's $2L$ and is not the same as his L .

indicate that the population variances are not the same? This will be answered by setting up the null hypothesis that the populations have the same variance and then determining whether this hypothesis can reasonably be accepted on the basis of the sample results. In making this test it will be presumed that the investigator is indifferent as to whether any possible difference between the variances is in one direction or the other.

The quantities required for the calculation of the statistic L are

$$\sigma^2 = \frac{10(9,834.5) + 10(18,832.1)}{10 + 10 - 2} = 15,926$$

$$\log_e \sigma^2 = 2.30259 \log_{10} 15,926 = 9.675730$$

$$\log_e \sigma_1^2 = \log_e \frac{N_1 \sigma_1^2}{N_1 - 1} = 2.30259 \log_{10} \frac{10(9,834.5)}{9} = 9.299017$$

$$\log_e \sigma_2^2 = \log_e \frac{N_2 \sigma_2^2}{N_2 - 1} = 2.30259 \log_{10} \frac{10(18,832.1)}{9} = 9.948582$$

$$\alpha = \left(\frac{1}{3}\right) \left(\frac{1}{9} + \frac{1}{9} - \frac{1}{9+9}\right) = \frac{1}{18}$$

$$\begin{aligned} \text{Numerator of } L &= (9 + 9)(9.675730) - [9(9.299017) \\ &\quad + 9(9.948583)] \\ &= 174.163140 - 173.228391 \\ &= .934749 \end{aligned}$$

and L equals this number divided by $1 + \alpha$, that is,

$$L = \frac{.934749}{\frac{19}{18}} = .88555$$

The coefficient of risk will be taken equal to .05 as previously, and this region of rejection will be the upper .05 tail of the χ^2 distribution for $n = 1$. The table of the χ^2 distribution in the Appendix (Table VIII) shows that the lower limit of this region is 3.841. Since the value of L is well below this limit, the null hypothesis is not rejected and it is concluded that the two samples may reasonably have come from populations with the same variance.

ARE TWO SAMPLES FROM THE SAME POPULATION?

The Problem. The problems so far considered have been concerned with whether the normal populations from which two samples have been drawn differ with respect to a single charac-

teristic, for example, with respect to their means or with respect to their variances. The questions posed were: (1) On the assumption that the variances of the populations are the same, are the means also the same? (2) Without any assumption regarding the means of the populations, are their variances the same? The question now to be raised is: Are both the means and variances the same? This question differs from (1) in that there the variances were assumed to be the same, whereas in the new question the equality of the variances is part of the hypothesis to be tested.

For example, suppose a manufacturer of automobile tires is comparing two processes. If he knows or has good reason to believe that the variability in mileage is the same for tires manufactured by one process as for those manufactured by the other, and if he is interested only in a possible difference in average mileage, he will use one of the procedures above for testing the difference between means. If he does not care about a possible difference in average mileage but is interested only in a possible difference in variability, he will use one of the procedures for testing the difference between variances. Finally, if he is interested in whether the tires manufactured by the two processes differ either with respect to average mileage or with respect to variability, or with respect to both, he will employ the test outlined below.

The Test. When the populations are normal, the best test that can be made of joint equality of means and variances appears to be offered by the statistic¹

$$\lambda_H = \left(\frac{\sigma_1}{\sigma_0} \right)^{N_1} \left(\frac{\sigma_2}{\sigma_0} \right)^{N_2} \quad (10)$$

in which σ_0 is the standard deviation of the two samples treated as a single sample and σ_1 and σ_2 are the two individual sample standard deviations.² The value of σ_0^2 and hence of σ_0 , may be found by throwing the two samples together and calculating the

¹ The symbol λ_H is used in the original article and is continued here. The H has no significance other than to distinguish the statistic.

² NEYMAN, J., and E. S. PEARSON, "On the Problem of Two Samples," *Bulletin international de L'Académie polonaise des sciences et des lettres, Classe des sciences mathématiques et naturelles. Series A: Sciences mathématiques* (1930), pp. 73-96.

value of $\frac{\sum(X_i - \bar{X})^2}{N_1 + N_2}$ where $\bar{X}$ is the mean of the combined samples. If only the value of the sample means and standard deviations are known, σ_0 may be computed from the relationship

$$\sigma_0^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2} + \frac{N_1N_2}{(N_1 + N_2)^2} (\bar{X}_1 - \bar{X}_2)^2 \quad (11)$$

Table 48 shows the lower¹ .05 and .01 points of the sampling distribution of λ_H for various values of N_1 and N_2 . The use of this distribution in testing a joint hypothesis will now be illustrated.

TABLE 48.—APPROXIMATE VALUES OF THE LOWER .05 (IN BOLDFACE TYPE) AND .01 POINTS OF THE SAMPLING DISTRIBUTION OF λ_H FOR SELECTED VALUES OF N_1 AND N_2 ¹

Approximate values of λ_H						
$N_1 \backslash N_2$	5	10	20	50	∞	Probability points $P(\lambda_H \geq)$
5	.0167 .0019	.0222 .0029	.0241 .0033	.0247 .0034	.0248 .0034	.05 .01
10	.0222 .0029	.0312 .0048	.0349 .0058	.0364 .0061	.0368 .0062	.05 .01
20	.0241 .0033	.0349 .0058	.0401 .0071	.0425 .0078	.0432 .0080	.05 .01
50	.0247 .0034	.0364 .0061	.0425 .0078	.0459 .0088	.0473 .0092	.05 .01
∞	.0248 .0034	.0368 .0062	.0432 .0080	.0473 .0092	.0500 .0100	.05 .01

¹ Adapted by permission from J. Neyman and E. S. Pearson, "On the Problem of Two Samples," *Bulletin international de l'Académie polonaise des sciences et des lettres, Classe des sciences mathématiques et naturelles. Série A: Sciences mathématiques* (1930), p. 92, Tables II and III.

An Example. Consider once again the employment data of small industrial companies in 1929 and 1935. The two samples of 10 already analyzed for differences in means and variances

¹ The greater the differences, the smaller the values of λ .

had means of 89.1 and 123.5 and variances of 9,834.5 ($= 99.16^2$) and 18,832.1 ($= 137.2^2$). The question is: In view of these sample results could these two samples have reasonably come from the same population? To answer this question the null hypothesis is set up that the populations are the same, and the value of λ_H is calculated. For the given data, the value of σ_0^2 is

$$\frac{(10)(9,834.5) + (10)(18,832.1)}{10 + 10} + \frac{(10)(10)}{(10 + 10)^2} (89.1 - 123.5)^2 = 14,629.1$$

which gives $\sigma_0 = 120.9$. Hence,

$$\log \lambda_H = (10)(\log 99.16 - \log 120.9) + (10)(\log 137.2 - \log 120.9) = -.31170 = 9.68830 - 10$$

and

$$\lambda_H = .4879$$

Table 48 indicates that, the greater the difference between the means and variances, the smaller the value of λ_H ; that is, small values of λ_H are significant. Since the foregoing value of λ_H is larger than either the .05 or the .01 value for $N_1 = 10$ and $N_2 = 10$, the null hypothesis must be accepted. In other words, there is little reason in this case to believe that the two samples are not from identical populations.

In conclusion, it should be noted that, if this test should show a significant difference between the two samples, there is nothing in the result itself that will tell whether the difference lies in the means or in the variances, or in both. In fact, the result is purely a joint product; for the significance of the difference between the means will depend to some extent on the amount of difference in the variances, and vice versa. In the present instance, the individual tests showed no significant difference between either the means or the variances, and it could not therefore be expected that the two samples as a whole would be deemed significantly different. In some cases, however, one of the individual tests might show a significant difference, while the joint test would fail to show any difference. As suggested above, the conclusions to be drawn from the various tests depend entirely upon the problem. Each type of problem has its own appropriate test and should not be confused with tests appropriate for other types of problem.

DIFFERENCE BETWEEN TWO INDEPENDENTLY DERIVED CORRELATION COEFFICIENTS

The difference between two sample correlation coefficients can readily be tested by making use of the z transformation.¹ A sample z , corresponding to a sample r , it will be recalled, has a sampling distribution that is approximately normal in form, with a variance equal to $\frac{1}{N-3}$. Accordingly, the difference between two sample z 's is also practically normal in form with a variance equal to²

$$\sigma_z^2 = \frac{1}{N-3} + \frac{1}{N'-3} \quad (12)$$

Thus to test whether z_{12} is significantly different from z'_{12} it is merely necessary to note whether³

$$\frac{z_{12} - z'_{12}}{\sqrt{\frac{1}{N-3} + \frac{1}{N'-3}}} \geq 1.96$$

The problem is the same as the difference between two sample means when the variances are known. Thus, if $r_{12} = .7658$ and $r'_{12} = .7398$, and hence $z_{12} = 1.01$ and $z'_{12} = .95$; and if $N = 40$ and $N' = 60$, the quantity

$$\frac{z_{12} - z'_{12}}{\sqrt{\frac{1}{N-3} + \frac{1}{N'-3}}} \quad \text{would be equal to} \quad \frac{1.01 - .95}{\sqrt{\frac{1}{37} + \frac{1}{57}}}$$

or 2.86. Since this is greater than 1.96, it may be concluded that the two correlation coefficients r_{12} and r'_{12} are significantly different.

DIFFERENCE BETWEEN TWO INDEPENDENTLY DERIVED REGRESSION PARAMETERS

The difference between two independently derived regression parameters can be tested in the same way as the difference

¹ See Chap. XII.

² If two variables are independent, the distribution of the difference between two normally distributed variables is normally distributed and has a variance that is the sum of the variances of the two variables (*cf.* Appendix to this chapter, pp. 419-421).

³ This assumes a region of rejection equally distributed at each end.

between two means. First, the sample higher-order variances are pooled to give an estimate of the population higher-order variance. Thus

$$\hat{\sigma}_{1.23 \dots}^2 = \frac{N'(\sigma'_{1.23 \dots})^2 + N''(\sigma''_{1.23 \dots})^2}{N' + N'' - 2k} \quad (13)$$

Then the standard error of the difference between the two regression parameters, say, between $b'_{12.3 \dots}$ and $b''_{12.3 \dots}$ would be given by

$$\hat{\sigma}_{b'_{12.3 \dots} - b''_{12.3 \dots}} = \sqrt{\hat{\sigma}_{b'_{12.3 \dots}}^2 + \hat{\sigma}_{b''_{12.3 \dots}}^2} \quad (14)$$

in which

$$\begin{aligned} \hat{\sigma}_{b'_{12.3 \dots}} &= \frac{\hat{\sigma}_{1.23 \dots}}{\sigma'_{2.34 \dots} \sqrt{N'}} \\ \hat{\sigma}_{b''_{12.3 \dots}} &= \frac{\hat{\sigma}_{1.23 \dots}}{\sigma''_{2.34 \dots} \sqrt{N''}} \end{aligned}$$

The test of the difference could then be made by comparing the difference between the b 's with the standard error of the difference, using the t distribution if the sample is small or the normal curve if the sample is large.

For example, suppose that, in a given problem, $N' = 50$, $b'_{12.3} = 2.7$, $(\sigma'_{1.23})^2 = 25$, and $(\sigma'_{2.3})^2 = 43$, while $N'' = 70$, $b''_{12.3} = 3.1$, $(\sigma''_{1.23})^2 = 36$, and $(\sigma''_{2.3})^2 = 48$. Then

$$\hat{\sigma}_{1.23}^2 = \frac{(50)(25) + (70)(36)}{50 + 70 - 6} = 33.07$$

since $k = 3$, which is the number of regression statistics in the regression equation. Also,

$$\hat{\sigma}_{b'_{12.3}}^2 = \frac{33.07}{(43)(70)} = .0154$$

and

$$\hat{\sigma}_{b''_{12.3}}^2 = \frac{33.07}{(48)(70)} = .0098$$

Hence,

$$\hat{\sigma}_{b'_{12.3} - b''_{12.3}}^2 = .0154 + .0098 = .0252$$

and

$$\hat{\sigma}_{b'_{12.3} - b''_{12.3}} = \sqrt{.0252} = .16$$

The difference between $b'_{12.3}$ and $b''_{12.3}$ is $-.4$, and the ratio of this difference to its σ is $-.4/.16 = -2.5$. Since the samples are large, the sampling distribution of $b'_{12.3} - b''_{12.3}$ may be assumed to be normal. Hence a deviate of -2.5 lies beyond both the .05 point

and the .025 point—i.e., beyond -1.645 and beyond -1.96 —and the hypothesis that $b'_{12.3}$ and $b''_{12.3}$ came from the same population is to be rejected. That is, $b'_{12.3}$ must be deemed “significantly different” from $b''_{12.3}$.

APPENDIX

THE MEAN AND VARIANCE OF THE SUM OR DIFFERENCE OF TWO VARIABLES

I. The mean. Let Z be equal to the sum (or difference) of the variables X and Y so that $Z_{ij} = X_i \pm Y_j$. Let X take on the values X_1, X_2 , and X_3 and Y the values Y_1, Y_2 , and Y_3 ; and let the joint relative frequencies, or probabilities, of pairs of X and Y be as follows:

Y_3	p_{13}	p_{23}	p_{33}
Y_2	p_{12}	p_{22}	p_{32}
Y_1	p_{11}	p_{21}	p_{31}
	X_1	X_2	X_3

Thus p_{12} means the probability of an X_1Y_2 combination, p_{31} the probability of an X_3Y_1 combination, etc. In the interests of simplicity, only three values are taken for each variable. The argument is equally valid, however, for any number of values for each variable and for continuous as well as for discrete distributions.

Since $Z_{ij} = X_i \pm Y_j$, the various values of Z and their probabilities are as follows:

Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$
$X_1 \pm Y_1$	p_{11}	$X_2 \pm Y_1$	p_{21}	$X_3 \pm Y_1$	p_{31}
$X_1 \pm Y_2$	p_{12}	$X_2 \pm Y_2$	p_{22}	$X_3 \pm Y_2$	p_{32}
$X_1 \pm Y_3$	p_{13}	$X_2 \pm Y_3$	p_{23}	$X_3 \pm Y_3$	p_{33}

This is the distribution of Z .

The individual distributions of X and Y are as follows:

X	$P(X)$	Y	$P(Y)$
X_1	$p_1 = p_{11} + p_{12} + p_{13}$	Y_1	$p'_1 = p_{11} + p_{21} + p_{31}$
X_2	$p_2 = p_{21} + p_{22} + p_{23}$	Y_2	$p'_2 = p_{12} + p_{22} + p_{32}$
X_3	$p_3 = p_{31} + p_{32} + p_{33}$	Y_3	$p'_3 = p_{13} + p_{23} + p_{33}$

The problem of this section is to express the mean of Z in terms of the means of X and Y .

By definition, $\bar{Z} = \Sigma P(Z)Z$, and when written out in full this becomes

$$p_{11}(X_1 \pm Y_1) + p_{21}(X_2 \pm Y_1) + p_{31}(X_3 \pm Y_1) \\ + p_{12}(X_1 \pm Y_2) + p_{22}(X_2 \pm Y_2) + p_{32}(X_3 \pm Y_2) \\ + p_{13}(X_1 \pm Y_3) + p_{23}(X_2 \pm Y_3) + p_{33}(X_3 \pm Y_3)$$

Upon removal of parentheses and collection of common terms, the mean of Z is seen to be equal to

$$[(p_{11} + p_{12} + p_{13})X_1 + (p_{21} + p_{22} + p_{23})X_2 \\ + (p_{31} + p_{32} + p_{33})X_3] \pm [(p_{11} + p_{21} + p_{31})Y_1 \\ + (p_{12} + p_{22} + p_{32})Y_2 + (p_{13} + p_{23} + p_{33})Y_3]$$

Hence,

$$\bar{Z} = (p_1 X_1 + p_2 X_2 + p_3 X_3) \pm (p'_1 Y_1 + p'_2 Y_2 + p'_3 Y_3) \\ = \bar{X} \pm \bar{Y}$$

That is, the mean of the sum or difference of two variables is the sum or difference of their means. (This is true whether the variables are independent or correlated.)

II. Variance. Let X and Y both be measured from their mean values, and let z be the sum (or difference) of X and Y when so measured. Hence $z_{ij} = x_i \pm y_j$. From I, it follows that $\bar{z} = \bar{x} \pm \bar{y}$; and since $\bar{x} = \bar{y} = 0$, $\bar{z}$ also equals 0. Thus the variances of the variables become

$$\sigma_x^2 = \Sigma p_i x_i^2 \quad \sigma_y^2 = \Sigma p'_j y_j^2 \quad \sigma_z^2 = \Sigma p_{ij} z_{ij}^2$$

The problem is to express σ_z^2 in terms of σ_x^2 and σ_y^2 .

When written out in full, σ_z^2 is as follows:

$$p_{11}(x_1 \pm y_1)^2 + p_{21}(x_2 \pm y_1)^2 + p_{31}(x_3 \pm y_1)^2 \\ + p_{12}(x_1 \pm y_2)^2 + p_{22}(x_2 \pm y_2)^2 + p_{32}(x_3 \pm y_2)^2 \\ + p_{13}(x_1 \pm y_3)^2 + p_{23}(x_2 \pm y_3)^2 + p_{33}(x_3 \pm y_3)^2$$

Upon clearing parentheses and collecting common terms, this becomes

$$(p_{11} + p_{12} + p_{13})x_1^2 + (p_{21} + p_{22} + p_{23})x_2^2 + (p_{31} + p_{32} + p_{33})x_3^2 \\ + (p_{11} + p_{21} + p_{31})y_1^2 + (p_{12} + p_{22} + p_{32})y_2^2 \\ + (p_{13} + p_{23} + p_{33})y_3^2 \pm 2(p_{11}x_1y_1 + p_{21}x_2y_1 + p_{31}x_3y_1 + p_{12}x_1y_2 \\ + p_{22}x_2y_2 + p_{32}x_3y_2 + p_{13}x_1y_3 + p_{23}x_2y_3 + p_{33}x_3y_3)$$

which reduces to

$$p_1x_1^2 + p_2x_2^2 + p_3x_3^2 + p'_1y_1^2 + p'_2y_2^2 + p'_3y_3^2 \pm 2\sum p_{ij}x_iy_j$$

Accordingly,

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 \pm 2r\sigma_x\sigma_y \quad \text{since } r = \frac{\sum p_{ij}x_iy_j}{\sigma_x\sigma_y}$$

That is, the variance of a sum of two variables equals the sum of their variances plus twice the correlation coefficient times the product of the two standard deviations. Also, the variance of a difference of two variables equals the sum of their variances minus twice the correlation coefficient times the product of the two standard deviations. Finally, if the two variables are independent, $r = 0$ and the variance of their sum or difference equals the sum of their variances.

CHAPTER XVII

ANALYSIS OF VARIANCE

In recent years, much use has been made of a so-called "analysis of variance."¹ This has been particularly true in the field of biological and agricultural experiments. The method has such wide application, however, that there is hardly any field of statistical investigation in which it cannot be employed.

Analysis of variance is essentially a method of testing for the existence of correlation or association. The technique consists in classification and cross classification on either a qualitative or a quantitative basis and in comparison of the variation from class to class with the variation within classes. The analysis is so arranged that variation within classes can be presumably attributed to chance. The test of association or correlation consists in comparing the variation between classes with the supposed chance variation within classes. The problems discussed below will illustrate the details of this procedure.

PROBLEMS INVOLVING A SINGLE BASIS OF CLASSIFICATION

Nature of Problems. In the simplest cases of analysis of variance there is only one basis of classification. This will accordingly be the first type of problem to be discussed.

In Table 49, on page 426 are listed the grades of 15 representative students² in an elementary course in economics. These are

¹ Historically, the method can be traced at least to the 1910 edition of G. Udney Yule's *Introduction to the Theory of Statistics*, in which Chap. V on Manifold Classification introduces the use of contingency tables that are the ancestors of modern analysis of variance tables. A certain similarity also is revealed between analysis of variance and Lexis's analysis of subnormal and supernormal dispersion. Cf. REITZ, L. *Mathematical Statistics*, pp. 146-155. For the original work see W. Lexis, "Über die Theorie der Stabilität statistischer Reihen," *Jahrbücher für Nationalökonomie und Statistik*, Vol. 32 (1879), pp. 60-98; and W. Lexis, *Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik*, (1903), Chaps. V-IX.

² Actually, each grade is the average of the grades of several students, but for the present analysis it will be considered a single individual grade.

classified according to the teacher of the student, and the problem is to determine whether the difference in teaching has an effect on the grades of the students. If the difference in teaching has an effect, it will be reflected in variation in the mean grades of the students as between teachers. The problem is therefore to determine whether the variation in mean student grades from teacher to teacher is greater than can reasonably be attributed to chance.

Theoretical Basis for the Analysis. It may be assumed that the variation between grades of students having the same teacher is due to the chance difference in ability of the students. This would be the case if the students were assigned to the teachers at random. If this variation of grades within each group is pooled for all groups, a good estimate will be secured of how much variation can be expected purely as a result of chance. This will form a standard with which to compare the variation in the mean student grade from teacher to teacher. Of course, the latter cannot be expected to vary as much as individual grades, since they are mean values.¹ Allowance for the smaller variation among means can be made, however, by proper weighting of the variation in the mean grades before the comparison is made.

The theoretical basis for the precise method of comparison that is used may be outlined as follows: The first step in the analysis is to set up the null hypothesis that the difference in teaching has no effect on the grades. Under these conditions both the variation in the means of student grades from teacher to teacher and the variation of grades within the groups of students having the same teacher will stem back to the variation in general in student grades, or to what may be called the "population variance." If this is large, variation in the mean student grades from teacher to teacher will tend to be large, as will also variation around these means. If, on the other hand, the population variance is small, the variation in mean student grades from teacher to teacher will tend to be small and so will variation around these means.

Accordingly, if the null hypothesis is correct, either the sample variation in the mean student grades from teacher to teacher or the variation around these mean grades could be used to esti-

¹ It will be recalled that the standard error of a mean is equal to $\sigma/\sqrt{N}$.

mate the underlying population variance. Of course, the two estimates could not be expected to be the same.¹ In one sample the estimate based upon variation in the means would be larger than that based upon variation around the means, and in another sample the opposite might be true. If the null hypothesis is true, however, the ratio of the first to the second would tend to average in the neighborhood of unity, and large deviations from unity would not be very likely. This suggests that, in any particular problem, the null hypothesis could be tested by computing the ratio of the two estimates of variance and seeing whether or not the sample ratio deviated unreasonably from unity.

Such, in general, is the theoretical basis of analysis of variance. To put it to use, however, requires more exact specifications regarding the nature of the two estimates of variance and the form of the sampling distribution of their ratio.

The estimates of variance that are used are maximum likelihood estimates. If a large number of this kind of sample estimates are obtained, their mean will have approximately the same value as that of the population variance. In other words, the mean of the sampling distribution of a maximum-likelihood estimate of variance is the population variance. For this reason a maximum-likelihood estimate of variance is also called an "unbiased" estimate.

Mathematical analysis shows that the maximum-likelihood estimate of the population variance that is based upon the variation in the means is given by $\frac{\sum N_r(\bar{X}_r - \bar{X})^2}{r - 1}$, that is, by the weighted sum of the squares of the deviations of the individual means about the mean of the entire sample of grades divided by the number of means minus 1. The N_r is the number of student grades used to calculate the r th mean; r is the number of means calculated and in this case would also be the number of teachers. Such an estimate is said to be based on $r - 1$ degrees of freedom.

There are $r - 1$ and not r degrees of freedom because the sample estimate is calculated from the variation around the grand mean of the entire sample and not the mean of the population. To allow for variation in the mean of the entire sample, the factor $r - 1$ is used instead of r .

¹ It can be shown that the two estimates are independent of each other, so that the value of one is not related to the value of the other.

A second maximum-likelihood estimate of the population variance is that based upon the variation around the means,

which is given by $\frac{\sum_i \sum_r (X_{ir} - \bar{X}_r)^2}{N - r}$ —that is, by the pooled sum of the squared deviations from the individual means divided by the number of cases minus the number of means. This is an estimate of the population variance based on $N - r$ degrees of freedom.

These are the two estimates of the population variance that are used in the analysis of variance. The next question is: What is the sampling distribution of their ratio? The answer given by mathematical analysis is as follows: If the original population is normal and if the null hypothesis is correct, the ratio of the two maximum-likelihood estimates of the population variance will tend to fluctuate from sample to sample in accordance with the F distribution. If the estimate based upon the variation in the means is put in the numerator of the ratio and the estimate based upon the variation around the means is put in the denominator, the appropriate F curve is that for which $n_1 = r - 1$ and $n_2 = N - r$.

The Numerical Analysis. Before proceeding to the application of the foregoing theory to a concrete problem, certain mathematical relationships should be noted that will be helpful in carrying out the numerical calculations. A study of the theoretical formulas shows that the numerator of each of the estimates of variance is a sum of squares. These could be calculated directly, of course, but it is usually easier to make use of the following identity,

$$\sum_i (X_i - \bar{X})^2 \equiv \sum_r N_r (\bar{X}_r - \bar{X})^2 + \sum_r \sum_i (X_{ir} - \bar{X}_r)^2 \quad (1)$$

which says that the total variation in the data, as represented by the total sum of squares, may be broken up into two parts, one consisting of the variation in the means (as represented by the weighted sum of the squared deviations of the individual means from the grand mean) and the other consisting of the variation around the means (as represented by the pooled sum of the squared deviations of the individual items from the mean of each group).

Since it is usually easier to calculate the total sums of squares and that pertaining to the means than it is to compute the sum of squares of the residual deviations, it is commonly the practice to calculate the latter by taking the difference between the other two more easily calculated values.

In calculating the sums of squares the following special formulas are found useful. Thus the total sum of squares can be computed from the equation¹

$$\sum (X_i - \bar{X})^2 = \sum X_i^2 - N\bar{X}^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{N} \quad (2)$$

and the weighted sums of squares of the deviations of the individual means from the grand mean can be computed from the equation²

$$\sum_r N_r(\bar{X}_r - \bar{X})^2 = \sum_r \left[\frac{[\sum X_{ir}]^2}{N_r} - \frac{(\sum X_i)^2}{N} \right] \quad (3)$$

in which $[\sum X_{ir}]^2$ refers to the square of the sum of the grades in the r th row.

TABLE 49.—GRADES OF REPRESENTATIVE STUDENTS¹ CLASSIFIED BY TEACHERS

Teachers	Grades of students			Mean grade
I	83.25	77.50	71.00	77.25
II	88.75	74.75	70.00	77.83
III	76.25	67.25	69.25	70.92
IV	78.75	68.75	62.25	69.92
V	81.50	75.75	64.75	74.00

¹ Grades are actually means of several students' grades, but they are considered as if they were individual student's grades—i.e., each teacher has three students.

The application of these equations to the data of Table 49 is carried out in the calculations of Table 50. This latter table shows that $\sum X_i^2 = 82,850.1875$ and $\frac{(\sum X_i)^2}{N} = 82,103.0042$. Hence the total sum of squares $\sum (X_i - \bar{X})^2$ is equal to $82,850.1875 - 82,103.0042 = 747.1833$.

$$^1 \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2 \sum X_i \bar{X} + N\bar{X}^2 = \sum X_i^2 - N\bar{X}^2, \text{ since } N\bar{X} = \sum X_i.$$

² The equation follows from the fact that $N_r \bar{X}_r = (\sum X_i)_r$ for each row, and the total for all rows is $\sum N_r \bar{X}_r = N\bar{X}$.

Table 50 also shows that

$$\sum_r \frac{[\sum X_i]_r^2}{N_r} - \frac{(\sum X_i)^2}{N} = 82,257.3125 - 82,103.0042 = 154.3083,$$

which is the sum of the weighted squared deviations of the individual means from the grand mean. The difference between

TABLE 50.—WORKSHEET FOR CALCULATING THE VARIOUS SUMS OF SQUARES
FOR ANALYSIS OF VARIANCE

Grades	Partial sums	Squares of	
		Individual grades	Partial sums
83.25	231.75	6,930.5625	53,708.0625
77.50		6,006.2500	
71.00		5,041.0000	
88.75	233.50	7,876.5625	54,522.2500
74.75		5,587.5625	
70.00		4,900.0000	
76.25		5,814.0625	
67.25	212.75	4,522.5625	45,262.5625
69.25		4,795.5625	
78.75		6,201.5625	
68.75	209.75	4,726.5675	43,995.0625
62.25		3,875.0625	
81.50		6,642.2500	
75.75	222.00	5,738.0625	49,284.0000
64.75		4,192.5625	
1,109.75		82,850.1875	246,771.9375
= $\sum X_i$		= $\sum X_i^2$	= $\sum [\sum X_i]_r^2$

$$(\sum X_i)^2 = (1,109.75)^2 = 1,231,545.0625$$

$$\frac{(\sum X_i)^2}{N} = \frac{(1,109.75)^2}{15} = 82,103.0042$$

$$\sum_r \frac{[\sum X_i]_r^2}{N_r} = \frac{246,771.9375}{3} = 82,257.3125$$

these two results gives the pooled sum of the squared deviations of the individual items from the row means. This last sum of squares,

$$\sum_r \sum_i (X_{ir} - \bar{X}_r)^2, \text{ is thus equal to } 747.1833 - 154.3083 \\ = 592.875.$$

To make the appropriate estimates of the population variance the last two sums of squares are each divided by the proper degrees of freedom. Thus 154.3083 is divided by $5 - 1 = 4$ to give 38.5771 as the estimate of the population variance based on the variation among teachers in mean student grades. Similarly, 592.8750 is divided by $15 - 5 = 10$ to give 59.2875 as the estimate of variance based on the variation around the means. The ratio of these two estimates is $38.5771/59.2875 = .65$.

To determine whether this sample ratio justifies the rejection or acceptance of the null hypothesis requires the selection of a suitable region of rejection. On the assumption that the risk of rejecting the null hypothesis when it is true will be put at the usual figures of 1 in 20, the size of the region of rejection will be .05. This is purely arbitrary, however, and might be set at .01 or .10 or any other figure, depending on the coefficient of risk adopted. A more significant question relates to the distribution of the region. Since the difference in teaching, if it had any effect, would tend to be revealed in a larger variation among the teachers in mean student grades, values of the ratio that are smaller than 1 would seem to be of little significance. The risk that is to be minimized by the proper choice of the region of rejection is the risk of accepting the null hypothesis when in fact the variation among teachers in the mean student grade is larger than may be due to chance. It would seem, therefore, that the proper region of rejection in this instance would be the upper .05 tail of the F distribution. This is the region that will be adopted in this and in all subsequent analyses of variance.

For $n_1 = 4$ and $n_2 = 10$, the upper .05 point of the F distribution is 3.478. The sample ratio in the present problem is .65 and obviously does not fall in the region of rejection.¹ Therefore the null hypothesis is accepted, and the variation in mean student grades from teacher to teacher is to be attributed to chance and not to any difference in teaching.

¹ If the ratio is less than unity, it will never fall in the region of rejection. In such cases, calculation of the two estimates of variance will of itself give sufficient evidence for acceptance of the null hypothesis.

PROBLEMS INVOLVING MORE THAN ONE BASIS OF CLASSIFICATION

A Single Case in Each Class. In the previous problem it might have been suspected that, if allowance were made for the general standing of the students, the difference in teaching might be revealed. Suppose, as is actually the case, that the first student of each group is a high-standing student, the second is one of average standing, and the third one of low standing. Then the 15 grades may be classified according to two criteria, standing and teacher, and mean grades may be calculated for the divisions of each classification. This is done in Table 51, on page 432.

Under such circumstances, two questions may be asked: (1) If allowance is made for the difference in standing, has the difference in teaching any significant effect on grades in the given course? (2) If allowance is made for the difference in teaching, has the difference in standing any significant effect on grades in the given course? The former question is the one that primarily concerns us here, but the second may also be of interest.

Theoretical Basis for the Analysis. In this more extended problem there are three types of variation to be considered. First, there is the variation in the means of the rows, the mean student grades for different teachers; second, the variation in the means of the columns, the mean grades of students of different standings; third, the variation in the individual grades about what would be expected from the combined row (teaching) and column (standing) effects.

This third variation needs further explanation. If a student's grade differed from the grand mean of all the students' grades by just the same amount that the average grades of all students having the same teacher differed from the grand mean plus the amount that the average grades of all students having the same standing differed from the grand mean, the grade of this student would be equal to $\bar{X} + (\bar{X}_r - \bar{X}) + (\bar{X}_c - \bar{X})$. In nearly every case a student's grade does not come exactly to the sum of this combined row and column effect but differs from it. This difference will be given by¹

$$X_{rc} - \bar{X} - (\bar{X}_r - \bar{X}) - (\bar{X}_c - \bar{X}) = X_{rc} - \bar{X}_r - \bar{X}_c + \bar{X}$$

¹ Individual grades are now designated by X_{rc} instead of X_i since any

It is the variation in such differences that constitutes the third type of variation distinguished above.

Whereas the variation in the row means might be due to the difference in teaching and the variation in the column means might be due to the difference in standing, this last variation is presumably due to chance. For brevity it will be called the "remainder" variation.

If the null hypothesis is set up that the difference in teaching has no effect on the grades, the size of both the variation in the row means and the size of the remainder variation will depend on the size of the variation in students' grades in general—*i.e.*, on the size of the population variance. As in the previous problem, each of these two kinds of variation could be used to make an independent estimate of the population variance, and their ratio could be used to test the null hypothesis. The analysis is essentially the same as in the previous instance except that now, according to the hypothesis, the measure of the chance variation does not contain any possible effect of the difference in standing. This test of teaching is independent of any possible effect of standing.

In the same manner, the null hypothesis that the difference in standing does not have any effect on the students' grades can be tested by comparing an estimate of the population variance based on the variation in the column (standing) means with an estimate based on the remainder (chance) variation. This is a test of standing that is independent of the possible effect of teaching.

In making the foregoing tests, the maximum-likelihood estimates of the population variance that are used are as follows.¹ The estimate of the population variance based upon the variation

in the row means is $\frac{\sum N_r(\bar{X}_r - \bar{X})^2}{r - 1}$. The estimate based upon

the variation in the column means is $\frac{\sum N_c(\bar{X}_c - \bar{X})^2}{c - 1}$, and the

individual student has both a teacher and a standing and there is only one student for any given combination of the two.

¹ That these are the proper formulas can be demonstrated by strict mathematical analysis similar to that of Chap. X. For further discussion, see J. O. Irwin, "Mathematical Theorems Involved in the Analysis of Variance," *Journal of the Royal Statistical Society*, Vol. 94 (1931), p. 284.

estimate based upon the remainder variation is

$$\frac{\sum_c \sum_r (\dot{X}_{rc} - \bar{X}_r - \bar{X}_c + \bar{X})^2}{(r-1)(c-1)}$$

If the null hypothesis that teaching has no effect is correct, and if the population is normal, the ratio of the first estimate to the last will have a sampling distribution that is of the form of the F distribution with $n_1 = r - 1$ and $n_2 = (r - 1)(c - 1)$. Likewise, if the null hypothesis that standing has no effect is correct and if the population is normal, the ratio of the second estimate to the last will have a sampling distribution of the form of the F -distribution with $n_1 = c - 1$ and $n_2 = (r - 1)(c - 1)$. It is upon these theoretical conclusions that the following numerical analysis rests.

The Numerical Analysis. In putting the foregoing theory into practice, use is made of the identity

$$\begin{aligned} \sum_r \sum_c (X_{rc} - \bar{X})^2 &= \sum_r N_r (\bar{X}_r - \bar{X})^2 + \sum_c N_c (\bar{X}_c - \bar{X})^2 \\ &\quad + \sum_r \sum_c (X_{rc} - \bar{X}_r - \bar{X}_c + \bar{X})^2 \quad (4) \end{aligned}$$

which merely says that the total sum of squares is equal to the sum of the sum of squares for each of the three variations—the variation in the means of the rows, the variation in the means of the columns, and the remainder variation. The total sum of squares and the sum of squares for the means of the rows have already been computed. If the sum of squares for the means of the columns is also computed, the remainder sum of squares can be calculated from the foregoing identity.

The equation for the calculation of the sum of squares for the means of the columns is similar to that for the means of the rows and takes the form

$$\sum_c N_c (\bar{X}_c - \bar{X})^2 = \sum_c \frac{\left[\sum_r X_{rc} \right]_c^2}{N_c} - \frac{(\sum X_{rc})^2}{N} \quad (5)$$

in which $\left[\sum_r X_{rc} \right]_c^2$ refers to the squared sum of the r grades in

the c th column. For the given data, this is equal to

$$82,621.1625 - 82,103.0047 = 518.1583$$

From Tables 51 and 52, it is seen that

$$\sum_c \frac{\left[\sum_r X_{rc} \right]^2}{N_c} = \frac{413,105.8125}{5} = 82,621.1625$$

From Table 50 it was found that

$$\frac{(\sum X_{rc})^2}{N} = 82,103.0042$$

Hence, by Eq. (5),

$$\sum N_c (\bar{X}_c - \bar{X})^2 = 82,621.1625 - 82,103.0042 = 518.1583$$

TABLE 51.—GRADES OF STUDENTS CLASSIFIED BY TEACHER AND STANDING

Teacher	Standing			Mean grade
	High	Medium	Low	
I	83.25	77.50	71.00	77.25
II	88.75	74.75	70.00	77.83
III	76.25	67.25	69.25	70.92
IV	78.75	68.75	62.25	69.92
V	81.50	75.75	64.75	74.00
Mean grade	81.70	72.80	67.45	73.98

TABLE 52.—CALCULATION OF SUM OF SQUARES FOR MEANS OF COLUMNS

	Sums of columns	Squares of sums
High standing.....	408.50	166,872.2500
Medium standing.....	364.00	132,496.0000
Low standing.....	337.25	113,737.5625
		413,105.8125
		$= \sum_c \left[\sum_r X_{rc} \right]^2$

The remainder variation, $\sum_r \sum_c (\bar{X}_{rc} - \bar{X}_r - \bar{X}_c + \bar{X})^2$, is cal-

culated from the identity (4) and is found to be equal to¹

$$747.1833 - 518.1583 - 154.3083 = 74.7167$$

The foregoing data may be assembled in an "analysis of variance" table (Table 53), in which the various estimates of σ^2 can be calculated.

TABLE 53.—ANALYSIS OF VARIANCE

	Sum of squares	Degrees of freedom	Unbiased estimate of σ^2
Means of rows.....	154.3083	$r - 1 = 4$	38.5771
Means of columns.....	518.1583	$c - 1 = 2$	259.0792
Remainder.....	74.7167	$(r - 1)(c - 1) = 8$	9.3396
Total.....	747.1833	$N - 1 = 14$	

To test the null hypothesis that the variation in the means of the columns is due to chance, *i.e.*, that the difference in standing has no effect on the grades in the given course, the ratio $259.0779/9.3507$ is calculated. The result is 27.707. For $n_1 = 2$ and $n_2 = 8$, the .05 value for the F distribution is 4.459. The sample ratio 27.707 is much greater than 4.459 and thus clearly lies in the region of rejection. The null hypothesis cannot be accepted, and it is to be concluded that the difference in standing does have definite bearing on the grades obtained in the given course.

To test the null hypothesis that the variation in the means of the rows is due to chance, *i.e.*, that the difference in teaching has no effect on the grades, the ratio $38.5765/9.3507$ is calculated. The result is 4.126. For $n_1 = 4$ and $n_2 = 8$, the .05 point of the F distribution is 3.838. The proximity of the sample value 4.126 to the .05 point, 3.838, suggests that this problem offers a borderline case. Nevertheless, it does lie in the region of rejection, since the sample value is the larger; and if the rule of procedure is strictly followed the second null hypothesis cannot be accepted. It may be concluded in this case that the difference in mean teacher grades is to be attributed to the difference in teaching.

¹ The total sum of squares is 747.1833 and was calculated on p. 426. The sum of squares for the row means is 154.3083 and was calculated on p. 427.

It will be recalled that in the previous case the hypothesis that the difference in teaching had no effect on the grades was not rejected. Here it is rejected. The reason for the different conclusion lies in the different measure of chance variation. In the former problem all the variation over and above the variation in the mean student grades from teacher to teacher was taken to be chance variation. This represented chance variation in student ability in the course, due to any cause whatever. Compared with such chance variation, the variation in mean student grades from teacher to teacher was not deemed significant. That is, the variation in mean student grades from teacher to teacher might easily have been explained by the chance difference in ability of the students assigned to each teacher.¹

In the second problem, in which not only different teachers but differences in standing are involved, the measure of chance variation is taken to be the variation remaining after the variation due to difference in teaching and the variation due to difference in standing have both been eliminated. That is, it is the chance variation in student ability within the groups of high, average, and low standing. It is the chance variation remaining after student standing has been roughly accounted for. Compared with such a chance variation the difference in teaching does appear to have some effect on the students' grades.

In conclusion, it should be pointed out that in the present problem the variation of the means of the rows is independent of whether the variation in the means of the columns is due to chance or not. The test is valid in any case. If it should appear from a previous test that variation in the means of the columns is due to chance, then it is better to combine this variation with the remainder variation so as to get an estimate of chance variation based upon a greater number of degrees of freedom. For it is better to use the test pertinent to a single basis of classification than to use the present test, if the variation in the means of the columns is due to chance, because the estimate based on a larger number of degrees of freedom gives a more reliable estimate of

¹ Strictly speaking, the analysis in that problem was not valid owing to the way in which the students grades were selected. The students were not assigned to the teachers entirely at random, but in order to serve for further analysis the students were so chosen that each teacher had an equal number of high-, average-, and low-standing students.

the chance variation. What has been said about variation in the means of the rows applies in turn to the variation in the means of the columns. The test is equally independent of whether the variation in the means of the rows is due to chance or not. If the latter is due to chance, however, the test of the former is best based upon a chance variation that includes the variation in the means of the rows.

More than One Case in Each Class. *The Problem.* In the preceding problems it was assumed that there was only one student of each standing assigned to each teacher. As a matter of fact, four students of each standing were assigned to each teacher, and the grades of the previous problem were actually the mean grades of these groups of four students, rather than grades of individual students. The full data are given in Table 54, on page 437.

The problem now to be considered is the analysis of variance of this full set of data shown in Table 54. Three questions may be asked: (1) Is the variation in the means of the rows greater than may be reasonably attributed to chance? (2) Is the variation in the column means greater than may reasonably be attributed to chance? (3) Are the individual cell means what may be reasonably expected on the assumption that each individual item is a mere sum of a row effect plus a column effect, or do these cell means give evidence of an interaction between row and cell effects? In other words, the three questions are: (1) Has the difference in teaching an effect on grades? (2) Does difference in general ability affect grades? (3) Is there any interaction between teacher and student ability—i.e., does one teacher do better with high- (or low-) standing students than another?

Theoretical Basis for the Analysis. In this problem there are four types of variation to be considered. There are the variation in the row means, the variation in the column means, the variation in the means of each cell about what may be expected from a linear combination of row and column effects (the so-called "interaction"), and, finally, the variation in the individual items about the means of each cell.

The variation in the individual items about the means of the cells is presumably due to chance and may be spoken of as the "chance" or "remainder" variation. It will be noted that this is not the same as the remainder variation of the preceding

problem. The remainder variation in the preceding problem is now the "interaction." Previously, the remainder variation was presumed to represent chance, for there was no other basis for measuring chance variation. Here there is another basis for measuring chance variation, and the former remainder variation may now be tested for the possible existence of correlation or interaction.

In the present problem three hypotheses may be tested. First there is the null hypothesis that the difference in teaching has no effect on grades; second, the null hypothesis that the difference in standing has no effect on grades; third, the null hypothesis that there is no real interaction between teaching and standing.

To test the first hypothesis an estimate is made of the population variance based upon the variation in the row means, and this is compared with an estimate based upon the remainder or chance variation. To test the second hypothesis an estimate of the population variance based upon the variation in the column means is compared with the estimate based upon chance variation. To test the third hypothesis an estimate based upon the interaction is compared with the chance estimate. The estimates based upon the different sources of variation are the following; each is a maximum-likelihood, or unbiased, estimate:
Estimate based on the variation in the means of rows:

$$\frac{\sum N_r(\bar{X}_r - \bar{X})^2}{r - 1}$$

Estimate based upon the variation in the means of the columns:

$$\frac{\sum N_c(\bar{X}_c - \bar{X})^2}{c - 1}$$

Estimate based upon the interaction:

$$\frac{\sum_r \sum_c N_{rc}(\bar{X}_{rc} - \bar{X}_r - \bar{X}_c - \bar{X})^2}{(r - 1)(c - 1)}$$

Estimate based upon the remainder or chance variation:

$$\frac{\sum_r \sum_c \sum_i (X_i - \bar{X}_{rc})^2}{N - rc}$$

Mathematical analysis shows that, if the first null hypothesis is correct and if the population is normal, the ratio of the first estimate to the last has a sampling distribution of the form of the F distribution with $n_1 = r - 1$ and $n_2 = N - rc$. If the second null hypothesis is correct and the population is normal, the ratio of the second estimate to the last has a sampling distribution of the form of the F distribution with $n_1 = c - 1$ and

TABLE 54.—GRADES OF STUDENTS IN A GIVEN COLLEGE COURSE CLASSIFIED ACCORDING TO THEIR GENERAL STANDING AND TEACHERS

Teacher	High standing	Medium standing	Low standing	Means of rows
I	90	87	80	77.25
	76	77	77	
	85	75	68	
	82	71	59	
	Average 83.25	77.50	71.00	
II	88	58	72	77.83
	85	77	58	
	90	81	80	
	92	83	70	
	Average 88.75	74.75	70.00	
III	84	62	61	70.92
	80	73	77	
	60	70	72	
	81	64	67	
	Average 76.25	67.25	69.25	
IV	80	65	63	69.92
	76	69	70	
	90	77	46	
	69	64	70	
	Average 78.75	68.75	62.25	
V	80	80	47	74.00
	80	80	62	
	75	76	81	
	91	67	69	
	Average 81.50	75.75	64.75	
Means of columns	81.70	72.80	67.45	Grand mean 73.98

$n_2 = N - rc$; and if the third null hypothesis is correct and the population is normal, the ratio of the third estimate to the last has a sampling distribution of the form of the F distribution with $n_1 = (r - 1)(c - 1)$ and $n_2 = N - rc$. The analysis thus proceeds in much the same manner as in the previous problems.

TABLE 55.—CALCULATIONS FOR ANALYSIS OF VARIANCE
Squares of individual grades

8,100	5,776	7,225	6,724	7,744	7,225	8,100
8,464	7,056	6,400	3,600	6,561	6,400	5,776
8,100	4,761	6,400	6,400	5,625	8,281	7,569
5,929	5,625	5,041	3,364	5,929	6,561	6,889
3,844	5,329	4,900	4,096	4,225	4,761	5,929
4,096	6,400	6,400	5,776	4,489	6,400	5,929
4,624	3,481	5,184	3,364	6,400	4,900	3,721
5,929	5,184	4,489	3,969	4,900	2,116	4,900
2,209	3,844	6,561	4,761	Total = 334,735 = $\sum X_i^2$		

Sums of row grades and squares of sums of row grades

927	859,329
934	872,356
851	724,201
839	703,921
888	788,544
Total 4,439 = $\sum X_i$	3,948,351 = $\sum_r [\sum X_i]^2$

Sums of column grades and squares of sums of column grades

1,634	2,669,956
1,456	2,119,936
1,349	1,819,801
Total 4,439 = $\sum X_i$	6,609,694 = $\sum_c [\sum X_i]^2$

The Numerical Analysis. In the numerical calculations use is again made of an identity in which the total sum of squares is broken up into its various components. This identity is as follows:

$$\begin{aligned} \sum (X_i - \bar{X})^2 &= \sum_r N_r (\bar{X}_r - \bar{X})^2 + \sum_c N_c (\bar{X}_c - \bar{X})^2 \\ &+ \sum_r \sum_c N_{rc} (\bar{X}_{rc} - \bar{X}_r - \bar{X}_c + \bar{X})^2 + \sum_r \sum_c \sum_i (X_i - \bar{X}_{rc})^2 \quad (6) \end{aligned}$$

As in the other problems, the first four sums of squares are calculated directly, and the remainder sum of squares is calculated as a residual. These calculations are shown in Table 55.

From Table 54 it is seen that the number of grades in each row is 12; each teacher has 4 students' grades of high, low, and medium standing, making 12 altogether, and hence, $N_r = 12$. The number of grades in each column is 20; therefore, $N_c = 20$. The number of grades in the column of one row is 4; thus, $N_{rc} = 4$. The total number of grades is 60, and $N = 60$. The following calculations may now be made:

The total sum of squares is best calculated from the formula

$$\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N} \quad (7)$$

Table 55 shows that this is equal to

$$\begin{aligned} \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N} &= 334,735 - \frac{(4,439)^2}{60} \\ &= 334,735 - 328,412.02 = 6,322.98 \end{aligned}$$

The weighted sum of the squared deviations of the means of the rows is calculated from the equation

$$\Sigma_r N_r (\bar{X}_r - \bar{X})^2 = \Sigma_r \frac{[\Sigma X_i]_r^2}{N_r} - \frac{(\Sigma X_i)^2}{N} \quad (8)$$

where $[\Sigma X_i]_r^2$ represents the square of the sum of the individual items in the r th row. Table 55 shows this second sum of squares to be

$$\begin{aligned} \Sigma_r \frac{[\Sigma X_i]_r^2}{N_r} - \frac{(\Sigma X_i)^2}{N} &= \frac{3,948,351}{12} - \frac{(4,439)^2}{60} \\ &= 329,029.25 - 328,422.02 = 617.23 \end{aligned}$$

The weighted sum of the squared deviations of the means of the columns may be calculated from the equation

$$\Sigma_c N_c (X_c - \bar{X})^2 = \Sigma_c \frac{[\Sigma X_i]_c^2}{N_c} - \frac{(\Sigma X_i)^2}{N} \quad (9)$$

in which $\left[\sum X_i\right]_c^2$ represents the square of the sum of the individual items in the c th column. Table 55 shows that this third sum of squares is equal to

$$\begin{aligned}\sum_c \frac{\left[\sum X_i\right]_c^2}{N_c} - \frac{\left(\sum_c X_i\right)^2}{N} &= \frac{6,609,694}{20} - \frac{(4,439)^2}{60} \\ &= 330,484.70 - 328,412.02 = 2,072.68\end{aligned}$$

The fourth sum of squares is best calculated indirectly. The first step is to calculate the weighted sum of the squared deviations of the cell means from the grand mean. This may be obtained from the equation

$$\sum_r \sum_c N_{rc}(\bar{X}_{rc} - \bar{X})^2 = \sum_r \sum_c N_{rc}(\bar{X}_{rc})^2 - \frac{\left(\sum X_i\right)^2}{N} \quad (10)$$

Since N_{rc} is constant in the present problem, the first term on the right may be written $N_{rc} \sum_r \sum_c (\bar{X}_{rc})^2$. To calculate this term it is thus necessary only to square each of the cell means, sum the results, and multiply by N_{rc} . This squaring and summing of the cell means was already done in Table 50, however,¹ where it was found that $\sum \bar{X}_{rc}^2$ (designated as $\sum X_c^2$ in Table 50) equaled 82,850.1875. Accordingly, as indicated in Table 55 the weighted sum of the squares of the cell means about the grand mean is equal to

$$\begin{aligned}\sum_c N_{rc}(\bar{X}_{rc})^2 - \frac{\left(\sum X_i\right)^2}{N} &= 4(82,850.1875) \\ &\quad - 328,412.02 = 2,988.73\end{aligned}$$

The second step in the calculation of the fourth sum of squares is to subtract from 2,988.73 the weighted sum of the squared deviations of the means of the rows and the weighted sum of the squared deviations of the means of the columns. The result, according to a modification of Eq. (4), is the fourth sum of squares.

¹ Because, in Table 50, the cell means of Table 54 were treated as if they were individual items; hence the sum of the squared items in Table 50 is equivalent to the sum of the squared means of Table 54.

Thus

$$\sum_r \sum_c N_{rc}(\bar{X}_{rc} - \bar{X}_c - \bar{X}_r + \bar{X})^2 = 2,988.73 - 2,072.68 \\ - 617.23 = 298.82$$

The final, or remainder, sum of squares is now obtained by subtraction of all the other component sums of squares from the total sum of squares. That is,

$$\sum_r \sum_c \sum_i (X_i - \bar{X}_{rc})^2 = 6,322.98 - 2,072.68 - 617.23 - 298.82 \\ = 3,334.25$$

All the foregoing sums of squares are collected in Table 56, where they are divided by the appropriate degrees of freedom to obtain various estimates of the population variance. Comparisons of these estimates afford tests of the various hypotheses to be considered.

TABLE 56.—ANALYSIS OF VARIANCE

	Sums of squares	Degrees of freedom	Estimates of σ^2
Means of rows.....	617.23	5 - 1 = 4	154.31
Means of columns.....	2,072.70	3 - 1 = 2	1,036.35
Means of cells, or interaction.....	298.82	4 × 2 = 8	37.35
Remainder.....	3,334.25	60 - 15 = 45	74.09
Total.....	6,322.98	60 - 1 = 59	

To test the null hypothesis that the variation in the row means is greater than can reasonably be attributed to chance—i.e., that the difference in teaching has no effect on grades—the ratio of the first estimate to the last estimate is computed. This is found to be 2.083. The .05 point of the F distribution for $n_1 = 4$ and $n_2 = 45$ lies between 2.5 and 2.69. Since the sample value of 2.083 is obviously less than this .05 point and hence does not lie in the region of rejection, the null hypothesis is accepted in this instance. The fluctuations in grades from teacher to teacher is thus apparently due to chance, and there is no basis for attributing it to a difference in teaching. This conclusion contradicts the result of the previous problem, but since it is based upon a larger set of data greater confidence is to be placed in its validity than in the result obtained in the preceding problem.

To test the null hypothesis that the variation in the column means is due to chance—i.e., that the difference in general standing has no effect upon student grades in the given course—the ratio of the second estimate to the last estimate is computed. This gives 13.99, which is far beyond the .05 point of the F distribution for $n_1 = 2$ and $n_2 = 45$.^{*} Hence the null hypothesis must be rejected, and it can be concluded that general standing does have an effect upon the grades obtained in the given course.

The variation in the cell means may be tested in a like manner by comparing the estimate of the population variance based upon this variation with that based upon the remainder variance. In the present problem the ratio is obviously less than 1 so that further analysis is unnecessary. The variation in cell means may be attributed to chance, and there is no basis for concluding that students of different standing react differently to different teachers.¹

TESTS OF CORRELATION COEFFICIENTS AS AN ANALYSIS OF VARIANCE

In Chap. XII certain tests were given to determine whether or not correlation coefficients were significantly different from zero. These tests were in reality an analysis of variance. The variance of points on a line or plane of regression, it will be recalled,² was given by $\sigma_x^2 = \sigma^2 r^2$. Similarly, the variation in the means of rows or columns was given by $\sigma_x^2 = \sigma_1^2 \eta_{12}^2$, and the variation of the points on a curve of regression was given by $\sigma_{x_1}^2 = \sigma_1^2 I_{12}^2$.[†] If there was correlation between the data, these variances might be taken as measures of that correlation. If there was no correlation, however, sample data might still yield values for these variances that would be merely the result of chance. Whether there was correlation or not, variation around the line or plane or curve of regression or around the progression of means could

^{*} Table IX of the Appendix shows that this point is in the neighborhood of 3.

¹ For further discussion of special problems dealing with the analysis of variance, see R. A. Fisher's *Statistical Methods for Research Workers* (1932) and *The Design of Experiments* (1935), also C. H. Gouliden's *Methods of Statistical Analysis* (1939) and Paul R. Rider's *An Introduction to Modern Statistical Methods* (1939).

² See p. 21.

[†] See SMITH, J. G., and A. J. DUNCAN, *Elementary Statistics and Application*, p. 396.

be taken as a measure of the amount of variation caused by chance forces.

Accordingly, to determine whether variation accounted for by a line, plane, curve, or progression of means is due to correlation or is merely the result of chance, it may be compared with the variation around the line, plane, curve, or progression of means. More specifically, the null hypothesis may be set up that all variation is due to chance. Then an estimate of the population variance based upon the variation of the points on the curve, line, or plane or upon the variation of the mean values may be compared with a similar estimate based upon the variation around these measures of correlation. In all cases, the ratio of these two estimates will be distributed like the F distribution.

For a line of regression, for example, the quantity $\frac{\sigma^2 r^2}{1}$ is a maximum-likelihood, or unbiased, estimate of the population variance—assuming no correlation—based on variation along the line of regression; and the quantity $\frac{\sigma^2(1 - r^2)}{N - 2}$ is a maximum-likelihood, or unbiased, estimate of the population variance based on variation around the line of regression. Since these two estimates have independent sampling distributions,¹ their ratio has a sampling distribution of the form of the F distribution, with $n_1 = 1$ and $n_2 = N - 2$. For a plane of regression, the two maximum-likelihood estimates of the population variance are $\frac{\sigma_1^2 R_{1.23}^2 \dots}{k - 1}$ and $\frac{\sigma_1^2(1 - R_{1.23}^2 \dots)}{N - k}$, and their ratio is distributed like the F distribution with $n_1 = k - 1$ and $n_2 = N - k$, where k is the number of regression statistics $a_{1.23}, b_{1.23} \dots$ used in the regression equation. For a progression of means the maximum-likelihood estimates are $\frac{\sigma_1^2 \eta_{12}^2}{k - 1}$ and $\frac{\sigma_1^2(1 - \eta_{12}^2)}{N - k}$, and their ratio is distributed like the F distribution with $n_1 = k - 1$ and $n_2 = N - k$, k being the number of means. Similar formulas apply to a curvilinear regression. The statistics that were used in Chap. XII to test the significance of the multiple correlation coefficient, the correlation ratio, and the correlation index will thus be recognized as the ratio of two maximum-likelihood estimates of the population variance—assuming no correlation—

¹ This cannot be proved here.

and the analysis there given will be recognized as being essentially an analysis of variance.

The test of linearity was also an analysis of variance, but the assumptions are somewhat different. In this case a linear correlation is assumed to exist, but not a curvilinear correlation. On the basis of this assumption an estimate of the population first-order variance is made from the variation of the means—or curve if such is used—from the line of regression and another estimate is made from the variation around the means—or around the curve of regression. These two estimates have independent sampling fluctuations, and their ratio is thus distributed like the F distribution. More specifically,

$$\frac{(\sigma_1^2 \eta_{12}^2 - \sigma_1^2 r_{12}^2)}{k - 2}, \quad \text{or} \quad \frac{(\sigma_1^2 I_{12}^2 - \sigma_1^2 r_{12}^2)}{N - k},$$

is a maximum-likelihood estimate of the population first-order variance based on the variation of the means or curve around the line of regression; and $\frac{\sigma_1^2(1 - \eta_{12}^2)}{N - k}$, or $\frac{\sigma_1^2(1 - I_{12}^2)}{N - k}$, is a maximum-likelihood estimate of the population first-order variance based on the variation around the means or curve, k being the number of means or the number of regression statistics. The ratios of these two estimates, respectively, are distributed like the F distribution with $n_1 = k - 2$ and $n_2 = N - k$. If the reader will turn back to page 307, he will note that this ratio is the statistic that was used to test the linearity of any regression. It should be evident now that this is another form of an analysis of variance.

CHAPTER XVIII

THE PROBLEM OF NONNORMALITY

Up to this point, most of the discussion has been based upon the assumption that the sampled population is normally distributed. This would appear to be a serious restriction, inasmuch as data that are not normally distributed are frequently encountered. It is the purpose of this chapter, therefore, to consider the effect of a departure from normality upon the sampling distributions of some of the more important statistics.

EXACT SAMPLING DISTRIBUTIONS

Exact sampling distributions have been worked out for certain statistics from particular nonnormal populations.¹ The populations studied have been of all sorts, rectangular, triangular, U-shaped populations, populations that conform to a type A Gram-Charlier series, populations that conform to a type III Pearsonian curve, and the like. The distributions derived were mainly for the mean and standard deviation.

Exact Distributions of the Mean. The distribution of the mean for samples of size N from a rectangular population, $y = \frac{1}{a}$, $x = 0$ to $x = a$, has been found to consist of a series of N polynomials, each of degree $N - 1$ and applicable to a subinterval of length a/N . The curve is bell shaped and resembles the normal curve when $N \geq 3$. For $N = 2$, the curve reduces to an isosceles triangle.² It has also been found that the distribution of means from a Pearsonian type III population is a type III curve,³ and

¹ This section is based primarily upon H. L. Rietz, "Topics in Sampling Theory," *Bulletin of the American Mathematical Society*, Vol. 43 (1937), pp. 209-230.

² *Ibid.*, p. 219.

³ Cf. CHURCH, A. E. R., "On the Means and Standard Deviations of Small Samples from any Population," *Biometrika*, Vol. 18 (1926), pp. 321-394, and CRAIG, C. C., "Distribution of Means of Samples from a Type A Population," *Annals of Mathematical Statistics*, Vol. 2 (1931), pp. 99-101.

distributions of means have been worked out for samples from triangular and U-shaped populations.¹

Integrations for distributions of means have been worked out for means of samples from the following populations:²

1. A rectangular population represented by

$$f(x) = \frac{1}{a} \quad (0 \leq x \leq a)$$

2. A J-shaped population represented by the declining exponential curve

$$f(x) = \frac{k}{\delta} \exp \left[-\frac{x}{\delta} \right] \quad (0 \leq x < \infty)$$

3. A positively skewed population represented by the Pearsonian type III or χ^2 type curve

$$f(x) = kx^{-\frac{1}{2}}e^{-\frac{x}{2}} \quad (0 \leq x < \infty)$$

4. A peaked population represented by the double declining exponential curve

$$f(x) = \frac{k}{\delta} \exp \left[\frac{-|x|}{\delta} \right] \quad (-\infty < x < \infty)$$

for $N = 2, 3$, and 4 only.

5. A triangular-shaped population represented by the two straight lines

$$f(x) = \frac{4x}{a^2} \quad \text{when } 0 \leq x \leq \frac{a}{2}$$

and

$$f(x) = \frac{4}{a^2} (a - x) \quad \text{when } \frac{a}{2} \leq x \leq a$$

for $N = 2$ only.

Exact Distributions of Standard Deviation and Variance. For samples of 2 from a rectangular population (represented by $y = 1, x = 0$ to $x = 1$) it has been found³ that the distribution of σ was $f(\sigma) = 4(1 - 2\sigma)$. The exact distribution has also been

¹ Cf. RIDER, PAUL, "On Small Samples from Certain Nonnormal Universes," *Annals of Mathematical Statistics*, Vol. 2 (1931), pp. 48-65.

² See CRAIG, A. T., "On the Distribution of Certain Statistics," *American Journal of Mathematics*, Vol. 54 (1932), pp. 353-366.

³ See RIDER, *op. cit.*

found for the standard deviation of samples of 3 from a rectangular population.¹

Studies of sampling experiments with a Pearsonian type III population suggest that the distribution of sample variances from such a population may be adequately described by a Pearsonian type VI distribution.² Studies have also been made of the moments of the sampling distribution of the variance with a view to obtaining Pearsonian curves to represent the sampling distribution when the population itself is a Pearsonian type curve. By this analysis, light is shed on how the sampling distribution changes with changes in population type and changes in size of sample.³

Exact Distributions of the Statistic $t = \frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$. Studies have been made of the distribution of sample t 's from a rectangular distribution. For samples of 2 the distribution of t was $f(t) = \frac{1}{2(1 + |t|)^2}$; for samples of 3 it was much more complicated. These studies indicated that the distribution of t was more peaked, i.e., had more cases near the middle and at the ends, when the population was rectangular than when it was normal.⁴ It has also been found that the distribution of t from a U-shaped population was similar to that for a rectangular population.⁵

In summarizing a recent account of sampling from nonnormal populations, Rietz concludes as follows: "Although the prospects of obtaining the exact distribution functions of such statistics as the standard deviation s or the 'Student' ratio z for samples from a considerable variety of nonnormal populations do not seem promising, nevertheless, by the use of moments of moments, and experimental sampling, along with the exact determination of

¹ See RIETZ, H. L., "Note on the Distribution of the Standard Deviation of Sets of Three Variates Drawn at Random from a Rectangular Population," *Biometrika*, Vol. 23 (1931), pp. 424-426.

² See SOPHISTER, "Discussion of Small Samples Drawn from an Infinite Skew Population," *Biometrika*, Vol. 20A (1928), pp. 389-423.

³ LE ROUX, J. M., "A Study of the Distribution of the Variance in Small Samples," *Biometrika*, Vol. 23 (1931), pp. 134-190.

⁴ PERLO, V., "On the Distribution of Student's Ratio for Samples of Three Drawn from a Rectangular Distribution," *Biometrika*, Vol. 25 (1933), pp. 203-204.

⁵ See RIDER, *op. cit.*

some distribution functions, significant contributions are being made toward an understanding of the probable nature of certain important features of the distributions in question."¹

The Use of Normal Theory for Nonnormal Populations. In cases in which the population is nonnormal, but the exact forms of the various sampling distributions are not known, it is quite a common practice to view the results obtained on the assumption that the population is normal as fairly good approximations to those that would be actually obtained from the nonnormal population. The extent to which this practice is justified will now be considered.

When the sample is large, as has been demonstrated above the distribution of sample means will tend to be normal in form, with a mean equal to the mean of the population and a variance equal to the variance of the population divided by N .^{*} Whatever the nature of the population the moments of the distribution of sample means are:²

$$\begin{aligned}\bar{X}_{\bar{X}} &= \bar{X} & \bar{X}\bar{u}_2 &= \frac{u_2}{N} & \bar{X}\bar{u}_3 &= \frac{u_3}{N^2} \\ \bar{X}\bar{u}_4 &= \frac{u_4}{N^3} + \frac{3(N-1)}{N^3} (u_2)^2\end{aligned} \quad (1)$$

from which it follows that

$$\bar{X}\beta_1 = \frac{\beta_1}{N} \quad \text{and} \quad \bar{X}\beta_2 - 3 = \frac{\beta_2 - 3}{N} \quad (2)$$

These suggest a fairly rapid approach to normality in general, while experiments in sampling from nonnormal populations that have only a moderate degree of skewness and kurtosis—for example, when $\beta_1 \leq .2$ and $\beta_2 - 3 \leq .4$ —suggest that the sampling distribution of the mean approaches normality very rapidly. This appears to be true, for example, of samples as small as 10.[†] Similar statements can be made for a regression coefficient, the difference between two means, and other statistics that, like the mean, are linear functions of the sample data. It

¹ RIETZ, H. L., "Topics in Sampling Theory," *Bulletin of the American Mathematical Society*, Vol. 43 (1937), p. 230.

^{*} See pp. 262–263.

² $\bar{X}\bar{u}_2$ means u_2 for $\bar{X}$, $\bar{X}\bar{u}_3$ means u_3 for $\bar{X}$, etc. Similarly, $\bar{X}\beta_1$ means β_1 for $\bar{X}$, etc.

[†] Cf. CHURCH, *op. cit.*

is assumed in all these cases that the variance of the population is known.*

The Statistic $t = \frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$. When the variance of the population is not known, it will be recalled that fluctuations in the mean are analyzed, through fluctuations in the statistic $\frac{\sqrt{N}(\bar{X} - \bar{X})}{\bar{\sigma}}$. A number of studies have been made of the sampling fluctuations in this statistic when the population is not normal.¹ When the population was rectangular or U-shaped, it was found that the distribution of t failed to agree with the distribution of t for samples from normal populations, especially near the middle and at the ends of the distribution.² In general it has been found that the agreement between t for nonnormal samples and t for normal samples is poorest when the nonnormal populations are very peaked.³ On the other hand, in the case of skew triangular populations it was found that, although the distribution of t was skewed, when the cumulative probability was considered (i.e., when the two tail areas were added) the results were about the same as for a normal population.⁴ This suggests that in the case of skew populations, at least, the use of the normal theory may not give bad approximations to the correct results. The danger of error apparently is greatest in very peaked populations.

Analysis of Variance. Although the foregoing conclusions are based on experiments relating to sample means and regression coefficients, they apply equally well to analyses of variance in which $n_1 = 1$, the case when the sampling distribution of F , or more precisely $\sqrt{F}$, reduces to the t distribution or half the t distribution because $\sqrt{F}$ may have no sign. That such conclusions can be extended to other analyses of variance requires additional evidence. This has already been pro-

¹ See references for section above, pp. 446-447.

² Also see SHEWHART, W. A., and WINTERS, F. W., "Small Samples—New Experimental Results," *Journal of the American Statistical Association*, Vol. 23 (1938), pp. 144-153.

³ Based on an extensive investigation of Egon S. Pearson. See "The Distribution of Frequency Constants in Small Samples from Nonnormal Symmetrical and Skew Populations," *Biometrika*, Vol. 21 (1929), pp. 259-286.

⁴ See RIDER, *op. cit.*

vided in part by other studies.¹ From a study of empirical data pertaining to a set of means grouped according to one principle of classification, Egon S. Pearson concludes that an analysis based on the assumption of normality gives fairly satisfactory results even when the population is not normal. This, he points out, arises from the correlation that may exist between the variation in the means and the variation about the means when the population is nonnormal.

Variances. General equations have also been derived for the moments and β -coefficients of the sampling distribution of the variance.² These also suggest that the sampling distribution of the variance tends toward the normal form as N increases. The approach to normality is not very rapid, however, and it is not safe to assume that the distribution of the variance is normal unless the sample is quite large. Furthermore, experimental evidence indicates that for smaller samples from certain types of nonnormal populations the sampling distribution of the variance does not necessarily approximate the χ^2 distribution, which is the form it takes when the population is normal.

Accordingly, it has been concluded by students of the subject that "we are liable to go far wrong when using the 'normal theory' variance distribution to represent $D(s^2)$ [i.e., the sampling distribution of the variance] in samples from nonnormal parent populations. . . ."³ There is some doubt also as to the adequacy of the F distribution for testing the ratio of two independent sample variances from nonnormal populations.⁴

¹ PEARSON, EGON S., "Analysis of Variance in Cases of Nonnormal Variation," *Biometrika*, Vol. 23 (1931), pp. 114-133.

² These are conveniently summarized by LeRoux, *op. cit.*

³ *Ibid.*, p. 189. The use of the moments of the sampling distribution of the variance to describe its general form is based on the assumption that it can be represented by a Pearsonian frequency curve. G. A. Baker has attacked the problem from another angle. He has derived a general formula for the sampling distribution of the variance on the assumption that the distribution of the population can be represented by a Gram-Charlier series. This formula is expressed in terms of the constants of the Gram-Charlier series representing the population and the number of terms of the series. Cf. "Note on the Distribution of the Standard Deviations and Second Moments of Samples from a Gram-Charlier Population," *Annals of Mathematical Statistics*, Vol. 2 (1931), pp. 48-65. Little use has yet been made of this approach to the problem.

⁴ Cf. PEARSON, *Biometrika*, Vol. 23 (1931), pp. 114-133.

Correlation Coefficients. The distribution of the correlation coefficient of samples from nonnormal populations has been studied primarily by sampling experiments. When the population correlation coefficient is zero, these experiments suggest that the distribution of the correlation coefficient has approximately the same form as when the population is normal.¹

When the population correlation coefficient is not zero, on the other hand, and when the samples are small, the agreement between the experimental distributions derived from nonnormal populations and the theoretical distribution for a normal population does not appear to be so good, although the closeness of the agreement improves as the size of the sample increases.² These conclusions are based on a very limited set of data and can at the most be considered as tentative only.

INDIRECT ATTACKS ON THE PROBLEM OF NONNORMALITY

Several indirect attacks on the problem of nonnormality have been attempted. One of these is the transformation of the original data. It has long been recognized, for example, that in certain cases the logarithms of given data are normally distributed although the data themselves are not. In such cases, the problem of nonnormality may readily be solved by a logarithmic transformation. A general transformation by which any nonnormal distribution might be made approximately normal has been proposed, and it has been suggested that the sampling distributions of estimates of the parameters of the nonnormal distributions might be expressed in terms of the transformation and the sampling distributions of estimates of the parameters of the normal distribution into which the nonnormal distribution has been transformed.³ This line of attack, however, has not as yet been developed to a point where it can be put to practical use.

Another line of attack has turned its attention to the qualitative or semiqualitative aspects of the data and by thus ignoring

¹ PEARSON, EGON S., "The Test of Significance for the Correlation Coefficient," *Journal of the American Statistical Association*, Vol. 26 (1931), pp. 128-134.

² CHESHIRE, LEONE, ELENA OLDIS, and EGON S. PEARSON, "Further Experiments on the Sampling Distribution of the Correlation Coefficient," *Journal of the American Statistical Association* Vol. 27 (1932), pp. 121-128.

³ Cf. *Annals of Mathematical Statistics*, Vol. 3 (1932), pp. 113-123.

some of the quantitative aspects has avoided the necessity of special assumptions as to the form of the population. Instead of determining, for example, whether two samples are from the same population by testing the difference between their means, variances, etc., it would be possible to group the data into classes and to apply a χ^2 test of independence. A method involving no assumption as to population form would thus be substituted for one necessitating such an assumption. It is to be noted, however, that this qualitative method is not as efficient as the quantitative method, since a certain amount of the information provided by the data is ignored. If the population appears to be normal, therefore, the more refined methods based upon normal assumptions are to be preferred. The qualitative method requires also that sufficient data be available to permit grouping to test for independence.

For large samples a semiquantitative method is available for testing the existence of correlation. If in correlating X_1 and X_2 , for example, only the rank or order of size of the variables is considered, it is possible to compute a coefficient of "rank correlation" whose significance may be tested without any assumption as to the form of the X_1X_2 population. If d_i represents the difference in rank between sample pairs of X_1 and X_2 values, then the coefficient of rank correlation is measured by

$$r' = 1 - \frac{6 \sum d_i^2}{N^3 - N}$$

For a population rank correlation of zero this has a sampling distribution that is approximately normal in form. The mean of this sampling distribution is zero, and its standard deviation is $\sigma_r = 1/\sqrt{N-1}$. The approximation is valid only for large samples. For a discussion of this procedure the student is referred to the work of Hotelling and Pabst.¹

It may be noted, however, that the discarding of certain quantitative information again reduces the efficiency of the test, although it is contended that the loss of efficiency is not very great—not more than, say, 9 per cent. It is also to be noted that this method merely affords a test of the existence of correlation. It does not offer a substitute for the sampling dis-

¹ "Rank Correlation and Tests of Significance Involving No Assumption of Normality," *Annals of Mathematical Statistics*, Vol. 7 (1936), pp. 29-43.

tribution of the Pearsonian correlation coefficient when it is to be assumed that the population correlation coefficient is other than zero.

It has also been suggested that the principle of rank correlation may be used to solve the problem of nonnormality in the analysis of variance.¹ This would involve the replacement of the actual value of a case by its rank in the conventional analysis of variance table. With such a plan, the efficiency of the test is again reduced in comparison with the ordinary analysis of variance applied to a sample from a normal population. This may be offset by the inaccuracy introduced when the usual test is applied to nonnormal populations. A great advantage of the rank method of handling analysis of variance is that it reduces greatly the amount of arithmetical computation involved in the test, since it converts all figures into simple small integers.

TCHÉBYCHEF'S AND OTHER INEQUALITIES

An interesting attack on the problem of nonnormality has been through the establishment of upper limits of probability for specified sampling fluctuations that are valid for almost any type of sampling distribution. The best known of these attempts is Tchébychef's inequality. This says that the probability of a variable deviating from its mean by as much as $\lambda\delta$ is equal to or less than $1/\lambda^2$. For example, if the mean of a variable is 100 and its standard deviation is 10, the probability that the variable will deviate from 100 by as much as 3δ is equal to or less than $\frac{1}{9} = .11$. The basis for this conclusion is as follows. By definition,

$$\sigma^2 = p_1(X_1 - \bar{X})^2 + p_2(X_2 - \bar{X})^2 + \dots + p_N(X_N - \bar{X})^2 \quad (3)$$

Let X' , X'' , X''' , . . . and p' , p'' , p''' , . . . represent those deviations and their probabilities that differ from $\bar{X}$ by as much as $\lambda\delta$. Then, since the part is less than or at the most equal to the whole,

$$\sigma^2 \geq p'(x' - \bar{X})^2 + p''(X'' - \bar{X})^2 + \dots \quad (4)$$

or if $(\lambda\delta)^2$ is put in place of $(X' - \bar{X})^2$, $(X'' - \bar{X})^2$, . . . which are equal to or greater than $(\lambda\delta)^2$,

¹ FRIEDMAN, MILTON, "The Use of Ranks to Avoid the Assumption of Normality," *Journal of the American Statistical Association*, Vol. 32 (1937), pp. 675-701.

$$\begin{aligned}\sigma^2 &\geq p'(\lambda\sigma)^2 + p''(\lambda\sigma)^2 + \dots \\ \sigma^2 &\geq (p' + p'' + \dots)(\lambda\sigma)^2\end{aligned}\quad (5)$$

and

$$P_{\lambda\sigma} \leq \frac{1}{\lambda^2} \quad (6)$$

where $P_{\lambda\sigma} = p' + p'' + \dots$ is the probability of a deviation equal to or greater than $\lambda\sigma$.¹

The problem of closer inequalities has been dealt with in recent papers by several mathematicians. Camp, Guldberg, Meidel, and Narumi have succeeded particularly well by placing certain mild restrictions on the nature of the population function $F(x)$. The restrictions are of such a nature as to leave the distribution function sufficiently general to be useful in the actual problems of statistics. The main restriction placed on $F(x)$ by Camp is that it is to be a monotonic decreasing function of $|x|$ when $|x| \geq c$, $c \geq 0$. The general effect of this restriction is to exclude distributions that are not represented by decreasing functions of $|x|$ at points more than a certain assigned distance from the origin.

With the origin so chosen that zero is at the mean, Camp reaches the general inequality²

$$P_{\lambda\sigma} \leq \frac{\beta_{2s-2}}{\lambda^{2s}} \frac{\left(\frac{2s}{2s+1}\right)^{2s}}{1+\varphi} + \frac{\varphi}{1+\varphi} P_{c\sigma} \quad (7)$$

where

$$\beta_{2s-2} = \frac{\mu_{2s}}{\sigma^{2s}} \quad \text{and} \quad \varphi = \frac{\left(\frac{c}{\lambda} \cdot \frac{2s}{2s+1}\right)^{2s}}{(2s+1)\left(\frac{\lambda}{c} - 1\right)}$$

¹ Cf. RIETZ, HENRY LEWIS, *Mathematical Statistics* (1936), pp. 28-30, 140-144; and ARNE FISHER, *The Mathematical Theory of Probabilities* (1922), pp. 108-109, 115-116. The original work on the subject is in Jules Bienaimé, "Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés," *Comptes Rendus des séances de l'Académie des Sciences*, Vol. 37 (1853), pp. 309-324; P. L. De Tchénychef, "Des Valeurs moyennes." Traduction du Russe par N. de Khanikof, *Journal de mathématiques pures et appliquées*, Deuxième Série, Vol. 12 (1867), pp. 177-184; and P. Pizzetti, "I fondamenti matematici per la critica dei risultati sperimentali," *Atti della Università di Genova* (1892), pp. 113-333.

² RIETZ, *op. cit.*, pp. 140-144.

and P_{∞} is the probability of a deviation equal to or greater than cd . If c is 0, this reduces to

$$P_{\lambda d} \cong \frac{\beta_{2s-2}}{\lambda^{2s}} \left(\frac{2s}{2s+1} \right)^{2s} \quad (8)$$

The difficulty with the use of either Tchébychef's formula or its extensions is that their use usually depends on knowledge of certain population parameters other than those for which hypotheses are being tested. For example, the variance of the sampling distribution of variances from any population equals approximately

$$\frac{\sigma^4}{N} (\beta_2 - 1) \quad (9)$$

From this it is evident that, to test a hypothesis regarding the population standard deviation, knowledge must be had of the population β_2 . This greatly reduces the usefulness of inequalities of this kind, since only limited knowledge regarding β_2 may be available. All that can validly be done in cases like this is to test various joint hypotheses regarding the population variance and the population β_2 .

APPENDIX

TABLE I.—FOUR-PLACE COMMON LOGARITHMS OF NUMBERS¹

	0	1	2	3	4	5	6	7	8	9	10	Tenths of the Tabular Difference 1 2 3 4 5
1.0	0.0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	0414	
1.1	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	0792	
1.2	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	1139	
1.3	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	1461	
1.4	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	1761	
1.5	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	2041	
1.6	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	2304	
1.7	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2553	
1.8	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2788	
1.9	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	3010	
2.0	0.3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	3222	2 4 6 8 11
2.1	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	3424	2 4 6 8 10
2.2	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	3617	2 4 6 8 10
2.3	3617	3636	3655	3673	3692	3711	3729	3747	3766	3784	3802	2 4 5 7 9
2.4	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	3979	2 4 5 7 9
2.5	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	4150	2 3 5 7 9
2.6	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	4314	2 3 5 7 8
2.7	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	4472	2 3 5 6 8
2.8	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	4624	2 3 5 6 8
2.9	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	4771	1 3 4 6 7
3.0	0.4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	4914	1 3 4 6 7
3.1	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	5051	1 3 4 6 7
3.2	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	5185	1 3 4 5 7
3.3	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	5315	1 3 4 5 6
3.4	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	5441	1 3 4 5 6
3.5	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	5563	1 2 4 5 6
3.6	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	5682	1 2 4 5 6
3.7	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	5798	1 2 3 5 6
3.8	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	5911	1 2 3 5 6
3.9	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	6021	1 2 3 4 6
4.0	0.6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	6128	1 2 3 4 5
4.1	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	6232	1 2 3 4 5
4.2	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	6335	1 2 3 4 5
4.3	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	6435	1 2 3 4 5
4.4	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	6532	1 2 3 4 5
4.5	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	6628	1 2 3 4 5
4.6	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	6721	1 2 3 4 5
4.7	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	6812	1 2 3 4 5
4.8	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	6902	1 2 3 4 4
4.9	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	6990	1 2 3 4 4
5.0	0.6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	7076	1 2 3 3 4
5.1	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	7160	1 2 3 3 4
5.2	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	7243	1 2 2 3 4
5.3	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	7324	1 2 2 3 4
5.4	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	7404	1 2 2 3 4

¹ Taken, with permission, from E. V. Huntington's *Four Place Tables of Logarithms and Trigonometric Functions* (Harvard Cooperative Society, Inc., 1907).

TABLE I.—FOUR-PLACE COMMON LOGARITHMS OF NUMBERS.—
(Continued)

	0	1	2	3	4	5	6	7	8	9	10	Tenths of the Tabular Difference 1 2 3 4 5				
5.5	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	7482	1	2	3	4	
5.6	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	7559	1	2	3	4	
5.7	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	7634	1	2	3	4	
5.8	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	7709	1	1	2	3	4
5.9	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	7782	1	1	2	3	4
6.0	0.7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	7853	1	1	2	3	4
6.1	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	7924	1	1	2	3	4
6.2	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	7993	1	1	2	3	3
6.3	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	8062	1	1	2	3	3
6.4	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	8129	1	1	2	3	3
6.5	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	8195	1	1	2	3	3
6.6	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	8261	1	1	2	3	3
6.7	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	8325	1	1	2	3	3
6.8	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	8388	1	1	2	3	3
6.9	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	8451	1	1	2	3	3
7.0	0.8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	8513	1	1	2	2	3
7.1	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	8573	1	1	2	2	3
7.2	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	8633	1	1	2	2	3
7.3	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	8692	1	1	2	2	3
7.4	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	8751	1	1	2	2	3
7.5	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	8808	1	1	2	2	3
7.6	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	8865	1	1	2	2	3
7.7	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	8921	1	1	2	2	3
7.8	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	8976	1	1	2	2	3
7.9	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	9031	1	1	2	2	3
8.0	0.9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	9085	1	1	2	2	3
8.1	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	9138	1	1	2	2	3
8.2	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	9191	1	1	2	2	3
8.3	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	9243	1	1	2	2	3
8.4	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	9294	1	1	2	2	3
8.5	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	9345	1	1	2	2	3
8.6	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	9395	1	1	2	2	3
8.7	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	9445	0	1	1	2	2
8.8	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	9494	0	1	1	2	2
8.9	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	9542	0	1	1	2	2
9.0	0.9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	9590	0	1	1	2	2
9.1	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	9638	0	1	1	2	2
9.2	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	9685	0	1	1	2	2
9.3	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	9731	0	1	1	2	2
9.4	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	9777	0	1	1	2	2
9.5	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	9823	0	1	1	2	2
9.6	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	9868	0	1	1	2	2
9.7	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	9912	0	1	1	2	2
9.8	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	9956	0	1	1	2	2
9.9	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996		0	1	1	2	2

TABLE I.—FOUR-PLACE COMMON LOGARITHMS OF NUMBERS.—
(Continued)

	0	1	2	3	4	5	6	7	8	9	10
1.00	0.0000	0004	0009	0013	0017	0022	0026	0030	0035	0039	0043
1.01	0043	0048	0052	0056	0060	0065	0069	0073	0077	0082	0086
1.02	0086	0090	0095	0099	0103	0107	0111	0116	0120	0124	0128
1.03	0128	0133	0137	0141	0145	0149	0154	0158	0162	0166	0170
1.04	0170	0175	0179	0183	0187	0191	0195	0199	0204	0208	0212
1.05	0212	0216	0220	0224	0228	0233	0237	0241	0245	0249	0253
1.06	0253	0257	0261	0265	0269	0273	0278	0282	0286	0290	0294
1.07	0294	0298	0302	0306	0310	0314	0318	0322	0326	0330	0334
1.08	0334	0338	0342	0346	0350	0354	0358	0362	0366	0370	0374
1.09	0374	0378	0382	0386	0390	0394	0398	0402	0406	0410	0414
1.10	0.0414	0418	0422	0426	0430	0434	0438	0441	0445	0449	0453
1.11	0453	0457	0461	0465	0469	0473	0477	0481	0484	0488	0492
1.12	0492	0496	0500	0504	0508	0512	0515	0519	0523	0527	0531
1.13	0531	0535	0538	0542	0546	0550	0554	0558	0561	0565	0569
1.14	0569	0573	0577	0580	0584	0588	0592	0596	0599	0603	0607
1.15	0607	0611	0615	0618	0622	0626	0630	0633	0637	0641	0645
1.16	0645	0648	0652	0656	0660	0663	0667	0671	0674	0678	0682
1.17	0682	0686	0689	0693	0697	0700	0704	0708	0711	0715	0719
1.18	0719	0722	0726	0730	0734	0737	0741	0745	0748	0752	0755
1.19	0755	0759	0763	0766	0770	0774	0777	0781	0785	0788	0792
1.20	0.0792	0795	0799	0803	0806	0810	0813	0817	0821	0824	0828
1.21	0828	0831	0835	0839	0842	0846	0849	0853	0856	0860	0864
1.22	0864	0867	0871	0874	0878	0881	0885	0888	0892	0896	0899
1.23	0899	0903	0906	0910	0913	0917	0920	0924	0927	0931	0934
1.24	0934	0938	0941	0945	0948	0952	0955	0959	0962	0966	0969
1.25	0969	0973	0976	0980	0983	0986	0990	0993	0997	1000	1004
1.26	1004	1007	1011	1014	1017	1021	1024	1028	1031	1035	1038
1.27	1038	1041	1045	1048	1052	1055	1059	1062	1065	1069	1072
1.28	1072	1075	1079	1082	1086	1089	1092	1096	1099	1103	1106
1.29	1106	1109	1113	1116	1119	1123	1126	1129	1133	1136	1139
1.30	0.1139	1143	1146	1149	1153	1156	1159	1163	1166	1169	1173
1.31	1173	1176	1179	1183	1186	1189	1193	1196	1199	1202	1206
1.32	1206	1209	1212	1216	1219	1222	1225	1229	1232	1235	1239
1.33	1239	1242	1245	1248	1252	1255	1258	1261	1265	1268	1271
1.34	1271	1274	1278	1281	1284	1287	1290	1294	1297	1300	1303
1.35	1303	1307	1310	1313	1316	1319	1323	1326	1329	1332	1335
1.36	1335	1339	1342	1345	1348	1351	1355	1358	1361	1364	1367
1.37	1367	1370	1374	1377	1380	1383	1386	1389	1392	1396	1399
1.38	1399	1402	1405	1408	1411	1414	1418	1421	1424	1427	1430
1.39	1430	1433	1436	1440	1443	1446	1449	1452	1455	1458	1461
1.40	0.1461	1464	1467	1471	1474	1477	1480	1483	1486	1489	1492
1.41	1492	1495	1498	1501	1504	1508	1511	1514	1517	1520	1523
1.42	1523	1526	1529	1532	1535	1538	1541	1544	1547	1550	1553
1.43	1553	1556	1559	1562	1565	1569	1572	1575	1578	1581	1584
1.44	1584	1587	1590	1593	1596	1599	1602	1605	1608	1611	1614
1.45	1614	1617	1620	1623	1626	1629	1632	1635	1638	1641	1644
1.46	1644	1647	1649	1652	1655	1658	1661	1664	1667	1670	1673
1.47	1673	1676	1679	1682	1685	1688	1691	1694	1697	1700	1703
1.48	1703	1706	1708	1711	1714	1717	1720	1723	1726	1729	1732
1.49	1732	1735	1738	1741	1744	1746	1749	1752	1755	1758	1761

TABLE I.—FOUR-PLACE COMMON LOGARITHMS OF NUMBERS.—
(Continued)

	0	1	2	3	4	5	6	7	8	9	10
1.50	0.1761	1764	1767	1770	1772	1775	1778	1781	1784	1787	1790
1.51	1790	1793	1796	1798	1801	1804	1807	1810	1813	1816	1818
1.52	1818	1821	1824	1827	1830	1833	1836	1838	1841	1844	1847
1.53	1847	1850	1853	1855	1858	1861	1864	1867	1870	1872	1875
1.54	1875	1878	1881	1884	1886	1889	1892	1895	1898	1901	1903
1.55	1903	1906	1909	1912	1915	1917	1920	1923	1926	1928	1931
1.56	1931	1934	1937	1940	1942	1945	1948	1951	1953	1956	1959
1.57	1959	1962	1965	1967	1970	1973	1976	1978	1981	1984	1987
1.58	1987	1989	1992	1995	1998	2000	2003	2006	2009	2011	2014
1.59	2014	2017	2019	2022	2025	2028	2030	2033	2036	2038	2041
1.60	0.2041	2044	2047	2049	2052	2055	2057	2060	2063	2066	2068
1.61	2068	2071	2074	2076	2079	2082	2084	2087	2090	2092	2095
1.62	2095	2098	2101	2103	2106	2109	2111	2114	2117	2119	2122
1.63	2122	2125	2127	2130	2133	2135	2138	2140	2143	2146	2148
1.64	2148	2151	2154	2156	2159	2162	2164	2167	2170	2172	2175
1.65	2175	2177	2180	2183	2185	2188	2191	2193	2196	2198	2201
1.66	2201	2204	2206	2209	2212	2214	2217	2219	2222	2225	2227
1.67	2227	2230	2232	2235	2238	2240	2243	2245	2248	2251	2253
1.68	2253	2256	2258	2261	2263	2266	2269	2271	2274	2276	2279
1.69	2279	2281	2284	2287	2289	2292	2294	2297	2299	2302	2304
1.70	0.2304	2307	2310	2312	2315	2317	2320	2322	2325	2327	2330
1.71	2330	2333	2335	2338	2340	2343	2345	2348	2350	2353	2355
1.72	2355	2358	2360	2363	2365	2368	2370	2373	2375	2378	2380
1.73	2380	2383	2385	2388	2390	2393	2395	2398	2400	2403	2405
1.74	2405	2408	2410	2413	2415	2418	2420	2423	2425	2428	2430
1.75	2430	2433	2435	2438	2440	2443	2445	2448	2450	2453	2455
1.76	2455	2458	2460	2463	2465	2467	2470	2472	2475	2477	2480
1.77	2480	2482	2485	2487	2490	2492	2494	2497	2499	2502	2504
1.78	2504	2507	2509	2512	2514	2516	2519	2521	2524	2526	2529
1.79	2529	2531	2533	2536	2538	2541	2543	2545	2548	2550	2553
1.80	0.2553	2555	2558	2560	2562	2565	2567	2570	2572	2574	2577
1.81	2577	2579	2582	2584	2586	2589	2591	2594	2596	2598	2601
1.82	2601	2603	2605	2608	2610	2613	2615	2617	2620	2622	2625
1.83	2625	2627	2629	2632	2634	2636	2639	2641	2643	2646	2648
1.84	2648	2651	2653	2655	2658	2660	2662	2665	2667	2669	2672
1.85	2672	2674	2676	2679	2681	2683	2686	2688	2690	2693	2695
1.86	2695	2697	2700	2702	2704	2707	2709	2711	2714	2716	2718
1.87	2718	2721	2723	2725	2728	2730	2732	2735	2737	2739	2742
1.88	2742	2744	2746	2749	2751	2753	2755	2758	2760	2762	2765
1.89	2765	2767	2769	2772	2774	2776	2778	2781	2783	2785	2788
1.90	0.2788	2790	2792	2794	2797	2799	2801	2804	2806	2808	2810
1.91	2810	2813	2815	2817	2819	2822	2824	2826	2828	2831	2833
1.92	2833	2835	2838	2840	2842	2844	2847	2849	2851	2853	2856
1.93	2856	2858	2860	2862	2865	2867	2869	2871	2874	2876	2878
1.94	2878	2880	2882	2885	2887	2889	2891	2894	2896	2898	2900
1.95	2900	2903	2905	2907	2909	2911	2914	2916	2918	2920	2923
1.96	2923	2925	2927	2929	2931	2934	2936	2938	2940	2942	2945
1.97	2945	2947	2949	2951	2953	2956	2958	2960	2962	2964	2967
1.98	2967	2969	2971	2973	2975	2978	2980	2982	2984	2986	2989
1.99	2989	2991	2993	2995	2997	2999	3002	3004	3006	3008	3010

TABLE II.—SQUARES OF NUMBERS¹

N	0	1	2	3	4	5	6	7	8	9
100	10000	10201	10404	10609	10816	11025	11236	11449	11664	11881
110	12100	12321	12544	12769	12996	13225	13456	13689	13924	14161
120	14400	14641	14884	15129	15376	15625	15876	16129	16384	16641
130	16900	17161	17424	17689	17956	18225	18496	18769	19044	19321
140	19600	19881	20164	20449	20736	21025	21316	21609	21904	22201
150	22500	22801	23104	23409	23716	24025	24336	24649	24964	25281
160	25600	25921	26244	26569	26896	27225	27556	27889	28224	28561
170	28900	29241	29584	29929	30276	30625	30976	31329	31684	32041
180	32400	32761	33124	33489	33856	34225	34596	34969	35344	35721
190	36100	36481	36864	37249	37636	38025	38416	38809	39204	39601
200	40000	40401	40804	41209	41616	42025	42436	42849	43264	43681
210	44100	44521	44944	45369	45796	46225	46656	47089	47524	47961
220	48400	48841	49284	49729	50176	50625	51076	51529	51984	52441
230	52900	53361	53824	54289	54756	55225	55696	56169	56644	57121
240	57600	58081	58564	59049	59536	60025	60516	61009	61504	62001
250	62500	63001	63504	64009	64516	65025	65536	66049	66564	67081
260	67600	68121	68644	69169	69696	70225	70756	71289	71824	72361
270	72900	73441	73984	74529	75076	75625	76176	76729	77284	77841
280	78400	78961	79524	80089	80656	81225	81796	82369	82944	83521
290	84100	84681	85262	85849	86436	87025	87616	88209	88804	89401
300	90000	90601	91204	91809	92416	93025	93636	94249	94864	95481
310	96100	96721	97344	97969	98596	99225	99856	100489	101124	101761
320	102400	103041	103684	104329	104976	105625	106276	106929	107584	108241
330	108900	109561	110224	110889	111556	112225	112896	113569	114244	114921
340	115600	116281	116964	117649	118336	119025	119716	120409	121104	121801
350	122500	123201	123904	124609	125316	126025	126736	127449	128164	128881
360	129600	130321	131044	131769	132496	133225	133956	134689	135424	136161
370	136900	137641	138384	139129	139876	140625	141376	142129	142884	143641
380	144400	145161	145924	146689	147456	148225	148996	149769	150544	151321
390	152100	152881	153664	154449	155236	156025	156816	157609	158404	159201
400	160000	160801	161604	162409	163216	164025	164836	165649	166464	167281
410	168100	168921	169744	170569	171396	172225	173056	173889	174724	175561
420	176400	177241	178084	178929	179776	180625	181476	182329	183184	184041
430	184900	185761	186624	187489	188356	189225	190096	190969	191844	192721
440	193600	194481	195364	196249	197136	198025	198916	199809	200704	201610
450	202500	203401	204304	205209	206116	207025	207936	208849	209764	210681
460	211600	212521	213444	214369	215296	216225	217156	218089	219024	219961
470	220900	221841	222784	223729	224676	225625	226576	227529	228484	229441
480	230400	231361	232324	233289	234256	235225	236196	237169	238144	239121
490	240100	241081	242064	243049	244036	245025	246016	247009	248004	249001
500	250000	251001	252004	253009	254016	255025	256036	257049	258064	259081
510	260100	261121	262144	263169	264196	265225	266256	267289	268324	269361
520	270400	271441	272484	273529	274576	275625	276676	277729	278784	279841
530	280900	281961	283024	284089	285156	286225	287296	288369	289444	290521
540	291600	292681	293764	294849	295936	297025	298116	299209	300304	301401

¹ Source: WAUGH, ALBERT E., *Laboratory Manual and Problems for Elements of Statistical Method* (McGraw-Hill Book Company, Inc., 1944).

TABLE II.—SQUARES OF NUMBERS.—(Continued)

N.	0	1	2	3	4	5	6	7	8	9
550	302500	303601	304704	305809	306916	308025	309136	310249	311364	312481
560	313600	314721	315844	316969	318096	319225	320356	321489	322624	323761
570	324900	326041	327184	328329	329476	330625	331776	332929	334084	335241
580	336400	337561	338724	339889	341056	342225	343396	344569	345744	346921
590	348100	349281	350464	351649	352836	354025	355216	356409	357604	358801
600	360000	361201	362404	363609	364816	366025	367236	368449	369664	370881
610	372100	373321	374544	375769	376996	378225	379456	380689	381924	383161
620	384400	385641	386884	388129	389376	390625	391876	393129	394384	395641
630	396900	398161	399424	400689	401956	403225	404496	405769	407044	408321
640	409600	410881	412164	413449	414736	416025	417316	418609	419904	421201
650	422500	423801	425104	426409	427716	429025	430336	431649	432964	434281
660	435600	436921	438244	439569	440896	442225	443556	444889	446224	447561
670	448900	450241	451584	452929	454276	455625	456976	458329	459684	461041
680	462400	463761	465124	466489	467856	469225	470596	471969	473344	474721
690	476100	477481	478864	480249	481636	483025	484416	485809	487204	488601
700	490000	491401	492804	494209	495616	497025	498436	498849	501264	502681
710	504100	505521	506944	508369	509796	511225	512656	514089	515524	516961
720	518400	519841	521284	522729	524176	525625	527076	528529	529984	531441
730	532900	534361	535824	537289	538756	540225	541696	543169	544644	546121
740	547600	549081	550564	552049	553536	555025	556516	558009	559504	561001
750	562500	564001	565504	567009	568516	570025	571536	573049	574564	576081
760	577600	579121	580644	582169	583696	585225	586756	588289	589824	591361
770	592900	594441	595984	597529	599076	600625	602176	603729	605284	606841
780	608400	609961	611524	613089	614656	616225	617796	619369	620944	622521
790	624100	625681	627264	628849	630436	632025	633616	635209	636804	638401
800	640000	641601	643204	644809	646416	648025	649636	651249	652864	654481
810	656100	657721	659344	660969	662596	664225	665856	667489	669124	670761
820	672400	674041	675684	677329	678976	680625	682276	683929	685584	687241
830	688900	690561	692224	693889	695556	697225	698896	700569	702244	703921
840	705600	707281	708964	710649	712336	714025	715716	717409	719104	720801
850	722500	724201	725904	727609	729316	731025	732736	734449	736164	737881
860	739600	741321	743044	744769	746496	748225	749956	751689	753424	755161
870	756900	758641	760384	762129	763876	765625	767376	769129	770884	772641
880	774400	776161	777924	779689	781456	783225	784996	786769	788544	790321
890	792100	793881	795664	797449	799236	801025	802816	804609	806404	808201
900	810000	811801	813604	815409	817216	819025	820836	822649	824464	826281
910	828100	829921	831744	833569	835396	837225	839056	840889	842724	844561
920	846400	848241	850084	851929	853776	855625	857476	859329	861184	863041
930	864900	866761	868624	870489	872356	874225	876096	877969	879844	881721
940	883600	885481	887364	889249	891136	893025	894916	896809	898704	900601
950	902500	904401	906304	908209	910116	912025	913936	915849	917764	919681
960	921600	923521	925444	927369	929296	931225	933156	935089	937024	938961
970	940900	942841	944784	946729	948676	950625	952576	954529	956484	958441
980	960400	962361	964324	966289	968256	970225	972196	974169	976144	978121
990	980100	982081	984064	986049	988036	990025	992016	994009	996004	998001

TABLE III.—SQUARE ROOTS OF NUMBERS FROM 10 TO 100¹

N	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	3.162	3.178	3.194	3.209	3.225	3.240	3.256	3.271	3.286	3.302
11	3.317	3.332	3.347	3.362	3.376	3.391	3.406	3.421	3.435	3.450
12	3.464	3.479	3.493	3.507	3.521	3.536	3.550	3.564	3.578	3.592
13	3.606	3.619	3.633	3.647	3.661	3.674	3.688	3.701	3.715	3.728
14	3.742	3.755	3.768	3.782	3.795	3.808	3.821	3.834	3.847	3.860
15	3.873	3.886	3.899	3.912	3.924	3.937	3.950	3.962	3.975	3.987
16	4.000	4.012	4.025	4.037	4.050	4.062	4.074	4.087	4.099	4.111
17	4.123	4.135	4.147	4.159	4.171	4.183	4.195	4.207	4.219	4.231
18	4.243	4.254	4.266	4.278	4.290	4.301	4.313	4.324	4.336	4.347
19	4.359	4.370	4.382	4.393	4.405	4.416	4.427	4.438	4.450	4.461
20	4.472	4.483	4.494	4.506	4.517	4.528	4.539	4.550	4.561	4.572
21	4.583	4.593	4.604	4.615	4.626	4.637	4.648	4.658	4.669	4.680
22	4.690	4.701	4.712	4.722	4.733	4.743	4.754	4.764	4.775	4.785
23	4.796	4.806	4.817	4.827	4.837	4.848	4.858	4.868	4.879	4.889
24	4.899	4.909	4.919	4.930	4.940	4.950	4.960	4.970	4.980	4.990
25	5.000	5.010	5.020	5.030	5.040	5.050	5.060	5.070	5.079	5.089
26	5.099	5.109	5.119	5.128	5.138	5.148	5.158	5.167	5.177	5.187
27	5.196	5.206	5.215	5.225	5.234	5.244	5.254	5.263	5.273	5.282
28	5.292	5.301	5.310	5.320	5.329	5.339	5.348	5.357	5.367	5.376
29	5.385	5.394	5.404	5.413	5.422	5.431	5.441	5.450	5.459	5.468
30	5.477	5.486	5.495	5.505	5.514	5.523	5.532	5.541	5.550	5.559
31	5.568	5.577	5.586	5.595	5.604	5.612	5.621	5.630	5.639	5.648
32	5.657	5.666	5.674	5.683	5.692	5.701	5.710	5.718	5.727	5.736
33	5.745	5.753	5.762	5.771	5.779	5.788	5.797	5.805	5.814	5.822
34	5.831	5.840	5.848	5.857	5.865	5.874	5.882	5.891	5.899	5.908
35	5.916	5.925	5.933	5.941	5.950	5.958	5.967	5.975	5.983	5.992
36	6.000	6.008	6.017	6.025	6.033	6.042	6.050	6.058	6.066	6.075
37	6.083	6.091	6.099	6.107	6.116	6.124	6.132	6.140	6.148	6.156
38	6.164	6.173	6.181	6.189	6.197	6.205	6.213	6.221	6.229	6.237
39	6.245	6.253	6.261	6.269	6.277	6.285	6.293	6.301	6.309	6.317
40	6.325	6.332	6.340	6.348	6.356	6.364	6.372	6.380	6.387	6.395
41	6.403	6.411	6.419	6.427	6.434	6.442	6.450	6.458	6.465	6.473
42	6.481	6.488	6.496	6.504	6.512	6.519	6.527	6.535	6.542	6.550
43	6.557	6.565	6.573	6.580	6.588	6.595	6.603	6.611	6.618	6.626
44	6.633	6.641	6.648	6.656	6.663	6.671	6.678	6.686	6.693	6.701
45	6.708	6.716	6.723	6.731	6.738	6.745	6.753	6.760	6.768	6.775
46	6.782	6.790	6.797	6.804	6.812	6.819	6.826	6.834	6.841	6.848
47	6.856	6.863	6.870	6.878	6.885	6.892	6.899	6.907	6.914	6.921
48	6.928	6.935	6.943	6.950	6.957	6.964	6.971	6.979	6.986	6.993
49	7.000	7.007	7.014	7.021	7.029	7.036	7.043	7.050	7.057	7.064
50	7.071	7.078	7.085	7.092	7.099	7.106	7.113	7.120	7.127	7.134
51	7.141	7.148	7.155	7.162	7.169	7.176	7.183	7.190	7.197	7.204
52	7.211	7.218	7.225	7.232	7.239	7.246	7.253	7.259	7.266	7.273
53	7.280	7.287	7.294	7.301	7.308	7.314	7.321	7.328	7.335	7.342
54	7.348	7.355	7.362	7.369	7.376	7.382	7.389	7.396	7.403	7.409

¹ Source: WAUGH, ALBERT E., *Laboratory Manual and Problems for Elements of Statistical Method* (McGraw-Hill Book Company, Inc., 1944):

TABLE III.—SQUARE ROOTS OF NUMBERS FROM 10 TO 100.—(Continued)

N	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
55	7.416	7.423	7.430	7.436	7.443	7.450	7.457	7.463	7.470	7.477
56	7.483	7.490	7.497	7.503	7.510	7.517	7.523	7.530	7.537	7.543
57	7.550	7.556	7.563	7.570	7.576	7.582	7.589	7.596	7.603	7.609
58	7.616	7.622	7.629	7.635	7.642	7.649	7.655	7.662	7.668	7.675
59	7.681	7.688	7.694	7.701	7.707	7.714	7.720	7.727	7.733	7.740
60	7.746	7.752	7.759	7.765	7.772	7.778	7.785	7.791	7.797	7.804
61	7.810	7.817	7.823	7.829	7.836	7.842	7.849	7.855	7.861	7.868
62	7.874	7.880	7.887	7.893	7.899	7.906	7.912	7.918	7.925	7.931
63	7.937	7.944	7.950	7.956	7.962	7.969	7.975	7.981	7.987	7.994
64	8.000	8.006	8.012	8.019	8.025	8.031	8.037	8.044	8.050	8.056
65	8.062	8.068	8.075	8.081	8.087	8.093	8.099	8.106	8.112	8.118
66	8.124	8.130	8.136	8.142	8.149	8.155	8.161	8.167	8.173	8.179
67	8.185	8.191	8.198	8.204	8.210	8.216	8.222	8.228	8.234	8.240
68	8.246	8.252	8.258	8.264	8.270	8.276	8.283	8.289	8.295	8.301
69	8.307	8.313	8.319	8.325	8.331	8.337	8.343	8.349	8.355	8.361
70	8.367	8.373	8.379	8.385	8.390	8.396	8.402	8.408	8.414	8.420
71	8.426	8.432	8.438	8.444	8.450	8.456	8.462	8.468	8.473	8.479
72	8.485	8.491	8.497	8.503	8.509	8.515	8.521	8.526	8.532	8.538
73	8.544	8.550	8.556	8.562	8.567	8.573	8.579	8.585	8.591	8.597
74	8.602	8.608	8.614	8.620	8.626	8.631	8.637	8.643	8.649	8.654
75	8.660	8.666	8.672	8.678	8.683	8.689	8.695	8.701	8.706	8.712
76	8.718	8.724	8.730	8.735	8.741	8.746	8.752	8.758	8.764	8.769
77	8.775	8.781	8.786	8.792	8.798	8.803	8.809	8.815	8.820	8.826
78	8.832	8.837	8.843	8.849	8.854	8.860	8.866	8.871	8.877	8.883
79	8.888	8.894	8.899	8.905	8.911	8.916	8.922	8.927	8.933	8.939
80	8.944	8.950	8.955	8.961	8.967	8.972	8.978	8.983	8.989	8.994
81	9.000	9.006	9.011	9.017	9.022	9.028	9.033	9.039	9.044	9.050
82	9.055	9.061	9.066	9.072	9.077	9.083	9.088	9.094	9.099	9.105
83	9.110	9.116	9.121	9.127	9.132	9.138	9.143	9.149	9.154	9.160
84	9.165	9.171	9.176	9.182	9.187	9.192	9.198	9.203	9.209	9.214
85	9.220	9.225	9.230	9.236	9.241	9.247	9.252	9.257	9.263	9.268
86	9.274	9.279	9.284	9.290	9.295	9.301	9.306	9.311	9.317	9.322
87	9.327	9.333	9.338	9.343	9.349	9.354	9.359	9.365	9.370	9.376
88	9.381	9.386	9.391	9.397	9.402	9.407	9.413	9.418	9.423	9.429
89	9.434	9.439	9.445	9.450	9.455	9.460	9.466	9.471	9.476	9.482
90	9.487	9.492	9.497	9.503	9.508	9.513	9.518	9.524	9.529	9.534
91	9.539	9.545	9.550	9.555	9.560	9.566	9.571	9.576	9.581	9.586
92	9.592	9.597	9.602	9.607	9.612	9.618	9.623	9.628	9.633	9.638
93	9.644	9.649	9.654	9.659	9.664	9.670	9.675	9.680	9.685	9.690
94	9.695	9.701	9.706	9.711	9.716	9.721	9.726	9.731	9.737	9.742
95	9.747	9.752	9.757	9.762	9.767	9.772	9.778	9.783	9.788	9.793
96	9.798	9.803	9.808	9.813	9.818	9.823	9.829	9.834	9.839	9.844
97	9.849	9.854	9.859	9.864	9.869	9.874	9.879	9.884	9.889	9.894
98	9.899	9.905	9.910	9.915	9.920	9.925	9.930	9.935	9.940	9.945
99	9.950	9.955	9.960	9.965	9.970	9.975	9.980	9.985	9.990	9.995

TABLE IV.—SQUARE ROOTS OF NUMBERS FROM 100 TO 1000¹

N	0	1	2	3	4	5	6	7	8	9
100	10.00	10.05	10.10	10.15	10.20	10.25	10.30	10.34	10.39	10.44
110	10.49	10.54	10.58	10.63	10.68	10.72	10.77	10.82	10.86	10.91
120	10.95	11.00	11.05	11.09	11.14	11.18	11.22	11.27	11.31	11.36
130	11.40	11.45	11.49	11.53	11.58	11.62	11.66	11.70	11.75	11.79
140	11.83	11.87	11.92	11.96	12.00	12.04	12.08	12.12	12.17	12.21
150	12.25	12.29	12.33	12.37	12.41	12.45	12.49	12.53	12.57	12.61
160	12.65	12.69	12.73	12.77	12.81	12.85	12.88	12.92	12.96	13.00
170	13.04	13.08	13.11	13.15	13.19	13.23	13.27	13.30	13.34	13.38
180	13.42	13.45	13.49	13.53	13.56	13.60	13.64	13.67	13.71	13.75
190	13.78	13.82	13.86	13.89	13.93	13.96	14.00	14.04	14.07	14.11
200	14.14	14.18	14.21	14.25	14.28	14.32	14.35	14.39	14.42	14.46
210	14.49	14.53	14.56	14.59	14.63	14.66	14.70	14.73	14.76	14.80
220	14.83	14.87	14.90	14.93	14.97	15.00	15.03	15.07	15.10	15.13
230	15.17	15.20	15.23	15.26	15.30	15.33	15.36	15.39	15.43	15.46
240	15.49	15.52	15.56	15.59	15.62	15.65	15.68	15.72	15.75	15.78
250	15.81	15.84	15.87	15.91	15.94	15.97	16.00	16.03	16.06	16.09
260	16.12	16.16	16.19	16.22	16.25	16.28	16.31	16.34	16.37	16.40
270	16.43	16.46	16.49	16.52	16.55	16.58	16.61	16.64	16.67	16.70
280	16.73	16.76	16.79	16.82	16.85	16.88	16.91	16.94	16.97	17.00
290	17.03	17.06	17.09	17.12	17.15	17.18	17.20	17.23	17.26	17.29
300	17.32	17.35	17.38	17.41	17.44	17.46	17.49	17.52	17.55	17.58
310	17.61	17.64	17.66	17.69	17.72	17.75	17.78	17.80	17.83	17.86
320	17.89	17.92	17.94	17.97	18.00	18.03	18.06	18.08	18.11	18.14
330	18.17	18.19	18.22	18.25	18.28	18.30	18.33	18.36	18.38	18.41
340	18.44	18.47	18.49	18.52	18.55	18.57	18.60	18.63	18.65	18.68
350	18.71	18.74	18.76	18.79	18.81	18.84	18.87	18.89	18.92	18.95
360	18.97	19.00	19.03	19.05	19.08	19.10	19.13	19.16	19.18	19.21
370	19.24	19.26	19.29	19.31	19.34	19.36	19.39	19.42	19.44	19.47
380	19.49	19.52	19.54	19.57	19.60	19.62	19.65	19.67	19.70	19.72
390	19.75	19.77	19.80	19.82	19.85	19.87	19.90	19.92	19.95	19.98
400	20.00	20.02	20.05	20.07	20.10	20.12	20.15	20.17	20.20	20.22
410	20.25	20.27	20.30	20.32	20.35	20.37	20.40	20.42	20.44	20.47
420	20.49	20.52	20.54	20.57	20.59	20.62	20.64	20.66	20.69	20.71
430	20.74	20.76	20.78	20.81	20.83	20.86	20.88	20.90	20.93	20.95
440	20.98	21.00	21.02	21.05	21.07	21.10	21.12	21.14	21.17	21.19
450	21.21	21.24	21.26	21.28	21.31	21.33	21.35	21.38	21.40	21.42
460	21.45	21.47	21.49	21.52	21.54	21.56	21.59	21.61	21.63	21.66
470	21.68	21.70	21.73	21.75	21.77	21.79	21.82	21.84	21.86	21.89
480	21.91	21.93	21.95	21.98	22.00	22.02	22.05	22.07	22.09	22.11
490	22.14	22.16	22.18	22.20	22.23	22.25	22.27	22.29	22.32	22.34
500	22.36	22.38	22.41	22.43	22.45	22.47	22.49	22.52	22.54	22.56
510	22.58	22.61	22.63	22.65	22.67	22.69	22.72	22.74	22.76	22.78
520	22.80	22.83	22.85	22.87	22.89	22.91	22.93	22.96	22.98	23.00
530	23.02	23.04	23.07	23.09	23.11	23.13	23.15	23.17	23.19	23.22
540	23.24	23.26	23.28	23.30	23.32	23.35	23.37	23.39	23.41	23.43
550	23.45	23.47	23.49	23.52	23.54	23.56	23.58	23.60	23.62	23.64

¹ Source: WAUGH, ALBERT E., *Laboratory Manual and Problems for Elements of Statistical Method* (McGraw-Hill Book Company, Inc., 1944).

TABLE IV.—SQUARE ROOTS OF NUMBERS FROM 100 TO 1000.—(Continued)

N	0	1	2	3	4	5	6	7	8	9
550	23.45	23.47	23.49	23.52	23.54	23.56	23.58	23.60	23.62	23.64
560	23.66	23.69	23.71	23.73	23.75	23.77	23.79	23.81	23.83	23.85
570	23.87	23.90	23.92	23.94	23.96	23.98	24.00	24.02	24.04	24.06
580	24.08	24.10	24.12	24.15	24.17	24.19	24.21	24.23	24.25	24.27
590	24.29	24.31	24.33	24.35	24.37	24.39	24.41	24.43	24.45	24.47
600	24.49	24.52	24.54	24.56	24.58	24.60	24.62	24.64	24.66	24.68
610	24.70	24.72	24.74	24.76	24.78	24.80	24.82	24.84	24.86	24.88
620	24.90	24.92	24.94	24.96	24.98	25.00	25.02	25.04	25.06	25.08
630	25.10	25.12	25.14	25.16	25.18	25.20	25.22	25.24	25.26	25.28
640	25.30	25.32	25.34	25.36	25.38	25.40	25.42	25.44	25.46	25.48
650	25.50	25.51	25.53	25.55	25.57	25.59	25.61	25.63	25.65	25.67
660	25.69	25.71	25.73	25.75	25.77	25.79	25.81	25.83	25.85	25.86
670	25.88	25.90	25.92	25.94	25.96	25.98	26.00	26.02	26.04	26.06
680	26.08	26.10	26.12	26.13	26.15	26.17	26.19	26.21	26.23	26.25
690	26.27	26.29	26.31	26.32	26.34	26.36	26.38	26.40	26.42	26.44
700	26.46	26.48	26.50	26.51	26.53	26.55	26.57	26.59	26.61	26.63
710	26.65	26.66	26.68	26.70	26.72	26.74	26.76	26.78	26.80	26.81
720	26.83	26.85	26.87	26.89	26.91	26.93	26.94	26.96	26.98	27.00
730	27.02	27.04	27.06	27.07	27.09	27.11	27.13	27.15	27.17	27.18
740	27.20	27.22	27.24	27.26	27.28	27.29	27.31	27.33	27.35	27.37
750	27.39	27.40	27.42	27.44	27.46	27.48	27.50	27.51	27.53	27.55
760	27.57	27.59	27.60	27.62	27.64	27.66	27.68	27.69	27.71	27.73
770	27.75	27.77	27.78	27.80	27.82	27.84	27.86	27.87	27.89	27.91
780	27.93	27.95	27.96	27.98	28.00	28.02	28.04	28.05	28.07	28.09
790	28.11	28.12	28.14	28.16	28.18	28.20	28.21	28.23	28.25	28.27
800	28.28	28.30	28.32	28.34	28.35	28.37	28.39	28.41	28.43	28.44
810	28.46	28.48	28.50	28.51	28.53	28.55	28.57	28.58	28.60	28.62
820	28.64	28.65	28.67	28.69	28.71	28.72	28.74	28.76	28.78	28.79
830	28.81	28.83	28.84	28.86	28.88	28.90	28.91	28.93	28.95	28.97
840	28.98	29.00	29.02	29.03	29.05	29.07	29.09	29.10	29.12	29.14
850	29.15	29.17	29.19	29.21	29.22	29.24	29.26	29.27	29.29	29.31
860	29.33	29.34	29.36	29.38	29.39	29.41	29.43	29.44	29.46	29.48
870	29.50	29.51	29.53	29.55	29.56	29.58	29.60	29.61	29.63	29.65
880	29.66	29.68	29.70	29.72	29.73	29.75	29.77	29.78	29.80	29.82
890	29.83	29.85	29.87	29.88	29.90	29.92	29.93	29.95	29.97	29.98
900	30.00	30.02	30.03	30.05	30.07	30.08	30.10	30.12	30.13	30.15
910	30.17	30.18	30.20	30.22	30.23	30.25	30.27	30.28	30.30	30.32
920	30.33	30.35	30.36	30.38	30.40	30.41	30.43	30.45	30.46	30.48
930	30.50	30.51	30.53	30.54	30.56	30.58	30.59	30.61	30.63	30.64
940	30.66	30.68	30.69	30.71	30.72	30.74	30.76	30.77	30.79	30.81
950	30.82	30.84	30.85	30.87	30.89	30.90	30.92	30.94	30.95	30.97
960	30.98	31.00	31.02	31.03	31.05	31.06	31.08	31.10	31.11	31.13
970	31.14	31.16	31.18	31.19	31.21	31.22	31.24	31.26	31.27	31.29
980	31.30	31.32	31.34	31.35	31.37	31.38	31.40	31.42	31.43	31.45
990	31.46	31.48	31.50	31.51	31.53	31.54	31.56	31.58	31.59	31.61

TABLE V.—RECIPROCAL OF NUMBERS¹

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.00	1.0000	.9901	.9804	.9709	.9615	.9524	.9434	.9346	.9259	.9174
1.10	.9091	.9009	.8929	.8850	.8772	.8696	.8621	.8547	.8475	.8403
1.20	.8333	.8264	.8197	.8130	.8065	.8000	.7937	.7874	.7812	.7752
1.30	.7692	.7634	.7576	.7519	.7463	.7407	.7353	.7299	.7246	.7194
1.40	.7143	.7092	.7042	.6993	.6944	.6897	.6849	.6803	.6757	.6711
1.50	.6667	.6623	.6579	.6536	.6494	.6452	.6410	.6369	.6329	.6289
1.60	.6250	.6211	.6173	.6135	.6098	.6061	.6024	.5988	.5952	.5917
1.70	.5882	.5848	.5814	.5780	.5747	.5714	.5682	.5650	.5618	.5587
1.80	.5556	.5525	.5495	.5464	.5435	.5405	.5376	.5348	.5319	.5291
1.90	.5263	.5236	.5208	.5181	.5155	.5128	.5102	.5076	.5051	.5025
2.00	.5000	.4975	.4950	.4926	.4902	.4878	.4854	.4831	.4808	.4785
2.10	.4762	.4739	.4717	.4694	.4673	.4651	.4630	.4608	.4587	.4566
2.20	.4545	.4525	.4504	.4484	.4464	.4444	.4425	.4405	.4386	.4367
2.30	.4348	.4329	.4310	.4292	.4274	.4255	.4237	.4219	.4202	.4184
2.40	.4167	.4149	.4132	.4115	.4098	.4082	.4065	.4049	.4032	.4016
2.50	.4000	.3984	.3968	.3953	.3937	.3922	.3906	.3891	.3876	.3861
2.60	.3846	.3831	.3817	.3802	.3788	.3774	.3759	.3745	.3731	.3717
2.70	.3704	.3690	.3676	.3663	.3650	.3636	.3623	.3610	.3597	.3584
2.80	.3571	.3559	.3546	.3534	.3521	.3509	.3496	.3484	.3472	.3460
2.90	.3448	.3436	.3425	.3413	.3401	.3390	.3378	.3367	.3356	.3344
3.00	.3333	.3322	.3311	.3300	.3289	.3279	.3268	.3257	.3247	.3236
3.10	.3226	.3215	.3205	.3195	.3185	.3175	.3165	.3155	.3145	.3135
3.20	.3125	.3115	.3106	.3096	.3086	.3077	.3067	.3058	.3049	.3040
3.30	.3030	.3021	.3012	.3003	.2994	.2985	.2976	.2967	.2959	.2950
3.40	.2941	.2933	.2924	.2915	.2907	.2899	.2890	.2882	.2874	.2865
3.50	.2857	.2849	.2841	.2833	.2825	.2817	.2809	.2801	.2793	.2786
3.60	.2778	.2770	.2762	.2755	.2747	.2740	.2732	.2725	.2717	.2710
3.70	.2703	.2695	.2688	.2681	.2674	.2667	.2660	.2653	.2646	.2639
3.80	.2632	.2625	.2618	.2611	.2604	.2597	.2591	.2584	.2577	.2571
3.90	.2564	.2558	.2551	.2545	.2538	.2532	.2525	.2519	.2513	.2506
4.00	.2500	.2494	.2488	.2481	.2475	.2469	.2463	.2457	.2451	.2445
4.10	.2439	.2433	.2427	.2421	.2415	.2410	.2404	.2398	.2392	.2387
4.20	.2381	.2375	.2370	.2364	.2358	.2353	.2347	.2342	.2336	.2331
4.30	.2326	.2320	.2315	.2309	.2304	.2299	.2294	.2288	.2283	.2278
4.40	.2273	.2268	.2262	.2257	.2252	.2247	.2242	.2237	.2232	.2227
4.50	.2222	.2217	.2212	.2208	.2203	.2198	.2193	.2188	.2183	.2179
4.60	.2174	.2169	.2164	.2160	.2155	.2151	.2146	.2141	.2137	.2132
4.70	.2128	.2123	.2119	.2114	.2110	.2105	.2101	.2096	.2092	.2088
4.80	.2083	.2079	.2075	.2070	.2066	.2062	.2058	.2053	.2049	.2045
4.90	.2041	.2037	.2033	.2028	.2024	.2020	.2016	.2012	.2008	.2004
5.00	.2000	.1996	.1992	.1988	.1984	.1980	.1976	.1972	.1968	.1965
5.10	.1961	.1957	.1953	.1949	.1946	.1942	.1938	.1934	.1930	.1927
5.20	.1923	.1919	.1916	.1912	.1908	.1905	.1901	.1898	.1894	.1890
5.30	.1887	.1883	.1880	.1876	.1873	.1869	.1866	.1862	.1859	.1855
5.40	.1852	.1848	.1845	.1842	.1838	.1835	.1832	.1828	.1825	.1821

¹ Source: WAUGH, ALBERT E., *Laboratory Manual and Problems for Elements of Statistical Method* (McGraw-Hill Book Company, Inc., 1944).

TABLE V.—RECIPROCAL OF NUMBERS.—(Continued)

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
5.50	.1818	.1815	.1812	.1808	.1805	.1802	.1799	.1795	.1792	.1789
5.60	.1786	.1783	.1779	.1776	.1773	.1770	.1767	.1764	.1761	.1757
5.70	.1754	.1751	.1748	.1745	.1742	.1739	.1736	.1733	.1730	.1727
5.80	.1724	.1721	.1718	.1715	.1712	.1709	.1706	.1704	.1701	.1698
5.90	.1695	.1692	.1689	.1686	.1684	.1681	.1678	.1675	.1672	.1669
6.00	.1667	.1664	.1661	.1658	.1656	.1653	.1650	.1647	.1645	.1642
6.10	.1639	.1637	.1634	.1631	.1629	.1626	.1623	.1621	.1618	.1616
6.20	.1613	.1610	.1608	.1605	.1603	.1600	.1597	.1595	.1592	.1590
6.30	.1587	.1585	.1582	.1580	.1577	.1575	.1572	.1570	.1567	.1565
6.40	.1562	.1560	.1558	.1555	.1553	.1550	.1548	.1546	.1543	.1541
6.50	.1538	.1536	.1534	.1531	.1529	.1527	.1524	.1522	.1520	.1517
6.60	.1515	.1513	.1511	.1508	.1506	.1504	.1502	.1499	.1497	.1495
6.70	.1493	.1490	.1488	.1486	.1484	.1481	.1479	.1477	.1475	.1473
6.80	.1471	.1468	.1466	.1464	.1462	.1460	.1458	.1456	.1453	.1451
6.90	.1449	.1447	.1445	.1443	.1441	.1439	.1437	.1435	.1433	.1431
7.00	.1429	.1427	.1424	.1422	.1420	.1418	.1416	.1414	.1412	.1410
7.10	.1408	.1406	.1404	.1403	.1401	.1399	.1397	.1395	.1393	.1391
7.20	.1389	.1387	.1385	.1383	.1381	.1379	.1377	.1376	.1374	.1372
7.30	.1370	.1368	.1366	.1364	.1362	.1361	.1359	.1357	.1355	.1353
7.40	.1351	.1350	.1348	.1346	.1344	.1342	.1340	.1339	.1337	.1335
7.50	.1333	.1332	.1330	.1328	.1326	.1324	.1323	.1321	.1319	.1318
7.60	.1316	.1314	.1312	.1311	.1309	.1307	.1305	.1304	.1302	.1300
7.70	.1299	.1297	.1295	.1294	.1292	.1290	.1289	.1287	.1285	.1284
7.80	.1282	.1280	.1279	.1277	.1276	.1274	.1272	.1271	.1269	.1267
7.90	.1266	.1264	.1263	.1261	.1259	.1258	.1256	.1255	.1253	.1252
8.00	.1250	.1248	.1247	.1245	.1244	.1242	.1241	.1239	.1238	.1236
8.10	.1235	.1233	.1232	.1230	.1228	.1227	.1225	.1224	.1222	.1221
8.20	.1220	.1218	.1217	.1215	.1214	.1212	.1211	.1209	.1208	.1206
8.30	.1205	.1203	.1202	.1200	.1199	.1198	.1196	.1195	.1193	.1192
8.40	.1190	.1189	.1188	.1186	.1185	.1183	.1182	.1181	.1179	.1178
8.50	.1176	.1175	.1174	.1172	.1171	.1170	.1168	.1167	.1166	.1164
8.60	.1163	.1161	.1160	.1159	.1157	.1156	.1155	.1153	.1152	.1151
8.70	.1149	.1148	.1147	.1145	.1144	.1143	.1142	.1140	.1139	.1138
8.80	.1136	.1135	.1134	.1132	.1131	.1130	.1129	.1127	.1126	.1125
8.90	.1124	.1122	.1121	.1120	.1119	.1117	.1116	.1115	.1114	.1112
9.00	.1111	.1110	.1109	.1107	.1106	.1105	.1104	.1103	.1101	.1100
9.10	.1099	.1098	.1096	.1095	.1094	.1093	.1092	.1091	.1089	.1088
9.20	.1087	.1086	.1085	.1083	.1082	.1081	.1080	.1079	.1078	.1076
9.30	.1075	.1074	.1073	.1072	.1071	.1070	.1068	.1067	.1066	.1065
9.40	.1064	.1063	.1062	.1060	.1059	.1058	.1057	.1056	.1055	.1054
9.50	.1053	.1052	.1050	.1049	.1048	.1047	.1046	.1045	.1044	.1043
9.60	.1042	.1041	.1040	.1038	.1037	.1036	.1035	.1034	.1033	.1032
9.70	.1031	.1030	.1029	.1028	.1027	.1026	.1025	.1024	.1022	.1021
9.80	.1020	.1019	.1018	.1017	.1016	.1015	.1014	.1013	.1012	.1011
9.90	.1010	.1009	.1008	.1007	.1006	.1005	.1004	.1003	.1002	.1001

TABLE VI.—AREAS, ORDINATES, AND DERIVATIVES OF THE NORMAL CURVE¹

The following table gives proportions of the area under the normal curve from $x/\delta = 0$ to values of x/δ given in column (1). Values of the ordinates and of the third and fourth derivatives are also given.

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
$\frac{x}{\delta}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\delta} \right)^*$	$\varphi_4 \left(\frac{x}{\delta} \right)^*$	$\frac{x}{\delta}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\delta} \right)^*$	$\varphi_4 \left(\frac{x}{\delta} \right)^*$
.00	.0000	.3989	.0000	1.1968	.50	.1915	.3521	.4841	.5501
.01	.0040	.3989	.0120	1.1965	.51	.1950	.3503	.4895	.5279
.02	.0080	.3989	.0239	1.1956	.52	.1985	.3485	.4947	.5056
.03	.0120	.3988	.0359	1.1941	.53	.2019	.3467	.4996	.4831
.04	.0160	.3986	.0478	1.1920	.54	.2054	.3448	.5043	.4605
.05	.0199	.3984	.0597	1.1894	.55	.2088	.3429	.5088	.4378
.06	.0239	.3982	.0716	1.1861	.56	.2123	.3411	.5131	.4150
.07	.0279	.3980	.0834	1.1822	.57	.2157	.3391	.5171	.3921
.08	.0319	.3977	.0952	1.1778	.58	.2190	.3372	.5209	.3691
.09	.0359	.3973	.1070	1.1727	.59	.2224	.3352	.5245	.3461
.10	.0398	.3970	.1187	1.1671	.60	.2258	.3332	.5278	.3231
.11	.0438	.3965	.1303	1.1609	.61	.2291	.3312	.5309	.3000
.12	.0478	.3961	.1419	1.1541	.62	.2324	.3292	.5338	.2770
.13	.0517	.3956	.1534	1.1468	.63	.2357	.3271	.5365	.2539
.14	.0557	.3951	.1648	1.1389	.64	.2389	.3251	.5389	.2309
.15	.0596	.3945	.1762	1.1304	.65	.2422	.3230	.5411	.2078
.16	.0636	.3939	.1874	1.1214	.66	.2454	.3209	.5431	.1849
.17	.0675	.3932	.1986	1.1118	.67	.2486	.3187	.5448	.1620
.18	.0714	.3925	.2097	1.1017	.68	.2518	.3166	.5463	.1391
.19	.0754	.3918	.2206	1.0911	.69	.2549	.3144	.5476	.1164
.20	.0793	.3910	.2315	1.0799	.70	.2580	.3123	.5486	.0937
.21	.0832	.3902	.2422	1.0682	.71	.2612	.3101	.5495	.0712
.22	.0871	.3894	.2529	1.0560	.72	.2642	.3079	.5501	.0487
.23	.0910	.3885	.2634	1.0434	.73	.2673	.3056	.5504	.0265
.24	.0948	.3876	.2737	1.0302	.74	.2704	.3034	.5506	.0043
.25	.0987	.3867	.2840	1.0165	.75	.2734	.3011	.5505	— .0176
.26	.1026	.3857	.2941	1.0024	.76	.2764	.2989	.5502	— .0394
.27	.1064	.3847	.3040	0.9878	.77	.2794	.2966	.5497	— .0611
.28	.1103	.3836	.3138	0.9727	.78	.2823	.2943	.5490	— .0825
.29	.1141	.3825	.3235	0.9572	.79	.2852	.2920	.5481	— .1037
.30	.1179	.3814	.3330	0.9413	.80	.2881	.2897	.5469	— .1247
.31	.1217	.3802	.3423	0.9250	.81	.2910	.2874	.5456	— .1455
.32	.1255	.3790	.3515	0.9082	.82	.2939	.2850	.5440	— .1660
.33	.1293	.3778	.3605	0.8910	.83	.2967	.2827	.5423	— .1862
.34	.1331	.3765	.3693	0.8735	.84	.2996	.2803	.5403	— .2063
.35	.1368	.3752	.3779	0.8556	.85	.3023	.2780	.5381	— .2260
.36	.1406	.3739	.3864	0.8373	.86	.3051	.2756	.5358	— .2455
.37	.1443	.3726	.3947	0.8186	.87	.3079	.2732	.5332	— .2646
.38	.1480	.3712	.4028	0.7996	.88	.3106	.2709	.5305	— .2835
.39	.1517	.3697	.4107	0.7803	.89	.3133	.2685	.5276	— .3021
.40	.1554	.3683	.4184	0.7607	.90	.3159	.2661	.5245	— .3203
.41	.1591	.3668	.4259	0.7408	.91	.3186	.2637	.5212	— .3383
.42	.1628	.3653	.4332	0.7206	.92	.3212	.2613	.5177	— .3559
.43	.1664	.3637	.4403	0.7001	.93	.3238	.2589	.5140	— .3731
.44	.1700	.3621	.4472	0.6793	.94	.3264	.2565	.5102	— .3901
.45	.1736	.3605	.4539	0.6583	.95	.3289	.2541	.5062	— .4066
.46	.1772	.3589	.4603	0.6371	.96	.3315	.2516	.5021	— .4228
.47	.1808	.3572	.4666	0.6156	.97	.3340	.2492	.4978	— .4387
.48	.1844	.3555	.4727	0.5940	.98	.3365	.2468	.4933	— .4541
.49	.1879	.3538	.4785	0.5721	.99	.3389	.2444	.4887	— .4692
.50	.1915	.3521	.4841	0.5501	1.00	.3413	.2420	.4839	— .4839

¹ Reproduced by permission from *Mathematical Tables from Handbook of Chemistry and Physics* compiled by Charles D. Hodgman, 7th ed., 1941, pp. 200–204.

* If the ordinate shown in column (3) is designated as $\varphi_0 \left(\frac{x}{\delta} \right)$, then $\varphi_1 \left(\frac{x}{\delta} \right)$ is defined as $\varphi_0 \left(\frac{x}{\delta} \right)$ multiplied by $\left(\frac{3x}{\delta} - \frac{x^3}{\delta^3} \right)$ and $\varphi_4 \left(\frac{x}{\delta} \right)$ is defined as $\varphi_0 \left(\frac{x}{\delta} \right)$ multiplied by

$$\left(3 - \frac{6x^2}{\delta^2} + \frac{x^4}{\delta^4} \right)$$

By successive differentiation of $\varphi_0 \left(\frac{x}{\delta} \right)$ it can be shown that $\varphi_1 \left(\frac{x}{\delta} \right)$ is also the third derivative of $\varphi_0 \left(\frac{x}{\delta} \right)$ and $\varphi_4 \left(\frac{x}{\delta} \right)$ is the fourth derivative of $\varphi_0 \left(\frac{x}{\delta} \right)$ (see Chap. VII, pp. 142–145).

TABLE VI.—AREAS, ORDINATES, AND DERIVATIVES OF THE NORMAL CURVE¹
(Continued)

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$	$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$
1.00	.3413	.2420	.4839	— .4839	1.50	.4332	.1295	.1457	— .7043
1.01	.3438	.2396	.4790	— .4989	1.51	.4345	.1276	.1387	— .6904
1.02	.3461	.2371	.4740	— .5122	1.52	.4357	.1257	.1317	— .6942
1.03	.3485	.2347	.4688	— .5257	1.53	.4370	.1238	.1248	— .6888
1.04	.3508	.2323	.4635	— .5389	1.54	.4382	.1219	.1180	— .6831
1.05	.3531	.2299	.4580	— .5516	1.55	.4394	.1200	.1111	— .6772
1.06	.3554	.2275	.4524	— .5639	1.56	.4406	.1182	.1044	— .6710
1.07	.3577	.2251	.4467	— .5758	1.57	.4418	.1163	.0977	— .6646
1.08	.3599	.2227	.4409	— .5873	1.58	.4430	.1145	.0911	— .6580
1.09	.3621	.2203	.4350	— .5984	1.59	.4441	.1127	.0846	— .6511
1.10	.3643	.2179	.4290	— .6091	1.60	.4452	.1109	.0781	— .6441
1.11	.3665	.2155	.4228	— .6193	1.61	.4463	.1092	.0717	— .6368
1.12	.3686	.2131	.4166	— .6292	1.62	.4474	.1074	.0654	— .6293
1.13	.3708	.2107	.4102	— .6386	1.63	.4485	.1057	.0591	— .6216
1.14	.3729	.2083	.4038	— .6476	1.64	.4495	.1040	.0529	— .6138
1.15	.3749	.2059	.3973	— .6561	1.65	.4505	.1023	.0468	— .6057
1.16	.3770	.2036	.3907	— .6643	1.66	.4515	.1006	.0408	— .5975
1.17	.3790	.2012	.3840	— .6720	1.67	.4525	.0989	.0349	— .5891
1.18	.3810	.1989	.3772	— .6792	1.68	.4535	.0973	.0290	— .5806
1.19	.3830	.1965	.3704	— .6861	1.69	.4545	.0957	.0233	— .5720
1.20	.3849	.1942	.3635	— .6926	1.70	.4554	.0941	.0176	— .5632
1.21	.3869	.1919	.3566	— .6980	1.71	.4564	.0925	.0120	— .5542
1.22	.3888	.1895	.3496	— .7042	1.72	.4573	.0909	.0065	— .5452
1.23	.3907	.1872	.3425	— .7094	1.73	.4582	.0893	.0011	— .5360
1.24	.3925	.1849	.3354	— .7141	1.74	.4591	.0878	— .0042	— .5267
1.25	.3944	.1827	.3282	— .7185	1.75	.4599	.0863	— .0094	— .5173
1.26	.3962	.1804	.3210	— .7224	1.76	.4608	.0848	— .0146	— .5079
1.27	.3980	.1781	.3138	— .7259	1.77	.4616	.0833	— .0196	— .4983
1.28	.3997	.1759	.3065	— .7291	1.78	.4625	.0818	— .0245	— .4887
1.29	.4015	.1736	.2992	— .7318	1.79	.4633	.0804	— .0294	— .4789
1.30	.4032	.1714	.2918	— .7341	1.80	.4641	.0790	— .0341	— .4692
1.31	.4049	.1692	.2845	— .7361	1.81	.4649	.0775	— .0388	— .4593
1.32	.4066	.1669	.2771	— .7376	1.82	.4656	.0761	— .0433	— .4494
1.33	.4082	.1647	.2697	— .7388	1.83	.4664	.0748	— .0477	— .4395
1.34	.4099	.1626	.2624	— .7395	1.84	.4671	.0734	— .0521	— .4295
1.35	.4115	.1604	.2550	— .7399	1.85	.4678	.0721	— .0563	— .4195
1.36	.4131	.1582	.2476	— .7400	1.86	.4686	.0707	— .0605	— .4095
1.37	.4147	.1561	.2402	— .7396	1.87	.4693	.0694	— .0645	— .3995
1.38	.4162	.1540	.2328	— .7389	1.88	.4700	.0681	— .0685	— .3894
1.39	.4177	.1518	.2254	— .7378	1.89	.4706	.0669	— .0723	— .3793
1.40	.4192	.1497	.2180	— .7364	1.90	.4713	.0656	— .0761	— .3693
1.41	.4207	.1476	.2107	— .7347	1.91	.4719	.0644	— .0797	— .3592
1.42	.4222	.1456	.2033	— .7326	1.92	.4726	.0632	— .0832	— .3492
1.43	.4236	.1435	.1960	— .7301	1.93	.4732	.0620	— .0867	— .3392
1.44	.4251	.1415	.1887	— .7274	1.94	.4738	.0608	— .0900	— .3292
1.45	.4265	.1394	.1815	— .7243	1.95	.4744	.0596	— .0933	— .3192
1.46	.4279	.1374	.1742	— .7209	1.96	.4750	.0584	— .0964	— .3093
1.47	.4292	.1354	.1670	— .7172	1.97	.4756	.0573	— .0994	— .2994
1.48	.4306	.1334	.1599	— .7132	1.98	.4762	.0562	— .1024	— .2895
1.49	.4319	.1315	.1528	— .7089	1.99	.4767	.0551	— .1052	— .2797
1.50	.4332	.1295	.1457	— .7043	2.00	.4773	.0540	— .1080	— .2700

¹ Reproduced by permission from *Mathematical Tables from Handbook of Chemistry and Physics* compiled by Charles D. Hodgman, 7th ed., 1941, pp. 200–204.

* If the ordinate shown in column (3) is designated as $\varphi_0 \left(\frac{x}{\sigma} \right)$, then $\varphi_3 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by $\left(\frac{3x}{\sigma} - \frac{x^3}{\sigma^3} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by

$$\left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right).$$

By successive differentiation of $\varphi_0 \left(\frac{x}{\sigma} \right)$ it can be shown that $\varphi_3 \left(\frac{x}{\sigma} \right)$ is also the third derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is the fourth derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ (see Chap. VII, pp. 142–145).

TABLE VI.—AREAS, ORDINATES, AND DERIVATIVES OF THE NORMAL CURVE¹
(Continued)

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_2 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$	$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_2 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$
2.00	.4773	.0540	— .1080	— .2700	2.50	.4938	.0175	— .1424	.0800
2.01	.4778	.0529	— .1106	— .2603	2.51	.4940	.0171	— .1416	.0838
2.02	.4783	.0519	— .1132	— .2506	2.52	.4941	.0167	— .1408	.0871
2.03	.4788	.0508	— .1157	— .2411	2.53	.4943	.0163	— .1399	.0905
2.04	.4793	.0498	— .1180	— .2316	2.54	.4945	.0159	— .1389	.0937
2.05	.4798	.0488	— .1203	— .2222	2.55	.4946	.0155	— .1380	.0968
2.06	.4803	.0478	— .1225	— .2129	2.56	.4948	.0151	— .1370	.0998
2.07	.4808	.0468	— .1245	— .2036	2.57	.4949	.0147	— .1360	.1027
2.08	.4812	.0459	— .1265	— .1945	2.58	.4951	.0143	— .1350	.1054
2.09	.4817	.0449	— .1284	— .1854	2.59	.4952	.0139	— .1339	.1080
2.10	.4821	.0440	— .1302	— .1765	2.60	.4953	.0136	— .1328	.1105
2.11	.4826	.0431	— .1320	— .1676	2.61	.4955	.0132	— .1317	.1129
2.12	.4830	.0422	— .1336	— .1588	2.62	.4956	.0129	— .1305	.1152
2.13	.4834	.0413	— .1351	— .1502	2.63	.4957	.0126	— .1294	.1173
2.14	.4838	.0404	— .1366	— .1416	2.64	.4959	.0122	— .1282	.1194
2.15	.4842	.0396	— .1380	— .1332	2.65	.4960	.0119	— .1270	.1213
2.16	.4846	.0387	— .1393	— .1249	2.66	.4961	.0116	— .1258	.1231
2.17	.4850	.0379	— .1405	— .1167	2.67	.4962	.0113	— .1245	.1248
2.18	.4854	.0371	— .1416	— .1086	2.68	.4963	.0110	— .1233	.1264
2.19	.4857	.0363	— .1426	— .1006	2.69	.4964	.0107	— .1220	.1279
2.20	.4861	.0355	— .1436	— .0927	2.70	.4965	.0104	— .1207	.1293
2.21	.4865	.0347	— .1445	— .0850	2.71	.4966	.0101	— .1194	.1306
2.22	.4868	.0339	— .1453	— .0774	2.72	.4967	.0099	— .1181	.1317
2.23	.4871	.0332	— .1460	— .0700	2.73	.4968	.0096	— .1168	.1328
2.24	.4875	.0325	— .1467	— .0626	2.74	.4969	.0094	— .1154	.1338
2.25	.4878	.0317	— .1473	— .0554	2.75	.4970	.0091	— .1141	.1347
2.26	.4881	.0310	— .1478	— .0484	2.76	.4971	.0089	— .1127	.1356
2.27	.4884	.0303	— .1483	— .0414	2.77	.4972	.0086	— .1114	.1363
2.28	.4887	.0297	— .1486	— .0346	2.78	.4973	.0084	— .1100	.1369
2.29	.4890	.0290	— .1490	— .0279	2.79	.4974	.0081	— .1087	.1375
2.30	.4893	.0283	— .1492	— .0214	2.80	.4974	.0079	— .1073	.1379
2.31	.4896	.0277	— .1494	— .0150	2.81	.4975	.0077	— .1059	.1383
2.32	.4898	.0271	— .1495	— .0088	2.82	.4976	.0075	— .1045	.1386
2.33	.4901	.0264	— .1496	— .0027	2.83	.4977	.0073	— .1031	.1389
2.34	.4904	.0258	— .1496	.0033	2.84	.4977	.0071	— .1017	.1390
2.35	.4906	.0252	— .1495	.0092	2.85	.4978	.0069	— .1003	.1391
2.36	.4909	.0246	— .1494	.0149	2.86	.4979	.0067	— .0990	.1391
2.37	.4911	.0241	— .1492	.0204	2.87	.4980	.0065	— .0976	.1391
2.38	.4913	.0235	— .1490	.0258	2.88	.4980	.0063	— .0962	.1389
2.39	.4916	.0229	— .1487	.0311	2.89	.4981	.0061	— .0948	.1388
2.40	.4918	.0224	— .1483	.0362	2.90	.4981	.0060	— .0934	.1385
2.41	.4920	.0219	— .1480	.0412	2.91	.4982	.0058	— .0920	.1382
2.42	.4922	.0213	— .1475	.0461	2.92	.4983	.0056	— .0906	.1378
2.43	.4925	.0208	— .1470	.0508	2.93	.4983	.0055	— .0893	.1374
2.44	.4927	.0203	— .1465	.0554	2.94	.4984	.0053	— .0879	.1369
2.45	.4929	.0198	— .1459	.0598	2.95	.4984	.0051	— .0865	.1364
2.46	.4931	.0194	— .1453	.0641	2.96	.4985	.0050	— .0852	.1358
2.47	.4932	.0189	— .1446	.0683	2.97	.4985	.0049	— .0838	.1352
2.48	.4934	.0184	— .1439	.0723	2.98	.4986	.0047	— .0825	.1345
2.49	.4936	.0180	— .1432	.0762	2.99	.4986	.0046	— .0811	.1337
2.50	.4938	.0175	— .1424	.0800	3.00	.4987	.0044	— .0798	.1330

¹ Reproduced by permission from *Mathematical Tables from Handbook of Chemistry and Physics* compiled by Charles D. Hodgman, 7th ed., 1941, pp. 200–204.

* If the ordinate shown in column (3) is designated as $\varphi_0 \left(\frac{x}{\sigma} \right)$, then $\varphi_2 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by $\left(3\frac{x}{\sigma} - \frac{x^3}{\sigma^3} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by

$$\left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right).$$

By successive differentiation of $\varphi_0 \left(\frac{x}{\sigma} \right)$ it can be shown that $\varphi_2 \left(\frac{x}{\sigma} \right)$ is also the third derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is the fourth derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ (see Chap. VII, pp.142–145).

TABLE VI.—AREAS, ORDINATES, AND DERIVATIVES OF THE NORMAL CURVE¹
(Continued)

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$	$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$
3.00	.4987	.0044	-.0798	.1330	3.50	.4998	.0009	-.0283	.0694
3.01	.4987	.0043	-.0785	.1321	3.51	.4998	.0008	-.0276	.0681
3.02	.4987	.0042	-.0771	.1313	3.52	.4998	.0008	-.0269	.0669
3.03	.4988	.0041	-.0758	.1304	3.53	.4998	.0008	-.0262	.0656
3.04	.4988	.0039	-.0745	.1294	3.54	.4998	.0008	-.0256	.0643
3.05	.4989	.0038	-.0732	.1285	3.55	.4998	.0007	-.0249	.0631
3.06	.4989	.0037	-.0720	.1275	3.56	.4998	.0007	-.0243	.0618
3.07	.4989	.0036	-.0707	.1264	3.57	.4998	.0007	-.0237	.0606
3.08	.4990	.0035	-.0694	.1254	3.58	.4998	.0007	-.0231	.0594
3.09	.4990	.0034	-.0682	.1243	3.59	.4998	.0006	-.0225	.0582
3.10	.4990	.0033	-.0669	.1231	3.60	.4998	.0006	-.0219	.0570
3.11	.4991	.0032	-.0657	.1220	3.61	.4999	.0006	-.0214	.0559
3.12	.4991	.0031	-.0645	.1208	3.62	.4999	.0006	-.0208	.0547
3.13	.4991	.0030	-.0633	.1196	3.63	.4999	.0006	-.0203	.0536
3.14	.4992	.0029	-.0621	.1184	3.64	.4999	.0005	-.0198	.0524
3.15	.4992	.0028	-.0609	.1171	3.65	.4999	.0005	-.0192	.0513
3.16	.4992	.0027	-.0598	.1159	3.66	.4999	.0005	-.0187	.0502
3.17	.4992	.0026	-.0586	.1146	3.67	.4999	.0005	-.0182	.0492
3.18	.4993	.0025	-.0575	.1133	3.68	.4999	.0005	-.0177	.0481
3.19	.4993	.0025	-.0564	.1120	3.69	.4999	.0004	-.0173	.0470
3.20	.4993	.0024	-.0552	.1107	3.70	.4999	.0004	-.0168	.0460
3.21	.4993	.0023	-.0541	.1093	3.71	.4999	.0004	-.0164	.0450
3.22	.4994	.0022	-.0531	.1080	3.72	.4999	.0004	-.0159	.0440
3.23	.4994	.0022	-.0520	.1066	3.73	.4999	.0004	-.0155	.0430
3.24	.4994	.0021	-.0509	.1053	3.74	.4999	.0004	-.0150	.0420
3.25	.4994	.0020	-.0499	.1039	3.75	.4999	.0004	-.0146	.0410
3.26	.4994	.0020	-.0488	.1025	3.76	.4999	.0003	-.0142	.0401
3.27	.4995	.0019	-.0478	.1011	3.77	.4999	.0003	-.0138	.0392
3.28	.4995	.0018	-.0468	.0997	3.78	.4999	.0003	-.0134	.0382
3.29	.4995	.0018	-.0458	.0983	3.79	.4999	.0003	-.0131	.0373
3.30	.4995	.0017	-.0449	.0969	3.80	.4999	.0003	-.0127	.0365
3.31	.4995	.0017	-.0439	.0955	3.81	.4999	.0003	-.0123	.0356
3.32	.4996	.0016	-.0429	.0941	3.82	.4999	.0003	-.0120	.0347
3.33	.4996	.0016	-.0420	.0927	3.83	.4999	.0003	-.0116	.0339
3.34	.4996	.0015	-.0411	.0913	3.84	.4999	.0003	-.0113	.0331
3.35	.4996	.0015	-.0402	.0899	3.85	.4999	.0002	-.0110	.0323
3.36	.4996	.0014	-.0393	.0885	3.86	.4999	.0002	-.0107	.0315
3.37	.4996	.0014	-.0384	.0871	3.87	.5000	.0002	-.0104	.0307
3.38	.4996	.0013	-.0376	.0857	3.88	.5000	.0002	-.0100	.0299
3.39	.4997	.0013	-.0367	.0843	3.89	.5000	.0002	-.0098	.0292
3.40	.4997	.0012	-.0359	.0829	3.90	.5000	.0002	-.0095	.0284
3.41	.4997	.0012	-.0350	.0815	3.91	.5000	.0002	-.0092	.0277
3.42	.4997	.0012	-.0342	.0801	3.92	.5000	.0002	-.0089	.0270
3.43	.4997	.0011	-.0334	.0788	3.93	.5000	.0002	-.0086	.0263
3.44	.4997	.0011	-.0327	.0774	3.94	.5000	.0002	-.0084	.0256
3.45	.4997	.0010	-.0319	.0761	3.95	.5000	.0002	-.0081	.0250
3.46	.4997	.0010	-.0311	.0747	3.96	.5000	.0002	-.0079	.0243
3.47	.4997	.0010	-.0304	.0734	3.97	.5000	.0002	-.0076	.0237
3.48	.4998	.0009	-.0297	.0721	3.98	.5000	.0001	-.0074	.0230
3.49	.4998	.0009	-.0290	.0707	3.99	.5000	.0001	-.0072	.0224
3.50	.4998	.0009	-.0283	.0694	4.00	.5000	.0001	-.0070	.0218

¹ Reproduced by permission from *Mathematical Tables from Handbook of Chemistry and Physics* compiled by Charles D. Hodgman, 7th ed., 1941, pp. 200-204.

* If the ordinate shown in column (3) is designated as $\varphi_0 \left(\frac{x}{\sigma} \right)$, then $\varphi_3 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by $\left(3\frac{x}{\sigma} - \frac{x^3}{\sigma^3} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by

$$\left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right).$$

By successive differentiation of $\varphi_0 \left(\frac{x}{\sigma} \right)$ it can be shown that $\varphi_3 \left(\frac{x}{\sigma} \right)$ is also the third derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is the fourth derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ (see Chap. VII, pp. 142-145).

TABLE VI.—AREAS, ORDINATES, AND DERIVATIVES OF THE NORMAL CURVE¹
(Concluded)

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$	$\frac{x}{\sigma}$	Area	Ordinate	$\varphi_3 \left(\frac{x}{\sigma} \right)^*$	$\varphi_4 \left(\frac{x}{\sigma} \right)^*$
4.00	.5000	.0001	— .0070	.0218	4.50	.5000	.0000	— .0012	.0047
4.01	.5000	.0001	— .0067	.0212	4.51	.5000	.0000	— .0012	.0045
4.02	.5000	.0001	— .0065	.0207	4.52	.5000	.0000	— .0012	.0044
4.03	.5000	.0001	— .0063	.0201	4.53	.5000	.0000	— .0011	.0042
4.04	.5000	.0001	— .0061	.0195	4.54	.5000	.0000	— .0011	.0041
4.05	.5000	.0001	— .0059	.0190	4.55	.5000	.0000	— .0010	.0039
4.06	.5000	.0001	— .0058	.0185	4.56	.5000	.0000	— .0010	.0038
4.07	.5000	.0001	— .0056	.0180	4.57	.5000	.0000	— .0010	.0037
4.08	.5000	.0001	— .0054	.0175	4.58	.5000	.0000	— .0009	.0035
4.09	.5000	.0001	— .0052	.0170	4.59	.5000	.0000	— .0009	.0034
4.10	.5000	.0001	— .0051	.0165	4.60	.5000	.0000	— .0009	.0033
4.11	.5000	.0001	— .0049	.0160	4.61	.5000	.0000	— .0008	.0032
4.12	.5000	.0001	— .0047	.0156	4.62	.5000	.0000	— .0008	.0031
4.13	.5000	.0001	— .0046	.0151	4.63	.5000	.0000	— .0008	.0030
4.14	.5000	.0001	— .0044	.0147	4.64	.5000	.0000	— .0007	.0028
4.15	.5000	.0001	— .0043	.0143	4.65	.5000	.0000	— .0007	.0027
4.16	.5000	.0001	— .0042	.0138	4.66	.5000	.0000	— .0007	.0026
4.17	.5000	.0001	— .0040	.0134	4.67	.5000	.0000	— .0006	.0026
4.18	.5000	.0001	— .0039	.0130	4.68	.5000	.0000	— .0006	.0025
4.19	.5000	.0001	— .0038	.0127	4.69	.5000	.0000	— .0006	.0024
4.20	.5000	.0001	— .0036	.0123	4.70	.5000	.0000	— .0006	.0023
4.21	.5000	.0001	— .0035	.0119	4.71	.5000	.0000	— .0006	.0022
4.22	.5000	.0001	— .0034	.0116	4.72	.5000	.0000	— .0005	.0021
4.23	.5000	.0001	— .0033	.0112	4.73	.5000	.0000	— .0005	.0020
4.24	.5000	.0001	— .0032	.0109	4.74	.5000	.0000	— .0005	.0020
4.25	.5000	.0001	— .0031	.0105	4.75	.5000	.0000	— .0005	.0019
4.26	.5000	.0001	— .0030	.0102	4.76	.5000	.0000	— .0005	.0018
4.27	.5000	.0000	— .0029	.0099	4.77	.5000	.0000	— .0004	.0018
4.28	.5000	.0000	— .0028	.0096	4.78	.5000	.0000	— .0004	.0017
4.29	.5000	.0000	— .0027	.0093	4.79	.5000	.0000	— .0004	.0016
4.30	.5000	.0000	— .0026	.0090	4.80	.5000	.0000	— .0004	.0016
4.31	.5000	.0000	— .0025	.0087	4.81	.5000	.0000	— .0004	.0015
4.32	.5000	.0000	— .0024	.0085	4.82	.5000	.0000	— .0004	.0015
4.33	.5000	.0000	— .0023	.0082	4.83	.5000	.0000	— .0003	.0014
4.34	.5000	.0000	— .0022	.0079	4.84	.5000	.0000	— .0003	.0013
4.35	.5000	.0000	— .0022	.0077	4.85	.5000	.0000	— .0003	.0013
4.36	.5000	.0000	— .0021	.0074	4.86	.5000	.0000	— .0003	.0012
4.37	.5000	.0000	— .0020	.0072	4.87	.5000	.0000	— .0003	.0012
4.38	.5000	.0000	— .0019	.0070	4.88	.5000	.0000	— .0003	.0012
4.39	.5000	.0000	— .0019	.0067	4.89	.5000	.0000	— .0003	.0011
4.40	.5000	.0000	— .0018	.0065	4.90	.5000	.0000	— .0003	.0011
4.41	.5000	.0000	— .0017	.0063	4.91	.5000	.0000	— .0002	.0010
4.42	.5000	.0000	— .0017	.0061	4.92	.5000	.0000	— .0002	.0010
4.43	.5000	.0000	— .0016	.0059	4.93	.5000	.0000	— .0002	.0009
4.44	.5000	.0000	— .0016	.0057	4.94	.5000	.0000	— .0002	.0009
4.45	.5000	.0000	— .0015	.0055	4.95	.5000	.0000	— .0002	.0009
4.46	.5000	.0000	— .0014	.0053	4.96	.5000	.0000	— .0002	.0008
4.47	.5000	.0000	— .0014	.0052	4.97	.5000	.0000	— .0002	.0008
4.48	.5000	.0000	— .0013	.0050	4.98	.5000	.0000	— .0002	.0008
4.49	.5000	.0000	— .0013	.0048	4.99	.5000	.0000	— .0002	.0007
4.50	.5000	.0000	— .0012	.0047					

¹ Reproduced by permission from *Mathematical Tables from Handbook of Chemistry and Physics* compiled by Charles D. Hodgman, 7th ed., 1941, pp. 200–204.

* If the ordinate shown in column (3) is designated as $\varphi_0 \left(\frac{x}{\sigma} \right)$, then $\varphi_3 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by $\left(\frac{3x}{\sigma} - \frac{x^3}{\sigma^3} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is defined as $\varphi_0 \left(\frac{x}{\sigma} \right)$ multiplied by

$$\left(3 - \frac{6x^2}{\sigma^2} + \frac{x^4}{\sigma^4} \right)$$

By successive differentiation of $\varphi_0 \left(\frac{x}{\sigma} \right)$ it can be shown that $\varphi_3 \left(\frac{x}{\sigma} \right)$ is also the third derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ and $\varphi_4 \left(\frac{x}{\sigma} \right)$ is the fourth derivative of $\varphi_0 \left(\frac{x}{\sigma} \right)$ (see Chap. VII, pp. 142–145).

TABLE VII.—TABLE OF t^*

The column headings in this table are probabilities. The figures in the body of the table are values of t . The stub of the table shows n , the degrees of freedom. The probabilities refer to the sum of the two tails of the t distribution, *i.e.*, for both $+$ and $-$ values of t . For example, for $n = 10$, probability .05 shows in the table $t = 2.228$, which means $P(t \geq |2.228|)$ equals .05; $P(t \geq +2.228)$ equals .025; and $P(t \geq -2.228)$ equals .025. The last row of the t table, for $n = \infty$, shows t values corresponding to the $\frac{x}{\sigma}$ values obtained from the area table of the normal curve (Table VI).

n	$P = .9$	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

* Reprinted from Table IV of R. A. Fisher, *Statistical Methods for Research Workers* (Oliver & Boyd, Ltd., Edinburgh), by kind permission of the author and publishers.

TABLE VIII.—TABLE OF χ^2 *

The column headings in this table are probabilities. The figures in the body of the table are values of χ^2 . The stub of the table shows n , the degrees of freedom. The probability of a value of χ^2 equal to or greater than the value specified is given. Thus for $n = 10$, $P(\chi^2 \geq 3.940)$ is .95; and accordingly by subtraction $P(\chi^2 \leq 3.940)$ is $1.00 - .95 = .05$.

n	$P = .99$	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.478	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

For larger values of n , the expression $\sqrt{2\chi^2} - \sqrt{2n - 1}$ may be used as a normal deviate with unit standard error.

* Reprinted from Table III of R. A. Fisher, *Statistical Methods for Research Workers* (Oliver & Boyd, Ltd., Edinburgh), by kind permission of the author and publishers.

TABLE IX.—TABLE OF F^*
(Values of F at 5 per cent points roman type, and at 1 per cent points boldface type.)

n_2	n_1 degrees of freedom (for greater mean square)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254	
	4.052	4.999	5.403	5.625	5.764	5.859	5.928	5.981	6.022	6.056	6.082	6.106	6.142	6.169	6.208	6.234	6.258	6.286	6.302	6.323	6.334	6.352	6.361	6.366	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50	
	98.49	99.01	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.50	99.50	99.50	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53	
	34.12	30.81	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63	
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67	
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23	
	12.25	9.55	8.45	7.85	7.45	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93	
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71	
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54	
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.62	2.58	2.55	2.52	2.49	2.47	2.45	2.42	2.40	
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60	
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30	
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36	
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21	
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.85	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16	

The values of F ($= \alpha$) were computed for this table in part from R. A. Fisher's table of values for the z distribution published in *Statistical Methods for Research Workers*, Table VI. Entries not so computed were found by interpolation, mostly graphical interpolation.

* Reprinted from George W. Snedecor, *Statistical Methods Applied to Experiments in Agriculture and Biology* (1937), pp. 184-187.

TABLE IX.—TABLE OF F^* —(Continued)

n_2	n_1 degrees of freedom (for greater mean square)																							n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	
14	4.60 8.86	3.74 6.61	3.24 5.66	3.11 5.03	2.96 4.69	2.85 4.46	2.77 4.28	2.70 4.14	2.65 4.03	2.60 3.94	2.56 3.86	2.53 3.80	2.48 3.70	2.44 3.62	2.39 3.51	2.35 3.34	2.31 3.34	2.27 3.26	2.24 3.21	2.21 3.14	2.19 3.11	2.16 3.06	2.14 3.02	2.13 3.00
15	4.54 8.68	3.68 6.56	3.29 5.42	3.06 4.89	2.90 4.56	2.79 4.32	2.70 4.14	2.64 4.00	2.59 3.89	2.55 3.80	2.51 3.73	2.48 3.67	2.43 3.56	2.39 3.48	2.33 3.36	2.25 3.29	2.21 3.20	2.18 3.07	2.15 2.97	2.12 2.90	2.10 2.86	2.07 2.82	2.05 2.78	2.03 2.81
16	4.49 8.53	3.63 6.33	3.24 5.29	3.01 4.77	2.85 4.44	2.74 4.20	2.66 4.03	2.59 3.89	2.54 3.78	2.49 3.69	2.45 3.61	2.42 3.55	2.37 3.45	2.33 3.37	2.28 3.25	2.24 3.18	2.20 3.10	2.16 3.01	2.13 2.96	2.09 2.89	2.07 2.86	2.04 2.80	2.02 2.77	2.01 2.75
17	4.45 8.40	3.59 6.11	3.20 5.18	2.96 4.67	2.80 4.34	2.70 4.10	2.62 3.93	2.55 3.79	2.50 3.68	2.45 3.59	2.41 3.52	2.38 3.45	2.33 3.35	2.29 3.27	2.23 3.16	2.19 3.08	2.15 3.00	2.11 2.92	2.08 2.86	2.04 2.79	2.02 2.76	1.99 2.70	1.97 2.67	1.96 2.65
18	4.41 8.28	3.55 6.01	3.16 5.09	2.93 4.56	2.77 4.25	2.66 4.01	2.58 3.85	2.51 3.71	2.46 3.60	2.41 3.51	2.37 3.44	2.34 3.37	2.29 3.27	2.25 3.19	2.19 3.07	2.15 3.00	2.11 2.91	2.07 2.83	2.04 2.78	2.00 2.71	1.98 2.68	1.95 2.62	1.93 2.59	1.92 2.57
19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50	2.74 4.17	2.63 3.94	2.55 3.77	2.48 3.63	2.43 3.52	2.38 3.43	2.34 3.36	2.31 3.30	2.26 3.19	2.21 3.12	2.15 3.00	2.11 2.92	2.07 2.84	2.02 2.75	1.99 2.69	1.96 2.63	1.94 2.58	1.91 2.47	1.88 2.44	1.87 2.42
20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43	2.71 4.10	2.60 3.87	2.52 3.71	2.45 3.56	2.40 3.45	2.35 3.37	2.31 3.30	2.28 3.23	2.23 3.13	2.18 3.05	2.12 2.94	2.08 2.86	2.04 2.77	1.99 2.69	1.96 2.63	1.92 2.56	1.90 2.53	1.87 2.47	1.85 2.44	1.84 2.42
21	4.32 8.02	3.47 5.78	3.07 4.87	2.84 4.37	2.68 4.04	2.57 3.81	2.49 3.65	2.42 3.51	2.37 3.40	2.32 3.31	2.28 3.24	2.25 3.17	2.20 3.07	2.15 2.99	2.09 2.88	2.05 2.80	2.00 2.72	1.96 2.68	1.93 2.58	1.89 2.53	1.87 2.47	1.84 2.42	1.82 2.38	1.81 2.36
22	4.30 7.94	3.44 5.72	3.05 4.82	2.82 4.31	2.66 3.99	2.55 3.76	2.47 3.59	2.40 3.45	2.35 3.35	2.30 3.26	2.26 3.18	2.23 3.12	2.18 3.02	2.13 2.94	2.07 2.83	2.03 2.75	1.98 2.67	1.93 2.58	1.91 2.53	1.87 2.46	1.84 2.42	1.81 2.37	1.80 2.33	1.78 2.31
23	4.28 7.88	3.42 5.66	3.03 4.76	2.80 4.26	2.64 3.94	2.53 3.71	2.45 3.54	2.38 3.48	2.32 3.30	2.28 3.21	2.24 3.14	2.20 3.07	2.15 2.97	2.10 2.89	2.04 2.78	2.00 2.70	1.96 2.62	1.91 2.53	1.88 2.48	1.84 2.41	1.82 2.37	1.79 2.32	1.77 2.28	1.76 2.26
24	4.26 7.82	3.40 5.61	3.01 4.72	2.78 4.22	2.62 3.90	2.51 3.67	2.43 3.50	2.36 3.36	2.30 3.25	2.26 3.17	2.22 3.09	2.18 3.03	2.13 2.93	2.09 2.85	2.02 2.74	1.98 2.66	1.94 2.58	1.89 2.49	1.86 2.44	1.82 2.38	1.80 2.33	1.76 2.27	1.74 2.23	1.73 2.21
25	4.24 7.77	3.38 5.57	2.99 4.68	2.76 4.18	2.60 3.86	2.49 3.63	2.41 3.46	2.34 3.32	2.28 3.21	2.24 3.13	2.20 3.05	2.16 2.99	2.11 2.89	2.06 2.81	2.00 2.70	1.96 2.62	1.92 2.54	1.87 2.45	1.84 2.41	1.80 2.32	1.77 2.29	1.74 2.23	1.72 2.19	1.71 2.17
26	4.22 7.72	3.37 5.53	2.98 4.64	2.74 4.14	2.59 3.82	2.47 3.69	2.39 3.43	2.32 3.29	2.27 3.17	2.22 3.09	2.18 3.02	2.15 2.96	2.10 2.86	2.05 2.77	1.99 2.66	1.95 2.58	1.90 2.50	1.85 2.41	1.82 2.36	1.78 2.28	1.76 2.25	1.72 2.19	1.69 2.15	1.68 2.13

TABLE IX.—TABLE OF F^* —(Continued)

n_2		n_1 degrees of freedom (for greater mean square)																				n_2		
		1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75		100	200
27	4.21 7.68	3.35 5.49	2.96 4.60	2.73 4.11	2.57 3.79	2.46 3.66	2.37 3.39	2.30 3.26	2.25 3.14	2.20 3.06	2.16 2.98	2.13 2.93	2.08 2.83	2.03 2.74	1.97 2.63	1.93 2.55	1.88 2.47	1.84 2.38	1.80 2.33	1.76 2.25	1.74 2.21	1.71 2.16	1.68 2.12	1.67 2.10
28	4.20 7.64	3.34 5.45	2.95 4.57	2.71 4.07	2.56 3.76	2.44 3.63	2.36 3.36	2.29 3.23	2.24 3.11	2.19 3.03	2.15 2.95	2.12 2.90	2.06 2.80	2.02 2.71	1.96 2.60	1.91 2.52	1.87 2.41	1.84 2.35	1.81 2.32	1.78 2.22	1.75 2.18	1.72 2.13	1.69 2.09	1.65 2.06
29	4.18 7.60	3.33 5.42	2.93 4.54	2.70 4.04	2.54 3.73	2.43 3.60	2.35 3.33	2.28 3.20	2.22 3.08	2.18 3.00	2.14 2.92	2.10 2.87	2.05 2.77	2.00 2.68	1.94 2.57	1.90 2.49	1.85 2.41	1.82 2.38	1.80 2.32	1.77 2.27	1.73 2.19	1.71 2.15	1.68 2.10	1.64 2.06
30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70	2.42 3.47	2.34 3.30	2.27 3.17	2.21 3.06	2.16 2.98	2.12 2.90	2.09 2.84	2.04 2.74	1.99 2.66	1.93 2.55	1.89 2.47	1.84 2.38	1.82 2.33	1.79 2.24	1.76 2.21	1.72 2.16	1.69 2.13	1.66 2.07	1.62 2.03
32	4.15 7.50	3.30 5.34	2.90 4.46	2.67 3.97	2.51 3.66	2.40 3.42	2.32 3.25	2.25 3.12	2.19 3.01	2.14 2.86	2.10 2.82	2.07 2.80	2.02 2.70	1.97 2.62	1.91 2.51	1.86 2.42	1.82 2.34	1.78 2.25	1.74 2.20	1.71 2.17	1.67 2.12	1.64 2.08	1.61 2.02	1.59 1.98
34	4.13 7.44	3.28 5.29	2.88 4.42	2.65 3.93	2.49 3.61	2.38 3.33	2.30 3.21	2.23 3.08	2.17 2.97	2.12 2.89	2.08 2.82	2.05 2.76	2.00 2.66	1.95 2.58	1.89 2.47	1.84 2.38	1.80 2.30	1.76 2.21	1.71 2.15	1.67 2.08	1.64 2.04	1.61 1.98	1.59 1.94	1.57 1.91
36	4.11 7.39	3.26 5.25	2.86 4.38	2.63 3.89	2.48 3.58	2.36 3.35	2.28 3.18	2.21 3.04	2.15 2.94	2.10 2.86	2.06 2.78	2.03 2.72	1.98 2.62	1.93 2.54	1.87 2.43	1.82 2.35	1.78 2.26	1.72 2.17	1.69 2.12	1.65 2.04	1.62 2.00	1.59 1.94	1.56 1.90	1.55 1.87
38	4.10 7.35	3.25 5.21	2.85 4.34	2.62 3.86	2.46 3.54	2.35 3.32	2.26 3.15	2.20 3.02	2.14 2.91	2.09 2.82	2.05 2.69	1.99 2.62	1.92 2.59	1.85 2.51	1.80 2.32	1.75 2.22	1.70 2.18	1.67 2.11	1.63 2.08	1.60 2.00	1.57 1.97	1.54 1.90	1.53 1.86	1.53 1.84
40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51	2.34 3.39	2.25 3.12	2.18 2.99	2.12 2.88	2.07 2.80	2.04 2.73	1.99 2.66	1.94 2.56	1.89 2.49	1.84 2.37	1.79 2.29	1.74 2.20	1.71 2.11	1.67 2.05	1.63 1.99	1.60 1.94	1.57 1.88	1.55 1.84	1.53 1.81
42	4.07 7.27	3.22 5.15	2.83 4.29	2.59 3.80	2.44 3.49	2.32 3.36	2.24 3.10	2.17 2.96	2.11 2.86	2.06 2.77	2.02 2.70	1.97 2.64	1.92 2.54	1.87 2.46	1.82 2.35	1.78 2.24	1.73 2.15	1.68 2.06	1.64 2.00	1.60 1.92	1.57 1.88	1.54 1.85	1.51 1.80	1.49 1.78
44	4.06 7.24	3.21 5.12	2.82 4.26	2.58 3.78	2.43 3.46	2.31 3.34	2.23 3.07	2.16 2.94	2.10 2.84	2.05 2.75	2.01 2.68	1.96 2.62	1.92 2.52	1.88 2.44	1.81 2.32	1.76 2.24	1.72 2.15	1.66 2.00	1.63 1.95	1.58 1.92	1.56 1.88	1.52 1.82	1.50 1.78	1.48 1.75
46	4.05 7.21	3.20 5.10	2.81 4.24	2.57 3.76	2.42 3.44	2.30 3.32	2.22 3.05	2.14 2.92	2.08 2.82	2.04 2.73	1.99 2.66	1.94 2.60	1.89 2.50	1.85 2.42	1.80 2.30	1.75 2.22	1.71 2.13	1.65 2.04	1.62 1.98	1.57 1.90	1.54 1.86	1.51 1.80	1.48 1.76	1.46 1.72
48	4.04 7.19	3.19 5.08	2.80 4.22	2.56 3.74	2.41 3.42	2.30 3.42	2.21 3.04	2.14 2.90	2.08 2.80	2.03 2.71	1.99 2.64	1.96 2.58	1.90 2.48	1.86 2.40	1.81 2.28	1.74 2.20	1.70 2.11	1.64 2.02	1.61 1.96	1.56 1.88	1.53 1.84	1.50 1.78	1.47 1.73	1.45 1.70

TABLE IX.—TABLE OF F^* —(Continued)

n_2	n_1 degrees of freedom (for greater mean square)																							n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	
50	4.03 7.17	3.18 5.06	2.79 4.20	2.56 3.72	2.40 3.41	2.29 3.18	2.20 3.02	2.13 2.88	2.07 2.78	2.02 2.70	1.98 2.62	1.95 2.56	1.90 2.46	1.85 2.39	1.78 2.26	1.74 2.18	1.69 2.10	1.63 2.00	1.60 1.94	1.55 1.86	1.52 1.83	1.48 1.76	1.46 1.71	1.44 1.68
55	4.02 7.12	3.17 5.01	2.78 4.16	2.54 3.68	2.38 3.37	2.27 3.15	2.18 2.98	2.11 2.85	2.05 2.75	2.00 2.66	1.97 2.63	1.93 2.59	1.88 2.53	1.83 2.43	1.85 2.35	1.76 2.23	1.72 2.06	1.67 2.01	1.61 1.96	1.58 1.90	1.52 1.82	1.46 1.71	1.43 1.66	1.41 1.64
60	4.00 7.08	3.15 4.98	2.76 4.13	2.52 3.65	2.37 3.34	2.25 3.12	2.17 2.95	2.10 2.82	2.04 2.72	1.99 2.68	1.95 2.56	1.92 2.50	1.86 2.40	1.81 2.32	1.81 2.20	1.75 2.12	1.65 2.03	1.59 1.93	1.56 1.87	1.50 1.81	1.48 1.74	1.44 1.68	1.41 1.63	1.39 1.60
65	3.99 7.04	3.14 4.95	2.75 4.10	2.51 3.62	2.36 3.31	2.24 3.09	2.15 2.93	2.08 2.79	2.02 2.70	1.98 2.61	1.94 2.54	1.90 2.47	1.85 2.37	1.80 2.30	1.73 2.18	1.68 2.09	1.63 2.00	1.57 1.90	1.54 1.84	1.49 1.76	1.46 1.71	1.42 1.64	1.39 1.60	1.37 1.56
70	3.98 7.01	3.13 4.92	2.74 4.08	2.50 3.60	2.35 3.29	2.23 3.07	2.14 2.91	2.07 2.77	2.01 2.67	1.97 2.59	1.93 2.51	1.89 2.45	1.84 2.35	1.79 2.28	1.72 2.15	1.67 2.07	1.62 1.98	1.56 1.88	1.53 1.82	1.47 1.74	1.45 1.69	1.40 1.62	1.37 1.56	1.35 1.53
80	3.96 6.96	3.11 4.88	2.72 4.04	2.48 3.56	2.33 3.25	2.21 3.04	2.12 2.87	2.05 2.74	1.99 2.64	1.95 2.55	1.91 2.48	1.88 2.41	1.82 2.32	1.77 2.24	1.70 2.11	1.65 2.03	1.60 1.94	1.54 1.84	1.51 1.78	1.45 1.70	1.42 1.65	1.38 1.57	1.35 1.53	1.32 1.49
100	3.94 6.90	3.09 4.82	2.70 4.02	2.46 3.51	2.30 3.20	2.19 2.99	2.10 2.82	2.03 2.69	1.97 2.59	1.92 2.51	1.88 2.43	1.85 2.36	1.79 2.26	1.75 2.19	1.68 2.06	1.63 1.98	1.57 1.89	1.51 1.79	1.48 1.73	1.42 1.64	1.39 1.61	1.34 1.59	1.30 1.46	1.28 1.43
125	3.92 6.84	3.07 4.78	2.68 3.94	2.44 3.47	2.29 3.17	2.17 2.95	2.08 2.79	2.01 2.65	1.95 2.56	1.90 2.47	1.86 2.40	1.83 2.33	1.77 2.23	1.72 2.15	1.65 2.03	1.59 1.94	1.55 1.85	1.49 1.75	1.45 1.70	1.39 1.65	1.36 1.54	1.31 1.46	1.27 1.40	1.25 1.37
150	3.91 6.81	3.06 4.76	2.67 3.91	2.43 3.44	2.27 3.14	2.16 2.92	2.07 2.76	2.00 2.62	1.94 2.53	1.89 2.44	1.85 2.37	1.82 2.30	1.76 2.20	1.71 2.13	1.64 2.00	1.59 1.91	1.54 1.83	1.47 1.72	1.44 1.66	1.37 1.51	1.34 1.49	1.29 1.43	1.25 1.37	1.22 1.33
200	3.89 6.76	3.04 4.71	2.65 3.88	2.41 3.41	2.26 3.11	2.14 2.90	2.05 2.73	1.98 2.60	1.92 2.50	1.87 2.41	1.83 2.34	1.80 2.28	1.74 2.17	1.69 2.09	1.62 1.97	1.57 1.88	1.52 1.79	1.45 1.69	1.42 1.62	1.35 1.48	1.32 1.43	1.26 1.39	1.22 1.33	1.19 1.28
400	3.86 6.70	3.02 4.66	2.62 3.83	2.39 3.36	2.23 3.06	2.12 2.85	2.03 2.69	1.96 2.55	1.90 2.49	1.85 2.37	1.81 2.33	1.78 2.23	1.72 2.13	1.67 2.04	1.60 1.92	1.54 1.84	1.49 1.74	1.42 1.64	1.38 1.61	1.32 1.47	1.28 1.42	1.22 1.32	1.16 1.24	1.13 1.19
1000	3.85 6.66	3.00 4.62	2.61 3.80	2.38 3.34	2.22 3.04	2.10 2.82	2.02 2.65	1.95 2.53	1.89 2.43	1.84 2.34	1.80 2.26	1.76 2.09	1.70 2.09	1.65 2.01	1.58 1.89	1.53 1.81	1.47 1.71	1.36 1.61	1.30 1.44	1.26 1.44	1.19 1.38	1.13 1.28	1.08 1.19	1.00 1.11
∞	3.84 6.64	2.99 4.60	2.60 3.78	2.37 3.32	2.21 3.02	2.09 2.80	2.01 2.64	1.94 2.51	1.88 2.41	1.83 2.32	1.79 2.24	1.75 2.18	1.69 2.07	1.64 1.99	1.57 1.87	1.52 1.79	1.46 1.69	1.40 1.59	1.35 1.59	1.28 1.41	1.24 1.35	1.17 1.25	1.11 1.15	1.00 1.00

TABLE X.—5 PER CENT AND 10 PER CENT POINTS OF THE SAMPLING DISTRIBUTION OF $\sqrt{\beta_1}$ FOR SAMPLES OF VARIOUS SIZES¹
(Approximate values)

Size of sample	Values of $\sqrt{\beta_1}$ for $P(\sqrt{\beta_1} \geq)$	
	.05	.10
25	.711	1.061
30	.661	.982
35	.621	.921
40	.587	.869
45	.558	.825
50	.533	.787
60	.492	.723
70	.459	.673
80	.432	.631
90	.409	.596
100	.389	.567

¹ Reproduced from P. Williams, "Note on the Sampling Distribution of $\sqrt{\beta_1}$, Where the Population Is Normal," *Biometrika*, Vol. 27 (1935), pp. 269-271.

TABLE XI.—LOWER AND UPPER 1 PER CENT AND 5 PER CENT POINTS OF THE SAMPLING DISTRIBUTION OF β_2 FOR SAMPLES OF VARIOUS SIZES¹
(Approximate values)

Size of sample	Values of β_2			
	For $P(\beta_2 \leq)$		For $P(\beta_2 \geq)$	
	.01	.05	.05	.01
100	2.18	2.35	3.77	4.39
125	2.24	2.40	3.70	4.24
150	2.29	2.45	3.65	4.14
175	2.33	2.48	3.61	4.05
200	2.37	2.51	3.57	3.98
250	2.42	2.55	3.52	3.87
300	2.46	2.59	3.47	3.79
400	2.52	2.64	3.41	3.67
500	2.57	2.67	3.37	3.60
800	2.65	2.74	3.29	3.46
1,000	2.68	2.76	3.26	3.41
2,000	2.77	2.83	3.18	3.28
5,000	2.85	2.89	3.12	3.17

¹ Abridged from E. S. Pearson, "A Further Development of Tests for Normality," *Biometrika*, Vol. 22 (1930-1931), pp. 239-249. Table assumes population to be normal.

TABLE XII.—LOWER AND UPPER 1 PER CENT, 5 PER CENT, AND 10 PER CENT POINTS OF SAMPLING DISTRIBUTION OF $a = A.D./\sigma$ FOR SAMPLES OF VARIOUS SIZES¹
(Approximate values)

	Values of a					
	For $P(a \leq)$			For $P(a \geq)$		
	.01	.05	.10	.01	.05	.10
10	.6675	.7153	.7409	.8899	.9073	.9359
15	.6829	.7236	.7452	.8733	.8884	.9137
20	.6950	.7304	.7495	.8631	.8768	.9001
25	.7040	.7360	.7530	.8570	.8686	.8901
30	.7110	.7404	.7559	.8511	.8625	.8827
35	.7167	.7440	.7583	.8468	.8578	.8769
40	.7216	.7470	.7604	.8436	.8540	.8722
45	.7256	.7496	.7621	.8409	.8508	.8682
50	.7291	.7518	.7636	.8385	.8481	.8648
60	.7347	.7554	.7662	.8349	.8434	.8592
70	.7393	.7583	.7683	.8321	.8403	.8549
80	.7430	.7607	.7700	.8298	.8376	.8515
90	.7460	.7626	.7714	.8279	.8353	.8484
100	.7487	.7644	.7726	.8264	.8344	.8460
200	.7629	.7738	.7796	.8178	.8229	.8322
300	.7693	.7781	.7828	.8140	.8183	.8260

¹ Abridged from R. C. Geary, "Moments of the Ratio of the Mean Deviation to the Standard Deviation for Normal Samples," *Biometrika*, Vol. 28 (1936), pp. 295-307.

* The size of sample is N , and $n = N - 1$.

TABLE XIII.—SAMPLING DISTRIBUTION OF THE RANGE $w = \frac{X_n - X_1}{\sigma}$ *

N	Mean w	$a_n \dagger$	σ_w	Lower tail of distribution $P(w \leq)$						Upper tail of distribution $P(w \geq)$					
				.001	.005	.010	.025	.050	.100	.100	.050	.025	.010	.005	.001
2	1.128	.886	.853	.00	.01	.02	.04	.09	.18	2.33	2.77	3.17	3.64	3.97	4.65
3	1.693	.591	.888	.06	.13	.19	.30	.43	.62	2.90	3.31	3.68	4.12	4.42	5.06
4	2.059	.486	.880	.20	.34	.43	.59	.76	.98	3.24	3.63	3.98	4.40	4.69	5.31
5	2.326	.430	.864	.37	.55	.66	.85	1.03	1.26	3.48	3.86	4.20	4.60	4.89	5.48
6	2.534	.395	.848	.54	.75	.87	1.06	1.25	1.49	3.66	4.03	4.36	4.76	5.03	5.62
7	2.704	.370	.833	.69	.92	1.05	1.25	1.44	1.68	3.81	4.17	4.49	4.88	5.15	5.73
8	2.847	.351	.820	.83	1.08	1.20	1.41	1.60	1.83	3.93	4.29	4.61	4.99	5.26	5.82
9	2.970	.337	.808	.96	1.21	1.34	1.55	1.74	1.97	4.04	4.39	4.70	5.08	5.34	5.90
10	3.078	.325	.797	1.08	1.33	1.47	1.67	1.86	2.09	4.13	4.47	4.79	5.16	5.42	5.97
11	3.173	.315	.787	1.20	1.45	1.58	1.78	1.97	2.20	4.21	4.55	4.86	5.23	5.49	6.04
12	3.258	.307	.778	1.30	1.55	1.68	1.88	2.07	2.30	4.29	4.62	4.92	5.29	5.54	6.09
13	3.336	.300	.770	1.38	1.64	1.77	1.97	2.16	2.39	4.35	4.69	4.99	5.35	5.60	6.15
14	3.407	.294	.762	1.47	1.72	1.86	2.06	2.24	2.47	4.41	4.74	5.04	5.40	5.65	6.20
15	3.472	.288	.755	1.55	1.80	1.93	2.14	2.32	2.54	4.47	4.80	5.09	5.45	5.70	6.24
16	3.532	.283	.749	1.63	1.88	2.01	2.21	2.39	2.61	4.52	4.85	5.14	5.49	5.74	6.28
17	3.588	.279	.743	1.69	1.94	2.07	2.27	2.45	2.67	4.57	4.89	5.18	5.54	5.79	6.31
18	3.640	.275	.738	1.75	2.01	2.14	2.34	2.51	2.73	4.61	4.93	5.22	5.58	5.82	6.35
19	3.689	.271	.733	1.82	2.07	2.20	2.39	2.57	2.79	4.65	4.97	5.26	5.61	5.86	6.38
20	3.735	.268	.729	1.88	2.13	2.25	2.45	2.63	2.84	4.69	5.01	5.30	5.65	5.89	6.41

* Reproduced by permission from E. S. Pearson's "The Probability Integral of the Range in Samples of N Observations from a Normal Population," *Biometrika*, Vol. 32 (1941-1942), pp. 301-308; and "The Percentage Limits for the Distribution of Range in Samples from a Normal Population," *ibid.*, Vol. 24 (1932), pp. 404-417. Table 2 in *Biometrika*, Vol. 32, (1941-1942), p. 308, has been extended from $N = 12$ to $N = 20$ by interpolating for the necessary values in Table 1, pp. 302-307, of the same article. The values for the mean w and the standard deviation of w were obtained from a table in *Biometrika*, Vol. 24 (1932), p. 416. It is to be noted that X_n refers to the largest and X_1 to the smallest X in the sample.

In 1925, tables giving the expected or mean value and the standard deviation of range in random samples from a normal population were calculated by L. H. C. Tippett, Department of Applied Statistics, University College, London. Since the probability distribution $f_n(w)$ is itself far from normal in form, it was evident that its mean and standard deviation alone would not provide all the information generally needed in practice. Tippett included in his paper some values of the constants β_1 and β_2 of the distribution, and his work was extended by E. S. Pearson in 1926 and 1932. E. S. Pearson also developed an approximate method of determining probability levels for w and provided some provisional tables of these. Finally, in 1942, a full and accurate table of the probability integral of the range was completed and published in the article cited below. The actual method of computation was planned by H. O. Hartley, and the calculations were carried out under his supervision by Scientific Computing Service, Ltd. The scope of the main table, like that of Table XIII, was limited to $N = 20$. As N increases beyond this value, there is an increasing risk that the table may be misleading in practice, since $f_n(w)$ becomes very sensitive to relatively slight departures from normality in the tails of the population distribution. Cf. E. S. Pearson, "The Probability Integral of the Range in Samples of n Observations from a Normal Population," *Biometrika*, Vol. 32 (1941-1942), p. 308.

TABLE XIV.—HYPERBOLIC TANGENTS¹

z	$r = \tanh z$	z	$r = \tanh z$	z	$r = \tanh z$
0.00	0.00000	0.55	0.50052	1.10	0.80050
0.10	.01000	0.56	.50798	1.11	.80406
0.02	.02000	0.57	.51536	1.12	.80757
0.03	.02999	0.58	.52267	1.13	.81102
0.04	.03998	0.59	.52990	1.14	.81441
0.05	0.04996	0.60	0.53705	1.15	0.81775
0.06	.05993	0.61	.54413	1.16	.82104
0.07	.06989	0.62	.55113	1.17	.82427
0.08	.07983	0.63	.55805	1.18	.82745
0.09	.08976	0.64	.56490	1.19	.83058
0.10	0.09967	0.65	0.57167	1.20	0.83365
0.11	.10956	0.66	.57836	1.21	.83668
0.12	.11943	0.67	.58498	1.22	.83965
0.13	.12927	0.68	.59152	1.23	.84258
0.14	.13909	0.69	.59798	1.24	.84546
0.15	0.14889	0.70	0.60437	1.25	0.84828
0.16	.15865	0.71	.61068	1.26	.85106
0.17	.16838	0.72	.61691	1.27	.85380
0.18	.17808	0.73	.62307	1.28	.85648
0.19	.18775	0.74	.62915	1.29	.85913
0.20	0.19738	0.75	0.63515	1.30	0.86172
0.21	.20697	0.76	.64108	1.31	.86428
0.22	.21652	0.77	.64693	1.32	.86678
0.23	.22603	0.78	.65271	1.33	.86925
0.24	.23550	0.79	.65841	1.34	.87167
0.25	0.24492	0.80	0.66404	1.35	0.87405
0.26	.25430	0.81	.66959	1.36	.87639
0.27	.26362	0.82	.67607	1.37	.87869
0.28	.27291	0.83	.68048	1.38	.88095
0.29	.28213	0.84	.68581	1.39	.88317
0.30	0.29131	0.85	0.69107	1.40	0.88535
0.31	.30044	0.86	.69626	1.41	.88749
0.32	.30951	0.87	.70137	1.42	.88960
0.33	.31852	0.88	.70642	1.43	.89167
0.34	.32748	0.89	.71139	1.44	.89370
0.35	0.33638	0.90	0.71630	1.45	0.89569
0.36	.34521	0.91	.72113	1.46	.89765
0.37	.35399	0.92	.72590	1.47	.89958
0.38	.36271	0.93	.73059	1.48	.90147
0.39	.37136	0.94	.73522	1.49	.90332
0.40	0.37995	0.95	0.73978	1.50	0.90515
0.41	.38847	0.96	.74428	1.51	.90694
0.42	.39693	0.97	.74870	1.52	.90870
0.43	.40532	0.98	.75307	1.53	.91042
0.44	.41364	0.99	.75736	1.54	.91212
0.45	0.42190	1.00	0.76159	1.55	0.91379
0.46	.43008	1.01	.76576	1.56	.91542
0.47	.43820	1.02	.76987	1.57	.91703
0.48	.44624	1.03	.77391	1.58	.91860
0.49	.45422	1.04	.77789	1.59	.92015
0.50	0.46212	1.05	0.78181	1.60	0.92167
0.51	.46995	1.06	.78566	1.61	.92316
0.52	.47770	1.07	.78946	1.62	.92462
0.53	.48538	1.08	.79320	1.63	.92606
0.54	.49299	1.09	.79688	1.64	.92747

¹ Source: HODGMAN, CHARLES C., *Mathematical Tables from Handbook of Chemistry and Physics* (1941).

TABLE XIV.—HYPERBOLIC TANGENTS.—(Continued)

z	$r = \tanh z$	z	$r = \tanh z$	z	$r = \tanh z$
1.65	0.92886	2.20	0.97574	2.75	0.99186
1.66	.93022	2.21	.97622	2.76	.99202
1.67	.93155	2.22	.97668	2.77	.99218
1.68	.93286	2.23	.97714	2.78	.99233
1.69	.93415	2.24	.97759	2.79	.99248
1.70	0.93541	2.25	0.97803	2.80	0.99263
1.71	.93665	2.26	.97846	2.81	.99278
1.72	.93783	2.27	.97888	2.82	.99292
1.73	.93906	2.28	.97929	2.83	.99306
1.74	.94023	2.29	.97970	2.84	.99320
1.75	0.94138	2.30	0.98010	2.85	0.99333
1.76	.94250	2.31	.98049	2.86	.99346
1.77	.94361	2.32	.98087	2.87	.99359
1.78	.94470	2.33	.98124	2.88	.99372
1.79	.94576	2.34	.98161	2.89	.99384
1.80	0.94681	2.35	0.98197	2.90	0.99396
1.81	.94783	2.36	.98233	2.91	.99408
1.82	.94884	2.37	.98267	2.92	.99420
1.83	.94983	2.38	.98301	2.93	.99431
1.84	.95080	2.39	.98335	2.94	.99443
1.85	0.95175	2.40	0.98367	2.95	0.99454
1.86	.95268	2.41	.98400	2.96	.99464
1.87	.95359	2.42	.98431	2.97	.99475
1.88	.95449	2.43	.98462	2.98	.99485
1.89	.95537	2.44	.98492	2.99	.99496
1.90	0.95624	2.45	0.98522	3.0	0.99505
1.91	.95709	2.46	.98551	3.1	.99595
1.92	.95792	2.47	.98579	3.2	.99668
1.93	.95873	2.48	.98607	3.3	.99728
1.94	.95953	2.49	.98635	3.4	.99777
1.95	0.96032	2.50	0.98661	3.5	0.99818
1.96	.96109	2.51	.98688	3.6	.99851
1.97	.96185	2.52	.98714	3.7	.99878
1.98	.96259	2.53	.98739	3.8	.99900
1.99	.96331	2.54	.98764	3.9	.99918
2.00	0.96403	2.55	0.98788	4.0	0.99933
2.01	.96473	2.56	.98812	4.1	.99945
2.02	.96541	2.57	.98835	4.2	.99955
2.03	.96609	2.58	.98858	4.3	.99963
2.04	.96675	2.59	.98881	4.4	.99970
2.05	0.96740	2.60	0.98903	4.5	0.99975
2.06	.96803	2.61	.98924	4.6	.99980
2.07	.96865	2.62	.98946	4.7	.99983
2.08	.96926	2.63	.98966	4.8	.99986
2.09	.96986	2.64	.98987	4.9	.99989
2.10	0.97045	2.65	0.99007	5.0	0.99991
2.11	.97103	2.66	.99026		
2.12	.97159	2.67	.99045		
2.13	.97215	2.68	.99064		
2.14	.97269	2.69	.99083		
2.15	0.97323	2.70	0.99101		
2.16	.97375	2.71	.99118		
2.17	.97426	2.72	.99136		
2.18	.97477	2.73	.99153		
2.19	.97520	2.74	.99170		

AUTHOR INDEX

B

Baker, G. A., 450
 Bienaymé, J., 454
 Bowley, A. L., 211

C

Charlier, C. V. L., 82, 90
 Cheshire, L. E. O., 451
 Church, A. E. R., 445, 448
 Courant, R., 70
 Craig, C. C., 445, 446
 Czuber, E., 39

D

Davenport, D. H., 223
 De Tchénychef, P. L., 454
 Deming, W. E., 162
 Dudding, B. P., 296

E

Elderton, W. P., 50, 59, 128, 131,
 134, 135, 148

F

Fisher, Arne, 92, 454
 Fisher, R. A., 11, 114, 162, 182, 211,
 242, 298, 305, 442, 474, 475
 Friedman, M., 453

G

Gallup, G., 393
 Geary, R. C., 244, 245, 481
 Goulden, C. H., 10, 11, 442
 Gram, J. P., 82

H

Hansen, M. H., 162
 Hartley, H. O., 242, 482
 Hodgman, C. D., 469-473, 483-484
 Hotelling, H., 452
 Huntington, E. V., 457-561

I

Irwin, J. O., 114, 430

J

Jenson, A., 184

K

Kendall, M. G., 155, 158, 160, 161
 Kenney, J. F., 255
 Khanikof, N. de, 454

L

Le Roux, J. M., 447, 450
 Lexis, W., 422

M

Markov, A., 103
 Mills, F. C., 223

N

Neyman, J., 345, 362, 409, 414, 415

P

Pabst, Margaret R., 452
 Pearson, E. S., 242, 244, 296, 345,
 362, 409, 414, 415, 449, 450, 451,
 480, 482

- Pearson, Karl, 50, 55, 58, 64, 74, 134, 148, 158, 213, 303, 366
 Perlo, V., 447
 Pitman, E. J. G., 412
 Pizzetti, P., 454
- R
- Rae, S. F., 393
 Réitz, H. L., 90, 103, 422, 445, 447, 448, 454
 Rider, P. R., 114, 411, 442, 446, 447, 449
 Robinson, G., 93
- S
- Sheppard, W. F., 132, 142
 Shewhart, W. A., 449
 Smith, B. B., 155, 158, 160
- Snedecor, G. W., 476-479
 Sophister, 447
 Stephan, F. F., 162
- T
- Tchébychef, P. L. de, 454
 Thiele, T. N., 92
 Tippett, L. H. C., 242, 482
- W
- Waugh, A. E., 462, 463-464, 465-466, 467-468
 Whittaker, E. T., 93
 Williams, P., 244, 480
 Winters, F. W., 449
- Y
- Yule, G. U., 155, 159, 161, 422

SUBJECT INDEX

A

Analysis of variance, chance variation, measure of, more than one basis of classification, more than one case in each class, 435-436
 single case in each class, 434
 one basis of classification, 434
F distribution, use of, more than one basis of classification, more than one case in each class, 437-438, 441-442
 one case in each class, 431, 433
 selection of region of rejection, 428
 single basis of classification, 425
 testing correlation coefficients, etc., 443
 testing linearity, 444
 "interaction," 435-436
 more than one basis of classification, more than one case in each class, analysis of variance table, 441
 numerical analysis, 438-442
 study of results, 441-442
 theoretical basis, 435-438
 single case in each class, analysis of variance table, 433
 comparison of results with those of single basis, 434
 numerical analysis, 431-433
 theoretical basis, 429-431
 nonnormal population, 449-450
 remainder variance, more than one case in each class, 435-436
 one case in each class, 429-433

Analysis of variance, single basis of classification, numerical analysis, 425-428
 theoretical basis, 423-425
 worksheet for calculating sums of squares, 427
 tests of correlation coefficients as, 442-444
 tests of linearity, 444
 Asymmetrical binomial distribution, betas, formulas for, 45
 derivation, 67
 characteristics, 44-46
 derivation, 40-44
 as distribution of sample percentages, 190
 effect of changing N , 49
 formula, 43
 graphs, 44, 45
 mean, formula for, 45
 derivation, 65-66
 mode, formula for, 45
 derivation, 68
 and the normal curve, 46-47, 68-74
 numerical examples, 42, 43
 and Pearson's type III curve, 47-50
 relative slope, 76-77
 second approximation, 72-73
 standard deviation, formula for, 45
 derivation, 66-67
 Average deviation, 12-13
 Averages (*see* individual titles such as: Mean, Median, and Mode)

B

Beta coefficients, β_1 and β_2 , 10-11
 sampling distribution of $\sqrt{\beta_1}$, 242-244

Beta coefficients, sampling distribution of $\sqrt{\beta_1}$, table of .05 and .10 points, 480
 sampling distribution of β_2 , 242-245
 table of .01 and .05 points, 480
 standard error of $\sqrt{\beta_1}$, 242, 244
 standard error of β_2 , 242, 244
 Binomial distribution (*see* Symmetrical binomial distribution; Asymmetrical binomial distribution)
 Binomial expansion, 25-26
 Bivariate frequency distribution, 13-22
 Boldfaced type, 10
 Breve, 10

C

Calculus of Observations, The, 93
 Camp, B. H., 454
 Chi square (χ^2) distribution (*see* Frequency curves, chi square (χ^2) distribution)
 Chi square test (*see* Frequency curves, testing goodness of fit, by χ^2 test)
 Coefficient, of correlation (*see* Correlation, coefficient of)
 of multiple correlation (*see* Multiple correlation, coefficient of)
 of risk (*see* Statistical inference, testing hypotheses, coefficient of risk)
 Combinations, 24-25
 Confidence coefficient (*see* Statistical inference, confidence intervals, confidence coefficient)
 Confidence intervals (*see* Statistical inference, confidence intervals)
 Contingency tables, 337, 422*n*.
 (*See also* Independence, test of)
 two-fold classification, 395-397
 Correlation, coefficient of, Pearsonian, 19
 and the breakup of variance, 20-21
 confidence limits for, 301-302

Correlation, coefficient of, Pearsonian, maximum likelihood estimate of, 302-303
 as measure of goodness of fit, 20
 as measure of proportion of total variance, 21
 relationship to first-order variance, 19-20
 sampling distribution, 298-300
 significance, 21
 testing hypotheses about, 300-301
 Correlation index, testing significance of, 306
 Correlation ratio, 21-22
 testing significance of, 306
 Cumulants, of contributory causes, relationship to cumulants of variable, 83, 94-96
 definition of, 83
 as semi-invariants, 83

D

Danish census (1923), 184
 Deciles, 8
 Degrees of freedom, in analysis of variance, 424, 434, 441
 in a contingency table, 337
 explained, 314, 318-319
 fitting a frequency curve, 332-333
 identified with n , 327-328
 Density of samples, in distribution of all possible samples, 256-258
 Dependent variable, estimation of, from sample line or plane of regression, 387-388
 with allowance for sampling errors, 388-389
 Design of experiments, 162
 Dispersion, subnormal and super-normal, 422*n*.

E

Elderton, W. P., 128, 134, 148
 Ezekiel, M., 305

F

F distribution (*see* Frequency curves, *F* distribution)

- First-order standard deviation, 18–19 (*See also* Higher-order variances)
 definition, 18
 relationship to r , 19–20
- Fisher, Arne, 92
- Fisher, R. A., 11, 114*n.*, 242, 299
- Flat space, 253
- Frequency curves, chi square (χ^2)
 distribution, description, 111–112
 formula, 111
 graphs, 112
 mean, 112
 mode, 112
 probabilities, table of, 475
 standard deviation, 112
 explanation of, 3–5
- F distribution, description, 112–114
 formula, 113
 graph, 113
 mean, 113
 mode, 113
 probabilities, table of, 476–479
 standard deviation, 113
- fitting of, Gram-Charlier curves, 133–134
 nonnormal curves, 132–137
 normal curve, 131–132
 Pearsonian curves, 134–137
 sampling curves, 137
 Sheppard's corrections, 10, 12, 131–132, 134
- Gram-Charlier (*see* Gram-Charlier frequency curves)
 graph, 4
 graphing of, labeling of vertical scale, 110*n.*
 normal curve, 115–116
 other curves, 118–119
 t , χ^2 and F curves, 116–117
- nonnormal, examples from everyday life, 105–106
 examples from sampling analysis, 107–114
- normal (*see* Normal frequency curves)
- Frequency curves, Pearsonian (*see* Pearsonian frequency curves)
 probabilities, computation of, 119–120
 for χ^2 curve, 125–126
 for F curve, 127
 for normal curve, 120–123
 for other curves, 127–131
 for t curve, 123–124
 by quadrature formulas, 128
 as probability distributions, 28
 sampling distributions (*see* Sampling distribution)
- t distribution, approximation by normal curve for n between 30 and 100, 401*n.*
 description, 109–110
 formula, 111
 graph, 110
 kurtosis, 111
 mean, 111
 probabilities, table of, 474
 variance, 111
- testing goodness of fit, 331–333
 by χ^2 test, Gram-Charlier curves, 144–145
 normal curve, 139–142
 Pearsonian curves, 151, 152
 by graphic comparison, Gram-Charlier curves, 142–144
 normal curve, 137–139
 Pearsonian curves, 146–151
- theory of, conditions leading to nonnormality, 100–105
 conditions leading to normality, 37–39, 100
 contributory causes, Gram-Charlier curves, 83–84, 90–92, 93–98
 nonnormal curves in general, 100–103
 normal curve, 35–39, 100
 Pearsonian curves, 60–65
 sampling distribution of the mean, 107–108
 summary, 100–105
 z distribution, 114*n.*

Frequency Curves and Correlation,
134, 148
Frequency distributions, 1-5
 bivariate, 13-14
 illustration of, 13
Frequency series, continuous, 1-2
 discrete, 1
Fourier integral theorem, 93, 97

G

Galton, Francis, 18
Gamma function, 78-79*n.*, 254-255
Goodness of fit, of frequency curves
 (see Frequency curves, testing
 goodness of fit)
Gram-Charlier frequency curves,
 assumptions, 82-83
 comparison with Pearsonian
 curves, 90-92
 comparison with Pearsonian type
 III curve, 82
 components of, 85-90
 combinations of, 86, 87, 88, 89
 graphs, 85, 87
 derivation, 82-85, 92-99
 formula for type A, 84
 general significance, 90-92
 probabilities, computation of, 145-
 146
 relationship to asymmetrical bi-
 nomial distribution, 91-92
 type B, 90
g statistics, 11-12, 243
 sampling distributions, 242-243
 standard errors, 243
Guldberg, M. A., 454

H

Higher-order variances, confidence
 limits, 386-387
 definition, 18
 maximum-likelihood estimates,
 387
 sampling distribution, 385-386
 testing hypotheses about, 386
 use in estimating dependent vari-
 ables, 387-390

Histogram, 2-3
 graph, 2
 of relative frequencies, graph, 4
Homogeneity, meaning, 101-103
 test of, 338-339
Hotelling, H., 452
Hyperbolic tangents (table), 483-
 484
Hypergeometrical distribution, char-
 acteristics, 55
 criterion for, 81
 derivation, 51-55
 as distribution of sample per-
 centages, 210-211
 formula, 54
 graph, 55
 mean, 55
 moments, 56
 numerical example, 53
 and the Pearsonian curves, 57-58
 relative slope, 79-81
Hypergeometrical series, 54*n.*
Hyperplane, 253
Hypersphere, 254
Hypotheses, testing of (see Statisti-
 cal inference, testing hypotheses)

I

Incomplete gamma function, 255
Independence, meaning of, 334-335
 test of, distribution of

$$\sum \frac{(N_i - Np_i)^2}{Np_i}$$

 use of, 337-338
 estimating population percent-
 ages, 335-337
 the null hypothesis, 334
 the problem, 333-334

J

*Journal of the Royal Statistical
 Society*, 159

K

Kendall, M. G., 159, 160
k statistics, 11-12, 242-243
Kurtosis, 11

L

- Law, of large numbers, 27-28
 - of small chances, 212
 - of small numbers, 212
- Least squares, method of, 15, 374
 - minimizing horizontal deviations; illustrated, 16
 - minimizing vertical deviations, illustrated, 15
- Leptokurtic distributions, 11
- Lexis' analysis, 422*n*.
- Likelihood ratio, 345
- Linear function, sampling distribution, 108-109
- Lines of regression, 16-18
- Literary Digest* poll, 157
- Logarithms, of factorials, 117
 - or numbers, four-place common (table), 457-460
 - representation by an infinite series, 70-71*n*.

M

- Mathematical Statistics*, 92
- Mathematical Theory of Probabilities*, 92
- Maximum-likelihood estimates (*see* Statistical inference, maximum likelihood estimates of population parameters)
- Mean, arithmetic, calculation, 6-7
 - definition, 5-6
 - joint sampling distribution of, mean and standard deviation, 340-344
 - sampling distribution, any population, 107-108
 - normal population, 231, 262-263
 - special nonnormal populations, 445-446
 - standard error, 231, 268
 - geometric, 9
 - harmonic, 9
- Mean value, of a sum or difference, formula, 392
 - proof of formula, 419-420

- Means, progressions of, 16
 - charts, 14, 15
- Median, definition, 7
 - sampling distribution, 241-242
 - standard error, 242
- Meidel, M. B., 454
- Methods of Correlation Analysis*, 305
- Mode, 8
- Moment coefficients, 9
- Moment generating function, 93-94
- Moments, 9
- Most powerful test, 198*n*.
- Multinomial distribution, 309-323
 - (*See also* Random sampling, from a discrete manifold population, distribution of sample percentages, and testing hypotheses about p 's)
- Multinomial expansion, 26
- Multiple correlation, coefficient of, confidence limits for, 305
 - maximum-likelihood estimate, 305-306
 - testing significance of, 305
- Multivariate frequency distribution, 13-14

N

- Narumi, S., 454
- N -dimensional geometry, 253-254
- Neyman, J., 353, 362
- Nonnormality, problem of, 445-455
 - indirect attacks on, 451-455
 - qualitative or semiquantitative methods, 451-453
 - Tchébycheff's and other inequalities, 453-455
 - transformation of the data, 451
- sampling distributions of various statistics for specific nonnormal populations, 445-448
- use of normal theory for nonnormal populations, in case of analysis of variance, 449-450
 - in case of correlation coefficients, 451

- Nonnormality, use of normal theory
 for nonnormal populations,
 in case of means, 448-449
 in case of $\sqrt{N}(\bar{X}-\bar{X})/\bar{s}$, 449
 in case of variances, 450
- Normal frequency curve, areas under
 (table), 469-473
 characteristics, 10, 12
 conditions leading to, 37-39
 derivation, 34-35
 derivatives of (table of φ s and
 φ_1), 469-473
 first approximation to binomial
 distribution, 73
 fitting of, 131-132
 formula, 35
 graph, 121
 graphing of, 115-116
 ordinates of (table), 469-473
 probabilities, computation of, 123-
 124
 significance, 35-37
 standard form, 137-139
 testing goodness of fit, by χ^2 test,
 139-142
 by graphic comparison, 137-139
 truncated, 104-105
- Normality, determining departure
 from, by fitting a normal
 curve, 131-132, 137-142
 by special statistics, 242-245,
 296-297
- Null hypothesis, in analysis of
 variance, 423-424, 430-431, 436
 in comparing two samples, 391-
 392
 in testing difference between two
 sample means, 398
 percentages, 393
 variances, 407-408
 in testing independence, 334
- P
- Pabst, Margaret R., 452
- Parameters, 5, 10
- Partial correlation, coefficient of,
 inferences about, 303-304
- Pearson, E. S., 353, 362
- Pearson, Karl, 57, 58, 64, 134, 148,
 158, 255
- Pearsonian coefficient of correlation
 (see Correlation, coefficient of,
 Pearsonian)
- Pearsonian frequency curves, equa-
 tions for, determination of,
 146-151
 as an explanation of nonnormal
 curves, 65
 general formula, 134
 and the hypergeometrical dis-
 tribution, 57-58
 computation of probabilities, 151,
 152
 relative slope, formula for, 57
 types, chart for distinguishing, 136
 criterion for distinguishing, 58,
 135
 main, 58-60
 table of, 135
 transitional, 60
 type I, graph, 59
 type III } derivation, 48-50,
 76-79,
 formula, 49
 graph, 49
 type IV, graph, 59
- Percentage, sampling distribution,
 190-194, 210, 211-212
 standard error, 190, 195, 210-211,
 215, 315, 319
 statistical inferences about (see
 Random sampling, from a
 discrete twofold population;
 Random sampling, from a
 discrete manifold population)
- Percentiles, 8
- Permutations, 23
- Platykurtic distributions, 11
- Poisson distribution, betas, 215, 217
 derivation, 215-216
 formula, 212
 mean, 215
 derivation, 217
 moments, 217-218

Poisson distribution, and the normal distribution, 215, 217
 as sampling distribution of percentages, 211-213
 testing hypotheses with, 213-215
 variance, 215
 derivation, 217-218
 Polls, public opinion (*see* Random sampling, of public opinion)
 Polygons of regression, 16
 Population, bivariate, samples from. 298
 meaning of, 5
 types, 154-155
 Power of a test, 198*n*.
 Probability, definition of mathematical probability, 26-27
 dependent probabilities, 30-31
 empirical approximated, 27
 equations involving, 28-29
 independent probabilities, 30
 and law of large numbers, 27-28
 probability set, 26-27, 30
 Probability calculus, addition theorem, 29
 meaning of mutually exclusive in, 29
 multiplication theorem in, 30
 dependent probabilities, 30-31
 independent probabilities, 30
 Probability curves (*see* Frequency curves)
 Probability distributions, definition, 28
 Probability set, 26, 30
 derived or second order, 30
 Progressions of the means (*see* Means, progressions of)
 Public opinion polls (*see* Random sampling, of public opinion)

Q

Quadrature formulas, 128
 Quality control, use of range in, 295-296
 Quartiles, 7-8

R

Random sampling, from a discrete manifold population, assumptions, 308-309
 confidence zone for p_i 's, 328-329
 distribution of sample

$$\sum \frac{(N_i - Np_i)^2}{Np_i},$$

 explanation, 323-326
 distribution of sample percentages, derivation, 309-312
 formula, 311
 illustration of a skewed distribution, 317-319
 illustration of symmetrical distribution, 312-314
 mean of, 314-317
 standard deviation of, 314-317
 maximum-likelihood estimates of p_i 's, 329-330
 testing hypotheses about p_i 's, using distribution of

$$\sum \frac{(N_i - Np_i)^2}{Np_i},$$
 326-328, 330-331
 using multinomial distribution, 319-323
 from a discrete twofold population, assumptions, 186-187
 confidence coefficient and the confidence interval, 207-208
 distribution of difference between two sample percentages, 393-394
 distribution of sample percentages, derivation, 188-190
 formula, 190
 and the population percentage, 191-194
 and the size of the sample, 190-191
 estimation of population percentage, confidence intervals, 202-207

- Random sampling, from a discrete twofold population, estimation of population percentage, maximum likelihood estimate, 208-209
- size of sample and the confidence interval, 207
- small population percentage, 211-215
- small populations, 209-211
- testing the difference between two sample percentages, 392-395
- alternative method, 395-397
- testing hypotheses, coefficient of risk, 195
- regions of rejection, 195-201
- the test, 202
- from a normal population, confidence limits for σ^2 , 287-290
- confidence limits for $\bar{X}$, relationship between interval, coefficient, and N , 272-273
- σ known, 269-272
- σ unknown, 280-283, 284
- distribution of all possible samples, derivation, 255-256
- geometrical representation, 256-258
- properties, 257-259
- in terms of their means and variances, 259-262
- distribution of samples of $N = 2$, derivation, 221-224, 246-247
- geometrical measurement of $\sqrt{N}(\bar{X} - \bar{X})/\bar{s}$, 228-229
- geometrical measurement of σ , 226-228
- geometrical measurement of σ^2 , 226-228, 248
- geometrical measurement of $\bar{X}$, 225-226, 248
- numerical illustration, facing 224
- properties, 224-229, 247-248
- in terms of their means and variances, 248-250
- Random sampling, from a normal population, distribution of sample a 's ($=A.D./\sigma$), 244-245
- use in testing departure from normality, 296-297
- distribution of sample means, derivation, 262-263
- formula, 231
- ($N = 2$), derivation, 229-231, 250-251
- use in making inferences about population mean, 267-273
- distribution of sample medians, 241-242
- distribution of sample $\sqrt{N}(\bar{X} - \bar{X})/\bar{s}$, derivation, 264-266
- formula, 241
- ($N = 2$), derivation, 236-241, 252-253
- use in making inferences about population mean, 273-284
- distribution of sample ranges, 242
- use in analysis of variance, 295
- use in making inferences about population standard deviation, 294-295
- use in quality control, 295-296
- distribution of sample standard deviations, derivation, 264
- formula, 236
- ($N = 2$), derivation, 236
- distribution of sample variances, derivation, 263-264
- formula, 234
- ($N = 2$), derivation, 232-233, 251-252
- relation to χ^2 distribution, 234-236, 264
- use in making inferences about population variance, 284-294

- Random sampling, from a normal population, distributions of g statistics, 243
 - use in testing departure from normality, 296-297
- distributions of sample betas, 242, 243-245
 - use in testing departure from normality, 296-297
- joint confidence zone for $\bar{X}$ and σ , 366-370
- joint distribution of means and standard deviations, derivation, 340-344
 - formula, 344
 - numerical illustrations, 343, 346
- joint maximum-likelihood estimates of $\bar{X}$ and σ , 370-371
- maximum-likelihood estimate of σ^2 , 290-294
- maximum-likelihood estimate of $\bar{X}$, σ known, 273
 - σ unknown, 283-284
- testing departures from normality, by use of β_1 , β_2 , and α , 296-297
- testing difference between means of two correlated samples, 403-406
- testing difference between means of two independent samples, a common known σ , 398-400
 - a common unknown σ , 400-402
 - unequal σ 's, 403
- testing the difference between variances of two independent samples, both directions considered, 411-413
 - only one direction considered, 406-411
- testing hypotheses about both $\bar{X}$ and σ , 344-366
 - equations for λ contours, 353
- Random sampling, from a normal population, testing hypotheses about both $\bar{X}$ and σ , lambda (λ) probability tables, 361, 362
 - regions of rejection, 344-361
 - using corner region, 364-366
 - using λ contours, 345-354
- testing hypotheses about σ^2 , 284-287, 289-290
- testing hypotheses about $\bar{X}$, σ known, 267-269
 - σ known compared with σ unknown, 277-279
 - σ unknown, 273-276, 284
- testing whether two samples are from same population, 413-416
- from a normal bivariate population, confidence limits for, correlation coefficient (r), 301-302
 - regression parameters, two variables, 377
 - more than two variables, 380
- confidence limits for X_1 , 387-390
- distribution of regression statistics, 375-376
- distribution of sample correlation coefficients, 298-300
- distribution of sample lines of regression, 380-381
- distribution of sample z 's, 299-300
- limiting loci for population line of regression, 382-384
- maximum-likelihood estimate of correlation coefficient (r), 302-303
 - regression statistics, 372-374
- testing hypotheses about, correlation coefficient (r), 300-301
 - population regression parameters, 376-377
- X'_1 , 381-382

- Random sampling, from a normal bivariate population, testing difference between correlation coefficients of two independent samples, 417
 regression statistics of two independent samples; 417-419
 testing for linearity, 306-307
 testing significance of η , 306
 testing significance of I , 306
 from a normal multivariate population, confidence limits for, higher order variances, 386-387
 multiple correlation coefficient ($R_{i,jk} \dots$), 305
 population regression parameters, 380
 distribution of higher order variances, 385-386
 distributions of regression statistics, 377-378
 distribution of X'_1 , 385
 inferences about partial correlation coefficients, 303-304
 limiting loci for the population plane of regression, 385
 maximum-likelihood estimates of, higher order variances, 387
 multiple correlation coefficient, $R_{i,jk} \dots$, 305-306
 regression statistics, 374-375
 testing hypotheses about, higher order variances, 386
 multiple correlation coefficient $R_{i,jk} \dots$, 304-305
 a regression parameter, 378-380
 X'_1 , 385
 meaning, 154, 155-156
 and probability, 154
 of public opinion, 330-331
 representative, 183-184
 statistical inferences from (*see* Statistical inference)
 stratified, 183-184
- Random sampling, technique of, mechanical randomizing devices, 157-158
 natural selection, 161-162
 ordinal selection, 156-157
 random sampling numbers, 159-161
 tables of numbers, 158-159
Random Sampling Numbers, 159
 Range, absolute, definition, 13
 relative, definition, 13
 sampling distribution, 242, 482
 probabilities, table of, 482
 use in quality control, 295-296
 Rank correlation, 452-453
 Reciprocals of numbers, table, 467-468
 Region of acceptance, 164.
 Region of rejection (*see* Statistical inference, testing hypotheses, region of rejection)
 Reitz, Henry L., 92
 Robinson, G., 93
- | S
- Sampling; purposive, 184-185
 random (*see* Random sampling)
 reasons for, 153-154
 Sampling distribution, of $a = A.D./\sigma$, 244-245
 table of .01, .05, and .10 points, 481
 of $\sqrt{\beta_1}$, 242-244
 table of .05 and .10 points, 480
 of β_2 , 242-245
 table of .01 and .05 points, 480
 of the correlation coefficient (r), 298-300
 of the difference between two means, 399, 401
 percentages, 393-394
 z 's, 417
 explanation of, 164
 of g statistics, 242-243
 of a higher order variance, 385-386
 of k statistics, 242-243
 of L , 412

- Sampling distribution, of λ_H , 415
 probabilities, table of, 415
 of a linear function, 108-109
 of the mean, any population, 107-108
 normal population, 231, 262-263
 special nonnormal populations, 445-446
 of the median, 241-242
 of the multiple correlation coefficient, 304-305
 of $\sqrt{N}(\bar{X} - \bar{X})/\bar{\sigma}$, nonnormal populations, 449
 normal population, 241, 264-266
 special nonnormal populations, 447-448
 one general distribution, $11\frac{1}{2}n$.
 of partial correlation coefficient, 303
 of a percentage, 190-194, 210, 211-212
 of the ratio of two maximum-likelihood estimates of variance, 409
 of a regression coefficient, 108-109
 of the standard deviation, normal population, 236, 264
 special nonnormal populations, 446-447
 standard error of, 164
 of $z = \tanh^{-1}r$, 299-300
 use of various distributions in making inferences (*see* Analysis of variance; Random sampling; Statistical inference)
 of the variance, nonnormal populations, 450
 normal population, 234-236, 263-264
 special nonnormal populations, 446-447
 Semi-invariants (*see* Cumulants)
 Sheppard, W. F., 132
 Sheppard's corrections (*see* Frequency curves, fitting of, Sheppard's corrections)
 Skewness, 11
 Smith, B. Babington, 159, 160
 Squares of numbers, 100-1000 table, 461-462
 Square roots of numbers, 100-1000 table, 689-690
 Standard deviation, definition, 12
 first order, 18-19
 higher order, 18
 sampling distribution, 236, 264, 446-447
 Standard error, of $\sqrt{\beta_1}$, 242, 244
 of β_2 , 242, 244
 definition, 164
 of the difference between two sample means, 399
 percentages, 394
 regression statistics, 418
 z 's, 417
 of g_1 , 243
 of g_2 , 243
 of the mean, 231, 268
 of the median, 173, 242
 of a percentage, 190, 195, 210-211, 215, 315, 319,
 of the range, 242, 482
 of regression statistics, more than two variables, 378
 two variables, 375
 of a sum or difference, correlated variables, formula, 406
 independent variables, formula, 392
 proof of formulas, 420-421
 of the variance, 289
 of X'_1 , 381, 382, 385
 or $z = \tanh^{-1}r$, 301
 Statistic, 5, 181
 Statistical inference, confidence intervals, confidence coefficient, 175, 176
 confidence limits, 176
 definition, 174-175
 determination of, 175-176
 arbitrary elements in, 177-178
 effect of size of sample, 179

- Statistical inference, estimation of
 population parameters, 174-185
 maximum-likelihood estimates of
 population parameters, consistency of, 182
 efficiency of, 182
 method, 179-182
 as "optimum" statistics, 182-183
 as "unbiased estimates," 424
 sufficiency of, 183
 symbolic representation, 10
 sampling distributions, explanation, 164
 standard error, 164
 testing hypotheses, arbitrary elements in, 165-168
 caution in, 339
 coefficient of risk, 164
 effect of size of sample, 171-172
 effect of statistic selected, 172-174
 error I, definition, 163-164
 limiting risk of, 164-168
 error II, definition, 164
 minimizing risk of, 168-171
 power of a test, 198, 199
 the problem, 163
 region of acceptance, 164
 region of rejection, definition, 164
 selection of, 166-171
 unbiased region, 201
 theory of, 163-185
 (See also Random sampling)
 Stirling's approximation for factorials, 70-71
 Symbols, table of, used in this book, xi
 Symmetrical binomial distribution, betas, 34
 Symmetrical binomial distribution, characteristics, 34
 conditions leading to, 37-38
 derivation of, 32-33
 formula, 33
 mean, 34
 numerical example, 33
 relative slope of, 74-75
 significance of, 35-36
 standard deviation, 34
- T
- Tables for Statisticians and Biometricians*, 148, 149
 Tchébychef's inequality, 453-455
 extension of, 454-455
 t distribution (see Frequency curves, t distribution)
 Test of goodness of fit (see Frequency curves, testing goodness of fit)
 Testing hypotheses (see Statistical inference, testing hypotheses)
 Test of independence (see Independence, test of)
 Tippett, L. H. C., 159
Tracts for Computers, 117, 159
- U
- United States census (1940), 162
 Universe (see Population)
- W
- Whittaker, E. T., 93
 Williams, P., 244
- Z
- z transformation, 299-300

