THE MATHEMATICAL PART OF ELEMENTARY STATISTICS

A TEXTBOOK FOR COLLEGE STUDENTS

BY

BURTON HOWARD CAMP

PROFESSOR OF MATHEMATICS
WESLEYAN UNIVERSITY

1934

D. C. HEATH AND COMPANY

BOSTON NEW YORK CHICAGO
ATLANTA SAN FRANCISCO DALLAS
LONDON

3A4

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

This is an elementary textbook, dealing with the mathematical part of statistics. Statistical methods are used very generally by investigators in widely different fields, by economists, biologists, psychologists, physicists, and astronomers. The emphasis in one field is different from that in another, but there is a well-defined body of material, mathematical in nature, which is common to all. These mathematical rudiments are set forth in this text.

The course for which this book is intended will usually be offered to sophomores, and may, like the book, be divided into two parts, the first of which may be taken without the second. The book is, in fact, a reduction to printed form of two half-year lecture courses, which the author has been giving for the past ten years, and it has been quite the custom for certain students to take the first part only, substituting for the second part an applied course in economic or educational statistics. This arrangement has been advantageous to all the departments concerned. The mathematics has been taught in the department of mathematics, and the applications in the departments in which they belonged.

A brief course in analytic geometry is presupposed, but no calculus — although an occasional notation is introduced which is borrowed from calculus, such as the integral sign to indicate the area under a curve. Analytics is inherently necessary to the proper understanding of statistical methods, and it seems to the author better to suppose that it has been taught already by the use of one of the excellent texts in analytics than to attempt to teach it along with statistics.

Part I is distinctly easier than Part II, and it is intended

to provide a sufficient mathematical introduction to most of the applied courses. It will probably be found about as difficult as a semester course in analytics. Part II begins with the theory of probability. This subject involves intrinsically difficult notions, comparable in difficulty with those of the calculus. It must be presented because it lies at the basis of the theory of sampling. However, students who would be characteristically unable to think through for themselves problems in probability such as are given here should not be encouraged to study the second part. It is not desirable to try to teach them sampling theory, for they would not really understand it, and it is better that they should not acquire a superficial facility in using the formulae. The subject of finite differences, briefly considered in the closing chapter, may be taught with Part I if desired, being logically detached from the rest of the theory. Part III comprises the more necessary statistical tables and a separate introduction, including a more complete discussion of the point binomial than might be desired in an elementary text. Many of the simplest problems in probability require for their solution the summation of a number of consecutive terms of a point binomial. There is no single short method of obtaining even a fair approximation to this sum which is available for all cases. One set of formulae works best in one case, another in another. In this introduction certain selected formulae are listed, with instructions as to when each is to be used. By the aid of the tables, the summation can be made quickly so as to obtain the required probability, correct to two or three places. Part III is also published separately. It is supposed throughout, but especially in the instructions as to the use of these tables, that a computing machine is at the disposal of the student; nevertheless most of the problems in this text can be handled satisfactorily with only a slide rule and a four-place table of logarithms.

PREFACE

In writing this text, the author has held in mind two principles. First, every idea presented must be illustrated with a numerical example, and this must be followed by short "exercises," which can be done — at the board if desired — without any mechanical aids, and in five or ten minutes each. These are artificial problems involving simple numbers, and have been invented solely for the purpose of teaching the methods. Longer numerical problems, having applications in various statistical fields, and problems in theory, are listed at the close of each chapter. The second principle is that tacit hypotheses underlying the various methods shall be exposed. This is the more necessary because in practice these hypotheses often fail to be realized, and it is important that, at the very outset of his training, the statistical worker should appreciate the tentative character of results in such cases. Moreover, it makes possible the maintenance of an attitude of rigorousness of treatment which is very desirable for the student's proper mathematical development.

Problem material and statistical data have been selected freely from several books and journals. The author's name is usually indicated where this has been done. I am deeply indebted to many of my students for their assistance in preparing the manuscript, particularly perhaps to H. G. Neebe, whose thorough work on the tables was very valuable to me, and to H. A. Lewis, who read with keenly critical mind the whole manuscript and made further special investigations at several points.

B. H. C.

MIDDLETOWN, CONNECTICUT,
SEPTEMBER, 1930.

CONTENTS AND FORMULAE

SECTION PAGE

$$\mu_y{}^r = \frac{1}{N}\sum_{x,y} y^r f(x, y) = \frac{1}{N}\sum_y y^r f(y),$$

$$p_{xy} = \frac{1}{N}\sum_{x,y} xyf(x, y),\ r = \frac{p_{xy}}{\sigma_x\sigma_y}.$$

$$u = \frac{X - A}{h},\ \bar{u} = \frac{1}{N}\Sigma u f(u),\ \sigma_u^2 = \frac{1}{N}\Sigma u^2 f(u) - \bar{u}^2,$$

$$\sigma_x = h\sigma_u.$$

$$U = \sum_u uf(u, v),\ V = \sum_v vf(u, v),\ \sum_v vU = \sum_u uV,$$

$$r = \frac{\frac{1}{N}\Sigma vU - \bar{u}\bar{v}}{\sigma_u\sigma_v}.$$

$$r = \frac{1}{\sqrt{\frac{1}{N}\Sigma u^2 - \bar{u}^2}\sqrt{\frac{1}{N}\Sigma v^2 - \bar{v}^2}}\left(\frac{1}{N}\Sigma uv - \bar{u}\bar{v}\right).$$

CHAPTER IX — REGRESSION, INTERPRETATION OF r

$$\frac{y}{\sigma_y} = r\frac{x}{\sigma_x}, \qquad \frac{x}{\sigma_x} = r\frac{y}{\sigma_y},$$

$$\frac{Y - \bar{Y}}{\sigma_y} = r\frac{X - \bar{X}}{\sigma_x}, \qquad \frac{X - \bar{X}}{\sigma_x} = r\frac{Y - \bar{Y}}{\sigma_y}.$$

$$\text{For } y = rx, \quad \frac{1}{N}\Sigma\delta^2 f = 1 - r^2;$$

$$\text{for } x = ry, \quad \frac{1}{N}\Sigma\delta^2 f = 1 - r^2;$$

$$\text{for } y = \pm x, \frac{1}{N}\Sigma\delta^2 f = 1 - |r|.$$

CHAPTER X — NORMAL SURFACE. CORRELATION OF NON-MEASURABLE CHARACTERS

PART II

CHAPTER I — PROBABILITY

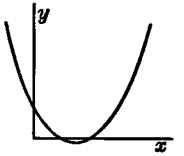
CHAPTER VII — FINITE DIFFERENCES

PART I: THE MATHEMATICAL PART OF ELEMENTARY STATISTICS

CHAPTER I

GRAPHS AND NOTATION

1. Function. The student is already familiar with certain functions and their graphs, although the name function may not have been applied to them. An example is the function represented by the parabolic curve



$$y = 2 - 3x + x^2.$$

In this example the expression $2 - 3x + x^2$, or the single variable y which is equal to it, is said to be a function of x. The curve is the graph of the function. As in analytics, the curve is also said to be the graph of the equation. Now, more generally, we shall say that any mathematical expression involving x is a "function of x." It need not be a quadratic expression like the example just considered, or even algebraic. Such expressions as $\log x$, $\tan x$, $\sqrt{x-2}$ are functions of x.

Example 1. If $3 + 2x + x^2$ is a certain function of x, what is the same function (*a*) of z? (*b*) of y? (*c*) of -3? (*d*) of $(x+1)$?

The answers are obtained by substituting for x the new variable or number suggested. They are: (*a*) $3 + 2z + z^2$, (*b*) $3 + 2y + y^2$, (*c*) $3 - 6 + 9 = 6$, (*d*) $3 + 2(x+1) + (x+1)^2 = 6 + 4x + x^2$.

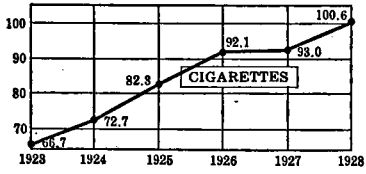
We shall also use the word function in a slightly more general sense: y is a function of x if the value of y is dependent on the value of x in the sense that to every one of a given set of values of x there corresponds a value of y, whether or not we can explicitly describe this correspondence by means of a familiar type of expression. Thus (see Example 2, p. 4)

the yearly production of cigarettes in the United States is dependent on and is therefore a function of the year, whether or not we know of, or, in fact, whether or not there exists, an expression defining that functional relationship.

2. Graphs. In general, the graphs of statistics are not essentially different from the graphs of analytics, but there are a few differences worth noting.

(*a*) Broken lines are more commonly used to connect points that have been plotted, instead of curves. This is because often these plotted points are not obtained by substitution in the equation of a curve, as in analytics, and are not thought of as special points of a curve. Often the function being represented does not exist except at these occasional points, and the broken line is drawn merely to aid the eye in locating the points.

Example 2. YEARLY PRODUCTION OF CIGARETTES IN THE UNITED STATES. (*The World Almanac.*)



YEAR	BILLIONS
1923	66.7
1924	72.7
1925	82.3
1926	92.1
1927	93.0
1928	100.6

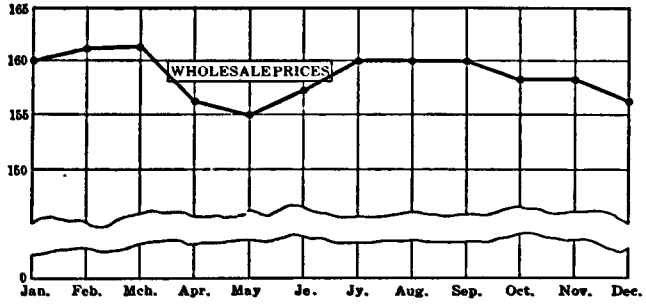
Here the function is defined by the table, and, as stated above, does not exist except at the points designated. For example, the number of cigarettes produced in the year ending December 31, 1925, was 82.3 billions, and the number produced in the year ending December 31, 1926, was 92.1 billions. If we wish to inquire what was the number produced in the year ending June 30, 1926, we have to say that the required information is not contained in our table. This is a case in which the function is not defined at the point midway between 1925 and 1926. Of course we may estimate this number by simple interpolation in the table, or, what is the same thing, by reading the ordinate of the straight line of the graph at the point halfway between 1925 and 1926; and our estimate might be better

if we were to draw a smooth curve through all the points instead of a broken line; but such an estimate would have to be recognized as an approximation, not as an exact figure. Often in statistics we are not concerned with this sort of question, and, unless we are, the lines are drawn merely to help us to appreciate visually the relative positions of the points.

(*b*) The distance from the diagram to the origin is often so great that the origin of a well-placed diagram would lie off the page. When this is the case, it is well to indicate the fact by a broken diagram as in the following example.

Example 3. INDEX OF WHOLESALE PRICES FOR THE YEAR 1925; 1913 = 100 (*C. Snyder*).

INDEX	160	161	161	156	155	157	160	160	160	158	158	156
MONTH	Jan.	Feb.	Mch.	Apr.	May	Je.	Jy.	Aug.	Sep.	Oct.	Nov.	Dec.



Frequently the lower portion of the figure is omitted, for the wavy base line of the upper portion is believed to be a sufficient warning to the reader. Without a warning the diagram may be misleading, although literally truthful. That is, one instinctively thinks of percentage or relative changes in looking at a diagram, and when the origin is off the page, these are apt to appear more pronounced than they really are. On the other hand, in Example 2 it is not necessary to break the diagram vertically so as to indicate that the origin of time is also off the page. The diagram is in no way misleading as it is, and the origin of time is quite arbitrary. It might

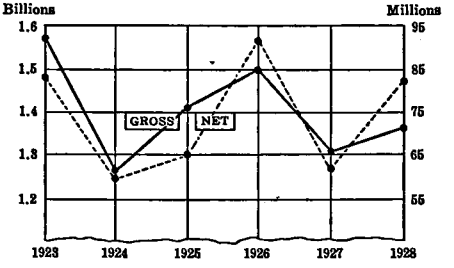
be taken at the end of the year 1923, just as in Example 3 it is taken at the beginning of the year 1925.

(*c*) The unit of the horizontal scale is commonly different from the unit of the vertical scale. This is the case [1] both in Example 2 and in Example 3.

(*d*) Two vertical scales may be used with the same horizontal scale.

Example 4. UNITED STATES STEEL CORPORATION. GROSS AND NET EARNINGS.

Year	*Gross Earnings, Millions of Dollars*	*After Preferred Dividends, Millions of Dollars*
1923	1571	83.3
1924	1264	59.8
1925	1407	65.4
1926	1508	91.1
1927	1310	62.1
1928	1374	82.8



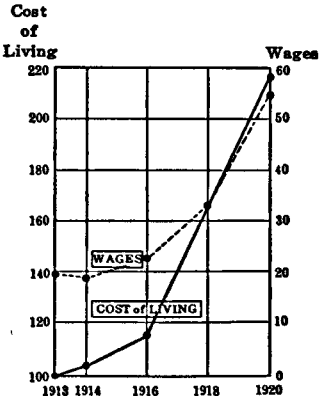
[1] The student may wish to add that the thing measured by the vertical scale does not have a common measure with the thing measured by the horizontal scale; that in Example 2, for instance, cigarettes and years are incommensurable. This is true but not exactly the point. We can avoid his objection by saying that the *number* of years elapsed since 1922 is related to the *number* of billions of cigarettes produced. Now these two *numbers* are commensurable but in our figure they were measured off on scales of different lengths.

By placing the origin off the page and shifting the units, one may produce a very misleading diagram. This is shown by the following example. It would appear from the diagram that the wage rate did not increase as rapidly as the cost of living, but from the table it may be seen that in the interval 1913–1920 the wage rate was multiplied by $\frac{55.7}{19.9} = 2.80$, and the cost of living by only 2.15. So the wage rate really increased more rapidly than the cost of living.

Example 5.

Year	*Cost of Living Index*	*Hourly Rate of Wages for Men, in Cents*
1913	100	19.9
1914	104	18.6
1916	114	22.2
1918	166	33.3
1920	215	55.7

In this diagram, if the origin of wages had been placed at the cost of living origin, and if the scale of wages had been made proportional to the cost of living scale, the picture would have been a true one.



A Misleading Diagram

The differences just considered between graphs in statistics and those which occur frequently in elementary analytics must not be thought of as separating the one subject from the other. Graphical statistics is in itself a part of analytic geometry, but the features just mentioned may not have been presented in a first course. However, it may be

desirable to use certain of these same methods in graphing very simple curves that do occur in elementary analytics. This is illustrated in Problems 2 and 3 at the close of this chapter.

3. Sums. In elementary algebra, when we add together a set of letters, such as $t_1, t_2, \cdots, t_n$, we represent their sum in the form $t_1 + t_2 + \cdots + t_n$. In more advanced mathematics, it is common to designate this sum in a more compact manner, making use of the Greek letter, capital sigma Σ, thus:

$$\sum_{i=1}^{n} t = t_1 + t_2 + \cdots + t_n. \tag{1}$$

The numbers represented by these letters need not be all different, but, if each of the numbers is repeated several times, it is customary to write the sum as follows.

Suppose

$$\left.\begin{aligned} &t_1 \text{ occurs } f_1 \text{ times,} \\ &t_2 \text{ occurs } f_2 \text{ times,} \\ &\cdots \\ &\cdots \\ &\cdots \\ &t_n \text{ occurs } f_n \text{ times} \end{aligned}\right\} \tag{2}$$

The sum of all these numbers, each counted as often as it occurs, is:[1]

$$\sum_{i=1}^{n} t_i f_i = t_1 f_1 + t_2 f_2 + \cdots + t_n f_n. \tag{3}$$

Usually, in cases of this sort, t_i is a measurement of some kind, and f_i is called its "frequency." Then the set (2) is called a frequency distribution. The total frequency is commonly designated by N:

$$N = \sum_{i=1}^{n} f_i. \tag{4}$$

[1] Some authors use Σt even in a case like this. It will not be our general practice, although we shall resort to it occasionally, when the f's are small and no ambiguity will be caused thereby.

Example 6. The length of a hall is measured fifty times by means of a small rule, with the following results:

Measurement t *(feet)*	*Frequency* f	tf
200.1	1	200.1
200.2	1	200.2
200.3	3	600.9
200.4	12	2440.8
200.5	14	2870.0
200.6	18	3700.8
200.7	0	0.0
200.8	1	200.8
Totals	50	10024.6

$N = 50$, $\Sigma tf = 10024.6$.

We shall find later that the frequency of an observation is also called its "weight." The sum indicated in (3) is therefore commonly called the "weighted sum."

4. Mean Value. The weighted sum divided by the total weight, or total frequency, is the "mean value." This is also called the arithmetic mean, or, more briefly, the mean or average. There are also other kinds of means and averages with which it should not be confused. This mean value will be denoted by $\bar{t}$, thus:

$$\bar{t} = \frac{1}{N} \sum_{i=1}^{n} t_i f_i.$$

Example 7. The mean length of the hall in Example 6 is $10024.6/50 = 200.492$ feet, for from the last column in that example we have $\Sigma tf = 10024.6$; and $N = \Sigma f = 50$.

5. Variables and Constants. In analytics we have learned that any letter which may take on different values is a variable and ordinarily it has been denoted by x or y. Let us now note that in the expression,

$$\sum_{i=1}^{n} t_i = t_1 + \cdots + t_n,$$

the subscript letter i is understood to vary from term to term, beginning with the value 1 and ending with the value n. It is therefore a variable. The letter t also varies, but its value depends on the value of i, thus: t_1 has one value, t_2 another, etc. So t is also a variable, but of a different kind from i. We shall say that i is an *independent* variable and that t is a *dependent* variable; in fact, t is a function of i. There is nothing mandatory, however, about the use of the letter t for the dependent and of i for the independent variable. Their rôles might be interchanged, thus:

$$\sum_{t=1}^{n} i_t = i_1 + \cdots + i_n.$$

It is mandatory only that there should be a clear indication as to which is which. This is the reason for writing the subscripts and the equations, $i = 1$ or $t = 1$, underneath the symbol Σ. Here we find always the independent variable. It is because of possible confusion of this sort that the shorter notation Σt is sometimes ambiguous. It may mean $\sum_{i=1}^{n} t_i = t_1 + \cdots + t_n$, as above, or it *might* mean

$$\sum_{t=1}^{n} t = 1 + 2 + \cdots + n.$$

Usually the context tells the reader clearly enough which is meant, but when it does not, the subscripts should be written. In the expression,

$$\sum_{i=1}^{n} ct_i = ct_1 + \cdots + ct_n,$$

the letter c is understood to have the same value in all the terms. It is therefore called a constant. We may now state two simple theorems which will be used frequently.

Theorem I. *A constant factor of the expression following the symbol* Σ *may be moved to the left of the symbol without affecting the value of the sum found, thus:*

$$\Sigma c t_i = c \Sigma t_i.$$

Theorem II. *If the expression following the symbol* Σ *is itself the sum of two (or more) terms, each may be considered separately and the results added, thus:*

$$\Sigma(t_i + u_i) = \Sigma t_i + \Sigma u_i.$$

The proofs of these theorems become self-evident as soon as the expressions are written in their expanded forms.

EXERCISES § 5

1. *Write in expanded forms:*

   (*a*) $\sum_{i=1}^{n} t_i^2$, (*b*) $\sum_{j=1}^{n} x_j f_j$, (*c*) $\sum_{i=3}^{n-1} (t_i - 1)(u_i + 1)$,

   (*d*) $\sum_{t=0}^{n} t^2$.

2. *Write in abbreviated forms, using* Σ:

   (*a*) $\frac{t_1^2}{f_1} + \frac{t_2^2}{f_2} + \cdots + \frac{t_n^2}{f_n}$,

   (*b*) $(\bar{t} - t_1)f_1 + (\bar{t} - t_2)f_2 + \cdots + (\bar{t} - t_n)f_n$,

   (*c*) $\frac{1}{N}[(t_1 - \bar{t})^2 f_1 + \cdots + (t_n - \bar{t})^2 f_n]$,

   (*d*) $\frac{1}{N}[|t_1 - \bar{t}| f_1 + \cdots + |t_n - \bar{t}| f_n]$.

3. (*a*) Express as the difference of two sums the compressed sum which is the result of Exercise 2 (*b*), and then show that that result equals zero.

   (*b*) Prove that $\sum_{i=1}^{n-1} (f_i - f_{i+1}) = f_1 - f_n$.

   (*c*) Show that $\sum_{t=0}^{n} t(t-1)(t-2)p^t = \sum_{t=3}^{n} t(t-1)(t-2)p^t$.

4. Compute the numerical values of Exercises 2 (*b*), (*c*), and (*d*) in the special case of Example 6, p. 9.

Ans., 0, .0151, .0946.

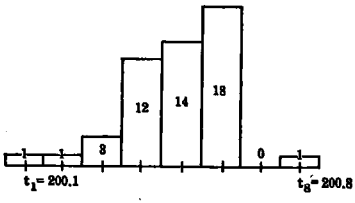
5. A frequency distribution is denoted, as on p. 8, by means of the letters (t, f). Express in compressed form a description of each of the following processes:

(*a*) Each t is to be squared and multiplied by its own f, and then all the results are to be added together.

(*b*) Each t is to be subtracted from the t following it, and then the numerical values of all these differences are to be added together.

(*c*) The sum of the squares of the t's is to be divided by the square of the sum of the t's.

6. Histograms. If we have a frequency distribution in which the intervals t_1 to t_2, t_2 to t_3, etc., are all equal, a common method of graphing it is indicated here. This graph is a histogram of the distribution in Example 6. The rectangles all have bases of equal widths. Their areas are equal (on a conveniently chosen scale) to the several frequencies. This is graphically accomplished by making the altitudes proportional to these frequencies. If the bases are all of unit width, then the altitudes will also be equal to the frequencies. Another method of graphing a frequency distribution consists in plotting the points (t_1, f_1), (t_2, f_2), etc., as in analytics, and drawing a broken line through them.



The diagram which we have called a histogram recognizes the fact that usually the several measurements do not lie precisely at the points indicated in the table, but are spread out over intervals of which the several points are centers. These intervals are called class intervals. They are the bases of

the rectangles. Often a frequency table is presented so as to indicate these intervals. Thus, the table of Example 6 might have read as follows:

t, in feet	f
$200.05 < t < 200.15$	1
$200.15 < t < 200.25$	1
$200.25 < t < 200.35$	3
$200.35 < t < 200.45$	12
$200.45 < t < 200.55$	14
$200.55 < t < 200.65$	18
$200.65 < t < 200.75$	0
$200.75 < t < 200.85$	1
Total	50

Since frequency tables are presented in several different ways some care must be used in finding the class intervals and their mid-points. Suppose our schedule of the values of t began 200.05 —, 200.15 —, etc. This would mean the same as $200.05 \leqq t < 200.15$, $200.15 \leqq t < 200.25$, etc., that is, the first end of each interval is included but not the last end. The reader would then naturally assume that the last point of the first interval was (to two-place accuracy) 200.14, so that the schedule for t might have been written: $200.05 \leqq t \leqq 200.14$, $200.15 \leqq t \leqq 200.24$, etc. Then the mid-points would have been 200.095, 200.195, etc., instead of 200.10, 200.20, etc., as above. But, in each of these cases, the length of the class interval would have been 10. The length of a class interval may always be found by finding the distance between its beginning point and the beginning point of the next interval. It is necessary to find the mid-points, because it is the t's of the mid-points that are used in computing the mean.

Example 8. Find the mean wage from the following data:

DATA		COMPUTATION	
Class	*f*	*Mid-t*	*ft*
\$4.50– 5.99	43	5.245	225.535
6.00– 7.49	99	6.745	667.755
7.50– 8.99	152	8.245	1253.240
9.00–10.49	178	9.745	1734.610
10.50–11.49	160	11.245	1799.200
12.00–13.49	41	12.745	522.545
13.50–14.99	25	14.245	356.125
15.00–16.49	3	15.745	47.235
Sums	701		6606.245
$\frac{1}{N}$ (Sums)	1		9.42

$\bar{t} = \frac{\Sigma ft}{N} = \$9.42.$ (Class interval = \$1.50.)

PROBLEMS CHAPTER I

1. Make graphs of the following sets of data, being careful so to choose both the zero points and the scales that the material will be plainly spread out over the whole page, and yet so that all the material will be on the page. Label clearly and simply.

(*a*) *U.S. Steel Corporation. Number of millions of dollars.*

Year	*Net Balance*	*Gross Income*
1920	176.7	755.5
1921	92.7	986.7
1922	101.5	1092.7
1923	179.6	1571.4
1924	153.1	1263.7
1925	165.5	1406.5
1926	199.1	1508.1
1927	164.3	1310.4
1928	193.3	1374.4
1929	258.7	1493.5

(*b*) *National Bureau of Economic Research. Internal revenue reports.* The number of thousands of income tax returns filed is represented by x, the number of millions of dollars of net income reported by y.

Year	x	y
1916	437	6 298
1917	3472	13 652
1918	4425	15 924
1919	5332	19 859
1920	7259	23 735
1921	6662	19 577
1922	6787	21 336
1923	7698	24 840
1924	7370	25 656

(*c*) *U. S. Census Bureau. Mortality Statistics, 1924.* Death rates per 100,000 population (based on civilian deaths in estimated civilian populations), 1910–1923, for the United States, France, and Australia.

Year	*U. S.*	*France*	*Australia*	*Year*	*U. S.*	*France*	*Australia*
1910	15.0	17.8	10.4	1917	14.3	20.2	9.7
1911	14.2	19.6	10.7	1918	18.1	24.6	10.0
1912	13.9	17.5	11.2	1919	12.9	19.3	12.7
1913	14.1	17.7	10.7	1920	13.1	17.2	10.5
1914	13.6	20.7	10.5	1921	11.6	17.7	9.9
1915	13.6	21.0	10.6	1922	11.8	17.5	9.2
1916	14.0	19.8	11.0	1923	12.3	16.8	9.9

2. Graph the following functions in the intervals indicated:

(*a*) 2^x, $5 \leq x \leq 10$. (*b*) cotan x, $0° 6' \leq x \leq 5° 0'$.

3. Graph the equation, $\frac{x^2}{4} + \frac{y^2}{1} = 1$, making the unit of x half as long as the unit of y.

4. A certain function of x is defined in Problem 2 (a). What is the same function of $-x$? of x^2? of $\log 2$? of 2^x?

5. Make histograms (see § 6) of the following:

(a) *Weights of college freshmen (Gavett).*

Mid-t (pounds)	*f*	*Mid-t (pounds)*	*f*	*Mid-t (pounds)*	*f*
105	15	149	129	193	5
116	43	160	82	204	3
127	138	171	35	215	1
138	162	182	16	Total	629

(b) *Heights of sons of tall fathers* (72.5 to 73.5 inches), *Yule.*

Inches	*f*	*Inches*	*f*	*Inches*	*f*
66.5–67.5	4	71.5–72.5	11	76.5–77.5	6
67.5–68.5	9	72.5–73.5	13	77.5–78.5	3
68.5–69.5	9	73.5–74.5	13	78.5–79.5	1
69.5–70.5	14	74.5–75.5	7		
70.5–71.5	20	75.5–76.5	4	Total	114

(c) *Deaths from tuberculosis by ages, U. S. Census Bureau, Mortality Statistics (1924).* (*Note:* When, as here, the class intervals are not all equal, one must exercise some care to make sure that nevertheless each rectangle is equal in area to the corresponding frequency. The data presented by the Census Bureau have been altered slightly to make the problem simpler. *E.g.*, the Bureau's final classification was 75 and over.)

Age of Death	*f*	*Age of Death*	*f*	*Age of Death*	*f*
Under 1	1450	15–19	6620	45–54	10542
1– 4	2700	20–24	11121	55–64	7123
5– 9	1267	25–34	19507	65–74	4469
10–14	1693	35–44	14703	75–84	1760

6. Find the mean age of death from tuberculosis by the use of the data in Problem 5 (*c*). Assume that the mid-points are: 0.5, 2.5, 7.0, etc. *Ans.*, 35.52 years.

7. (*a*) If $f(x) = 3^{-x^2}$, what is $f(-x)$?
(*b*) If $\phi(x) = -2x \cdot 3^{-x^2}$, what is $\phi(-x)$?
(*c*) If $F(x) = (x^2 - 1) \cdot 3^{-x^2}$, what is $F(-x)$?
(*d*) If $\psi(x) = (-x^3 + 3x) \cdot 3^{-x^2}$, what is $\psi(-x)$?
(*e*) If $g(x) = (x^4 - 6x^2 + 3) f(x)$, what is $g(-x)$?

CHAPTER II

MOMENTS

1. Moments about Any Given Origin. The frequency distributions of statistics are not all alike. Some have a symmetrical form, others are skewed one way or the other. The form of a distribution can be expressed pretty well by means of certain constants or parameters called moments, just as, in analytic geometry, the form of an ellipse is determined by means of the constants or parameters: a, the semi-major axis; and b, the semi-minor axis. Except in a certain special case (§ 3), it is customary to denote these moments by means of the Greek letter nu, ν. The first moment is ν_1, the second ν_2, etc., and they are defined as follows:

Relative to the t origin, in the t unit of measurement:

$$\left.\begin{aligned} \nu_1 &= \frac{1}{N}\sum_{i=1}^{n} t_i f_i, \\ \nu_2 &= \frac{1}{N}\sum_{i=1}^{n} t_i^2 f_i, \\ \nu_3 &= \frac{1}{N}\sum_{i=1}^{n} t_i^3 f_i, \\ &\text{etc.} \end{aligned}\right\} \tag{1}$$

Sometimes one also uses the "zero*th*" moment,

$$\nu_0 = \frac{1}{N}\sum_{i=1}^{n} t_i^0 f_i.$$

Since $t^0 = 1$, and since $\Sigma f = N$,

$$\nu_0 = 1. \tag{2}$$

The definition of the first moment is the same as that given for the mean value. Hence,

$$\nu_1 = \bar{t}. \tag{3}$$

EXERCISES § 1

1. Find ν_0, ν_1, ν_2, and ν_3 in the following cases:

(*a*)

t	f	t	f
0	2	3	15
1	5	4	20
2	10	5	3

(*b*)

t	f	t	f
−2	1	1	10
−1	5	2	5
0	10	3	1

(*c*)

t	f	t	f
0.5	6	3.5	30
1.5	12	4.5	10
2.5	20	5.5	2

Ans. for (*a*): $\nu_0 = 1$, $\nu_1 = 3$, $\nu_2 = 10.45$, $\nu_3 = 39$.

2. (*a*) Write the expression for the rth moment.

(*b*) Show that the even moments are always positive in value, but that the odd moments may be negative as well as positive.

(*c*) Show that the odd moments are all zero if both the f's and the t's are symmetrical with respect to the origin of t, as, *e.g.*,

t	−5	−3	−1	1	3	5
f	1	5	10	10	5	1

(*d*) Show that $\sum_{i=1}^{n} (t_i - \bar{t})f_i = 0$.

2. Short Methods of Computing ν_1. In certain cases the method of computing $\bar{t}$ used in Chapter I can be much simplified. To prove this we first develop an alternative formula for $\bar{t}$.

Let c and A be any constants (if $c \neq 0$), and make the following substitutions in the formula for $\bar{t} = \nu_1$ given in (1):

$$u = \frac{t - A}{c}, \text{ i.e., } t = cu + A. \tag{4}$$

Then

$$\bar{t} = \frac{1}{N}\sum_{i=1}^{n} (cu_i + A)f_i.$$

By Theorems I and II of Chapter I, this equals

$$\frac{c}{N}\sum_{i=1}^{n} u_i f_i + \frac{A}{N}\sum_{i=1}^{n} f_i. \tag{5}$$

But the first of these terms is by definition c times the mean value of u, and the second is, by (2), simply A. So

$$\bar{t} = c\bar{u} + A, \text{ where } \bar{u} = \frac{1}{N}\Sigma uf. \qquad (6)$$

This is the new formula sought.

Case a (class intervals equal). If the class intervals are all equal, let c equal the class interval, and choose A as one of the given values of mid-t, usually the one opposite the greatest frequency; and it will be found that formula (6) is much easier to use in computation than formula (1). By (4), A will become automatically the origin of u, for when $t = A$, $u = 0$.

Example 1. Find the mean value of t in Example 8 of Chapter I, page 14.

Mid-t in dollars	f	u	uf
5.245	43	−3	−129
6.745	99	−2	−198
8.245	152	−1	−152
$A \rightarrow$ 9.745	178	0	0
11.245	160	1	160
12.745	41	2	82
14.245	25	3	75
15.745	3	4	12
Sums	701		−150

$c = \$1.50, \qquad A = 9.745,$

$\bar{u} = -150/701 = -.214,$

$\bar{t} = c\bar{u} + A = (1.50)(-.214) + 9.745,$

$= -.321 + 9.745 = 9.424.$

Here $A = 9.745$, but it would have been about as simple to have chosen A as any other value of mid-t near to this one. It is only a matter of so choosing the origin of u that on the whole the small values of u will be the multipliers of the larger values of f. Formula (6) was developed before the values of c and A were chosen, and therefore all choices will result in exactly the same final value of $\bar{t}$. Not all choices are equally convenient, but all are equally valid.

EXERCISES § 2 a

Find by the shortest method the mean values in the following cases:

(*a*) Exercise 1 (*a*) of § 1. *Ans.*, 3.

(*b*) Exercise 1 (*c*) of § 1. *Ans.*, 2.9.

(*c*) Exercise 2 (*c*) of § 1. *Ans.*, 0.

(*d*) The first half of the data ($t < 10.495$) in Example 1. *Ans.*, 8.223.

(*e*) The second half of the data in Example 1. *Ans.*, 11.900.

(*f*)

t	−12	− 6	0	6	12	18	24
f	18	27	99	127	154	82	19

Ans., 7.916.

Case b (*class intervals unequal*). If the class intervals are of various lengths, we choose $c = 1$. Then by a proper choice of A the work can be simplified a little. The formula becomes $\bar{t} = \bar{u} + A$, which says merely that the distance of the mean from the origin of t is equal to its distance from A plus the distance of A from the origin of t. This statement is obvious and the simplification is one that would naturally occur to the computer independently of the formula.

Example 2. Find the weighted mean of the following micrometer measurements. We choose $A = 193$, because this number makes the u's both small and easy to compute. In this case we do not choose A as the value of t, 194.171, opposite the greatest frequency, because, although this would make the u's smaller, it would make necessary more difficult subtractions in order to obtain them.

t	f	u	fu
194.032	11	1.032	11.352
193.790	3	.790	2.370
194.151	6	1.151	6.906
193.850	4	.850	3.400
194.221	5	1.221	6.105
194.171	22	1.171	25.762
Sums	51		55.895

$\bar{u} = 55.895/51$
$= 1.0959,$
$\bar{t} = \bar{u} + A$
$= 1.096 + 193.000$
$= 194.096.$

The method of Case *a* amounts to a shifting of the origin to A, and, in addition, a change in the unit of measurement. (In Example 1, the unit of measurement of the t coördinate was one dollar, but the unit of measurement of the u coördinate was c dollars.) The method of Case *b* amounts to a shifting of the origin merely.

Case c (groups of equal intervals). Where the class intervals are not all alike but are separable into groups in each of which they are alike, the method of Case *b* could be used, but it is better to use a modification of the method of Case *a*. For this we need a new theorem:

Theorem. *The general mean $\bar{t}$ of any frequency distribution can be found by separating the frequencies into groups, finding the weighted mean of each group, and then the weighted mean of these means. Briefly stated, the mean of the means is the mean of the whole, thus:*

$$\bar{t} = \frac{T_1 g_1 + T_2 g_2 + \cdots + T_m g_m}{N}, \tag{7}$$

where T_1 is the mean and g_1 the total frequency of the first group, T_2 is the mean and g_2 the total frequency of the second group, etc. The number of groups is m, and N is the total of all the frequencies.

To prove this, it is only necessary to notice exactly what it says in terms of our previous notation. Let us consider first the case where there are only two such groups. The distribution will appear as indicated on the next page.

t	f	Totals	Means
t_1	f_1		
t_2	f_2		
.	.	g_1	T_1
.	.		
.	.		
t_a	f_a		
t_{a+1}	f_{a+1}		
t_{a+2}	f_{a+2}		
.	.	g_2	T_2
.	.		
.	.		
t_{a+b}	f_{a+b}		
Sums	N	N	

Here g_1 is the sum of the first a frequencies, and T_1 is their mean:

$$g_1 = \sum_{i=1}^{a} f_i, \; T_1 = \frac{1}{g_1} \sum_{i=1}^{a} t_i f_i. \qquad (8)$$

Similarly:

$$g_2 = \sum_{i=a+1}^{a+b} f_i, \; T_2 = \frac{1}{g_2} \sum_{i=a+1}^{a+b} t_i f_i. \qquad (9)$$

By the formula for $\bar{t}$:

$$N\bar{t} = \sum_{i=1}^{a+b} t_i f_i = \sum_{i=1}^{a} t_i f_i + \sum_{i=a+1}^{a+b} t_i f_i$$
$$= T_1 g_1 + T_2 g_2, \text{ by (8) and (9).}$$

Hence:

$$\bar{t} = \frac{T_1 g_1 + T_2 g_2}{N},$$

which is equation (7) in the special case where the number of groups is two. But this proof could be used quite as well for m groups as for two. The only change would be an increase in the number of symbols needed.

We will now exhibit the method in a special example. The saving of labor is not very marked in this example, but some very practical cases arise (cf. Problem 2) in which the other methods of finding the mean are so cumbersome that they could hardly be used at all. Then this theorem becomes very important.

Example 3. Find the mean age of death from diphtheria (*Mortality Tables, U. S. Census Bureau*, 1924).

Class	*f*	*mid-t*	*g*	*T*	*gT*
Under 1	643	.5	643	.50	321.50
1– 4	4378	2.5	4378	2.50	10945.00
5– 9	2407	7.0			
10–14	612	12.0	3297	8.92	29409.24
15–19	180	17.0			
20–24	98	22.0			
25–34	161	29.5			
35–44	107	39.5			
45–54	67	49.5	403	41.29	16639.87
55–64	48	59.5			
65–74	10	69.5			
75–84	10	79.5			
Sums	8721		$N = 8721$		57315.61

In this example $T_3 = 8.92$ and is the mean of the group for which $7 \leqq \text{mid-}t \leqq 22$; $T_4 = 41.29$ and is the mean of the group for which $29.5 \leqq \text{mid-}t \leqq 79.5$. Formula (7) is in detail:

$$\bar{t} = \frac{T_1g_1 + T_2g_2 + T_3g_3 + T_4g_4}{N}$$

$$= \frac{321.50 + 10{,}945.00 + 29{,}409.24 + 16{,}639.87}{8721}$$

$$= 6.57 \text{ years.}$$

EXERCISES § 2 b, c

1. In the example above: (*a*) show that $T_4 = 41.29$; (*b*) find the mean age of death of children under 9.5; (*c*) find the mean age of death of persons older than 9.5. *Ans.*, (*b*) 3.787.

2. From the results of Exercises (*d*) and (*e*) of § 2 *a* obtain the mean value for Example 1.

3. Find the means of the three groups, *a*, *b*, and *c* (on page 25), and thence obtain the mean of the whole. *Ans.*, 35.592.

a		b		c	
t	f	t	f	t	f
1	6	10	43	75	40
2	11	15	92	100	30
3	15	20	100	125	20
4	17	25	120	150	4
5	28	50	130	200	4

3. Moments about the Mean. Before the definition of the moments given in § 1, the following words were inserted: "relative to the t origin and in the t unit of measurement." These words were necessary. If another origin or another unit had been chosen, the moments would have had other values. For purposes of computation it is seldom desirable to use the given (t) origin or unit. As with the mean, so with the higher moments, it is better to use the u origin and u unit. Relative to the *u origin* and the *u unit* (*viz.*, the class interval), we have the following definitions:

$$\left.\begin{aligned} \nu_1 &= \frac{1}{N}\Sigma uf = \bar{u}, \\ \nu_2 &= \frac{1}{N}\Sigma u^2 f, \\ \nu_3 &= \frac{1}{N}\Sigma u^3 f, \text{ etc.} \end{aligned}\right\} \tag{10}$$

In the very special case where the mean is chosen as the origin, the moments are denoted by the Greek letter mu, μ. When this letter is used, it is unnecessary to mention the origin chosen, because the mean is always understood, but it is still necessary to mention the unit. Thus, *in the u unit*,

$$\left.\begin{aligned} \mu_1 &= \frac{1}{N}\Sigma(u-\bar{u})f = 0, \\ \mu_2 &= \frac{1}{N}\Sigma(u-\bar{u})^2 f, \\ \mu_3 &= \frac{1}{N}\Sigma(u-\bar{u})^3 f, \textit{ etc.} \end{aligned}\right\} \tag{11}$$

4. Short Methods of Computing μ's. We consider here only the case where the class intervals are all equal. The ν's are easy to compute as written in (10). The μ's can be computed more readily if we express them in terms of the ν's. This is done by expanding the expressions following the Σ's in (11), thus:

$$\mu_2 = \frac{1}{N}\Sigma u^2 f - \frac{2}{N}\Sigma u\bar{u}f + \frac{1}{N}\Sigma \bar{u}^2 f$$

$$= \frac{1}{N}\Sigma u^2 f - \frac{2\bar{u}}{N}\Sigma uf + \frac{\bar{u}^2}{N}\Sigma f \qquad \text{by Theorem I,}$$

$$= \nu_2 - 2\nu_1\bar{u} + \nu_1^2 \qquad \text{by (10),}$$

$$= \nu_2 - \nu_1^2 = \nu_2 - \bar{u}^2, \text{ since } \bar{u} = \nu_1. \qquad (12)$$

$$\mu_3 = \frac{1}{N}\Sigma u^3 f - \frac{3}{N}\Sigma u^2\bar{u}f + \frac{3}{N}\Sigma u\bar{u}^2 f - \frac{1}{N}\Sigma \bar{u}^3 f$$

$$= \nu_3 - 3\nu_2\bar{u} + 3\nu_1\bar{u}^2 - \bar{u}^3$$

$$= \nu_3 - 3\nu_2\bar{u} + 2\bar{u}^3. \qquad (13)$$

$$\mu_4 = \nu_4 - 4\nu_3\bar{u} + 6\nu_2\bar{u}^2 - 4\nu_1\bar{u}^3 + \bar{u}^4$$

$$= \nu_4 - 4\nu_3\bar{u} + 6\nu_2\bar{u}^2 - 3\bar{u}^4. \qquad (14)$$

As will be shown, these formulae enable us to compute rather easily μ_2, μ_3, and μ_4 in the u unit.

Example 4. Find the μ's (in the unit of u) of Example 1.

Data		Computation				
Mid-t	f	u	uf	u^2f	u^3f	u^4f
5.245	43	− 3	− 129	387	− 1161	3483
6.745	99	− 2	− 198	396	− 792	1584
8.245	152	− 1	− 152	152	− 152	152
9.745	178	0	0	0	0	0
11.245	160	1	160	160	160	160
12.745	41	2	82	164	328	656
14.245	25	3	75	225	675	2025
15.745	3	4	12	48	192	768
Sums	701		− 150	1532	− 750	8828
$\frac{1}{N}$(Sums)	1		−.214	2.185	−1.070	12.593
			$\bar{u}$	ν_2	ν_3	ν_4

$\bar{u}^2 = .046$, $\mu_2 = 2.185 - .046 = 2.139$,
$\bar{u}^3 = -.010$, $\mu_3 = -1.070 - 3(2.185)(-.214) + 2(-.010) = .313$,
$\bar{u}^4 = .002$, $\mu_4 = 12.593 - 4(-1.070)(-.214) + 6(2.185)(.046) - 3(.002) = 12.274$.

EXERCISES § 4

1. Find the μ's, in the u unit, in the following cases:

(a)

t	f
0	1
1	3
2	3
3	1

(b)

t	f
0	1
1	4
2	6
3	4
4	1

(c)

t	f
3	1
5	5
7	10
9	10
11	5
13	1

Ans., (a) $\mu_2 = .75$, $\mu_3 = 0$, $\mu_4 = 1.31$. *Ans.*, (b) $\mu_2 = 1$, $\mu_3 = 0$, $\mu_4 = 2.5$. *Ans.*, (c) $\mu_2 = 1.25$, $\mu_3 = 0$, $\mu_4 = 4.06$.

2. Find μ_2, μ_3, in the u unit, in the following cases:

(a)

t	f
2	1
3	6
4	12
5	8

(b)

t	f
20	1
30	8
40	24
50	32
60	16

Ans., (a) $\mu_2 = \frac{3}{4}$, $\mu_3 = -\frac{3}{8}$. *Ans.*, (b) $\mu_2 = \frac{8}{9}$, $\mu_3 = -\frac{5}{27}$.

Moments in Various Units. If we wish the μ's in some other unit, say the t unit, they may be found from the following relations:

$$\left.\begin{aligned} \mu_2(t \text{ unit}) &= c^2\mu_2(u \text{ unit}) \\ \mu_3 \quad " &= c^3\mu_3 \quad " \\ \mu_4 \quad " &= c^4\mu_4 \quad " \end{aligned}\right\} \tag{15}$$

Here the u unit is c times the t unit. In Problem 5 at the end of this chapter the student is asked to prove the second and third of these relations. The proof of the first is as follows:

$$\mu_2(t \text{ unit}) = \frac{1}{N}\Sigma(t - \bar{t})^2 f,$$

for this would be the definition of μ_2 in the t unit. If in this we make the substitution of equation (4), we get:

$$\mu_2(t \text{ unit}) = \frac{1}{N}\Sigma(cu + A - c\bar{u} - A)^2 f$$

$$= \frac{c^2}{N}\Sigma(u - \bar{u})^2 f = c^2\mu_2(u \text{ unit}), \text{ by (11).}$$

5. Standard Deviation. The most important of the μ's is μ_2. It is sometimes called the *variance*. More commonly, its square root is found and this is called the *standard deviation*, and is denoted by the small Greek letter sigma, σ:

$$\sigma = \sqrt{\mu_2}. \tag{16}$$

In stating the value of σ, of course its unit must be indicated. Quite commonly this is done by a subscript, thus:

$$\sigma_u = \sigma(u \text{ unit}).$$

It follows from (15) and (16) that

$$\sigma_t = c\sigma_u. \tag{17}$$

Example 5. The standard deviation in Example 1 is, by Example 4:

$$\sigma_u = \sqrt{2.139} = 1.463, \quad \sigma_t = 2.194 \text{ dollars.}$$

6. α_3, α_4. As in the preceding examples, it is commonly desirable to reduce both the mean and the standard deviation (essentially the first and second moments) to the given or t unit. The third and fourth moments, however, are used almost exclusively in terms of the standard deviation as the unit. When this is done, they are called α_3 and α_4. By the relations of (15) we have,

$$\left.\begin{aligned} \alpha_3 &= \frac{\mu_3(u \text{ unit})}{\sigma_u^3} \\ \alpha_4 &= \frac{\mu_4(u \text{ unit})}{\sigma_u^4} \end{aligned}\right\} \tag{18}$$

These may be thought of as expressed in a "natural unit," because, when they have been obtained, it turns out that it did not matter what the u unit was; that is:

$$\alpha_3 = \frac{\mu_3}{\sigma^3}, \quad \alpha_4 = \frac{\mu_4}{\sigma^4},$$

independently of the unit originally chosen for the μ's and σ's. To prove this, we need merely to note that, if we changed the unit in the first equation of (18) in both numerator and denominator, we should, by equations (15) and (17), multiply both by the cube of the same number; this would not affect the ratio. Similarly, in the second equation of (18), we should multiply both numerator and denominator by the fourth power of the same number, and this would not affect the ratio.

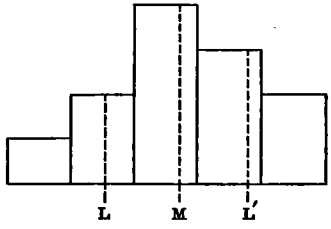
Example 6. In Examples 1 and 4: $\alpha_3 = \dfrac{.313}{3.129} = .100$,

$\alpha_4 = \dfrac{12.274}{4.575} = 2.683.$

7. Meaning of σ, α_3, α_4. In general, σ is a measure of what is called "dispersion." It tells us over how great a range, "on the average," the data are spread out on either side of the mean. There are also other measures of dispersion which tell us the same thing, and, after we have studied them, we shall be able to make this general statement about σ more precise.

In a physical sense, however, we can make it precise now: σ is the physicist's "radius of gyration." If we regard a histogram as a piece cut out from a flat sheet of metal, and suppose it to be spinning about a vertical axis M which goes through its center of gravity (the abscissa of the center of

gravity is our mean value), its motion under any forces would not be disturbed if all the material were concentrated equally on two vertical lines, L, L', each at a distance σ from the axis. Moreover, $N\mu_2 = N\sigma^2$ is the physicist's "moment of inertia" (about the gravity axis M).



The quantity α_3 is a measure of "skewness," and α_4 is a measure of "kurtosis." A distribution has "skewness" if it bulges on one side more than on the other. It has "kurtosis" if the material is spread out on either side to a much greater distance than the extent of the standard deviation; it lacks kurtosis if the material is concentrated near the center. More precisely:

$$\textit{Skewness} = \frac{\alpha_3}{2},$$

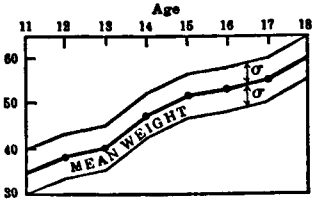
$$\textit{Kurtosis} = \frac{\alpha_4 - 3}{2}.$$

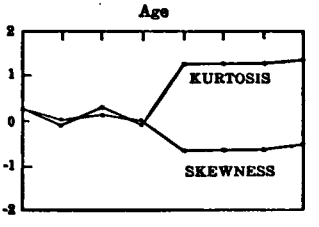
Some authors use $-\alpha_3/2$ and some use $|\alpha_3|$ for skewness, and, in place of kurtosis, use the term excess for $(\alpha_4 - 3)/8$. For a symmetrical distribution the skewness is obviously zero. It will be shown later that the kurtosis is zero for what is to be called a normal distribution.

8. Application. The constants, σ, α_3, and α_4, are called parameters of the frequency distribution, and they indicate the shape of the curve that fits it. Although their chief value lies in their relation to the later theory, it is interesting to note some uses to which they may be put now. The table on page 31 shows the weights of schoolboys at various ages. The data are similar to some gathered at a school in England, but the figures have been altered in order to bring out more forcibly the significance of the parameters. Let us suppose,

WEIGHTS OF SCHOOLBOYS
(Kilograms)

Age / Weight	11	12	13	14	15	16	17	18
25–28	8	2						
28–31	16	7	2		1			
31–34	25	16	7		1	1		
34–37	26	25	16	2	1	2	1	
37–40	18	26	25	7	1	1	2	1
40–43	8	18	29	16	3	2	1	1
43–46	2	8	19	25	7	2	2	1
46–49	1	2	8	26	16	7	2	1
49–52	1	1	2	18	25	16	7	3
52–55			1	8	30	25	16	7
55–58			1	2	20	20	25	16
58–61				1	8	8	20	25
61–64						1	8	30
64–67							1	20
67–70								8
Mean	34.6	37.3	40.5	46.3	51.4	52.5	55.5	60.4
σ	4.7	4.6	4.8	4.6	5.4	5.6	5.6	5.4
$\alpha_3/2$	.24	.05	.15	.05	−.62	−.69	−.69	−.62
$(\alpha_4 - 3)/2$	.25	−.05	.26	−.05	1.28	1.22	1.22	1.28





however, that this table did represent material actually gathered at a school, or, better, that a much larger body of material of this sort had been gathered in a group of schools. Let us suppose further that a statistician is given these figures, and that he has no other information concerning the boys whose weights were found. What can he get out of the data? The simplest, and most important, and the obvious thing to do first is to find the mean value of the weights for each age. These mean values are given on page 31 and plotted on the graph. To what uses may these means be put? Would it be proper, for example, for a physician, examining an English schoolboy of fourteen and finding that he weighed only 44 kilograms (97 pounds) instead of 46.3 kilograms (101.6 pounds), to conclude that this boy was subnormal? The answer to this question depends partly on σ. In the diagram a belt is drawn, about the mean line, of breadth 2σ (measured parallel to the y-axis). It is noticeable that 44 kilograms at fourteen years is a point well within the belt. There is a considerable fluctuation in weight among boys of the same age, and it does not appear very unusual to have found a specimen as far from the average as this one was. Whether subnormal is a proper term to apply to such a case is of course a matter of definition of the term, and this would have to be agreed on. Perhaps the limits of normality should be the limits of this belt, perhaps of a belt twice as broad, but in any case they should depend on how wide the belt is, and if it is narrower in some places than in others, the limits should be closer together in those places than in the others. It is not enough, then, in such a case to know merely the mean value. It is necessary to have also a measure of variability; σ is such a measure.

The next question we ask has to do with α_3 and α_4. If the matter of variability is taken care of, may the physician then use the data freely? Usually he may, but not always, and not in this case. Let us make the more careful analysis,

finding the skewness and kurtosis. These also appear on the graph. Since they almost never exceed 2 numerically, there is no undue distortion in plotting them on a diagram 4 units high. From the graph it is quite obvious that the boys whose ages run from fifteen to eighteen constitute a different sort of group from that of the younger boys. The difference may be due to normal biological changes natural to the greater maturity, but one strongly suspects some other cause, such as a radically different environment, or the admixture of a different race. An investigation of the original material is therefore indicated, and without such an investigation it would not be proper to use this table as though the material were homogeneous. The values of the skewness and kurtosis have shown that the shapes of the curves which would fit the frequency distributions of the later years are quite different from those which would fit the frequency distributions of the earlier years. If the material were homogeneous we should expect a varying mean, perhaps also a varying σ, but we should expect the general shapes of the curves to remain more nearly constant. Thus any considerable changes in α_3 and in α_4 point to hidden changes in the character of the data on which the whole investigation is based.

PROBLEMS CHAPTER II

1. Find by the shortest method the mean value in each case:

(*a*) Problem 5 (*a*) of Chapter I, page 16. *Ans.*, 142.25.

(*b*) Problem 5 (*b*) of Chapter I, page 16. *Ans.*, 72.

(*c*) Problem 6 of Chapter I, page 17, using the theorem of § 2.

(*d*) Bessel's observations on the diameter of one of Saturn's rings:

Seconds of Arc	39 + .179	.285	.294	.407	.410	.320	.377	.310	.127	.448
Weight	7	4	5	4	1	3	3	4	3	6

Ans., 39.31.

2. Find the mean income from the following data, using the theorem of § 2. The material has been condensed from a report on sala-

ries of families in the United States for the year 1918. Each of the groups given below was subdivided into many classes in the original report. Within each group the class intervals were equal, but manifestly the class interval for one group could not have been the same as for all the others. Thus, in the second group the class interval was \$100, and in the seventh it was \$500,000.

Mean of Group	*Number in Group*	*Mean of Group*	*Number in Group*	*Mean of Group*	*Number in Group*
−\$75.5	182,000	20,660	87,000	486,100	1,119
1,255	36,000,000	48,600	39,850	1,905,000	152
6,450	1,137,400	160,000	6,180	Total	37,453,701

3. Find the mean, σ, α_3, α_4, for the data of Example [1] 1, choosing A at another place instead of at \$9.745, as in the text.

4. Find the mean, σ, α_3, α_4, for each of the following cases, and express them relative to the given origin and given unit.

(*a*) Problem 1 (*a*). (*b*) Problem 1 (*b*); *ans.*, 72, 2.81, .34, 2.48.

5. Prove: (*a*) the second, (*b*) the third, of the relations of (15).

6. Give a definition of the moments in the t unit relative to t_1 as origin.

7. By making the substitution (4) in formula (1) for ν_2, show that $\nu_2(t) = c^2\nu_2(u) + 2Ac\bar{u} + A^2$.

8. Find σ for the data of Problem 1 (*d*). *Ans.*, .098.

9. Derive a formula analogous to that of the theorem of § 2 for second moments, *viz.*: $N\nu_2 = g_1\nu_2^{(1)} + \cdots + g_m\nu_2^{(m)}$, where ν_2 is the second moment of the entire distribution about any given origin, $\nu_2^{(1)}$ is the second moment of the first group about the same origin, etc., $\nu_2^{(m)}$ being the second moment of the mth group about the same origin; and where the g's are the frequencies of the several groups.

[1] Throughout this book, the word *example* is used to indicate an illustrative example embodied in the text; the word *exercise* to indicate one of the short problems that are given at the close of various sections; and the word *problem* to indicate one of the relatively longer problems that occur at the close of the chapters.

10. Use the formula of Problem 9 to obtain σ in Problem 6 of Chapter I, page 17.

11. Prove that μ_2 is less than or at most equal to ν_2, the same unit being used in both cases. Hence it follows that $\sqrt{\mu_2} \leqq \sqrt{\nu_2}$, and therefore that the standard deviation, defined as it is with reference to the mean, has a smaller value than it would have if defined with reference to any other point.

12. Write the expression for ν_3 in (10) thus:

$$\nu_3 = \frac{1}{N}\Sigma[(u - \bar{u}) + \bar{u}]^3 f.$$

Expand this expression, and obtain a formula somewhat analogous to (13) which will give the value of ν_3 in terms of $\bar{u}$ and the μ's. Check the result by eliminating ν_2 from the two simultaneous equations (12) and (13).

CHAPTER III

CUMULATIVE FREQUENCY

1. Cumulative Frequency Tables. A cumulative frequency table may be formed from an ordinary frequency table by successive additions of the several frequencies, thus: f_1, $f_1 + f_2$, $f_1 + f_2 + f_3$, etc., as illustrated.

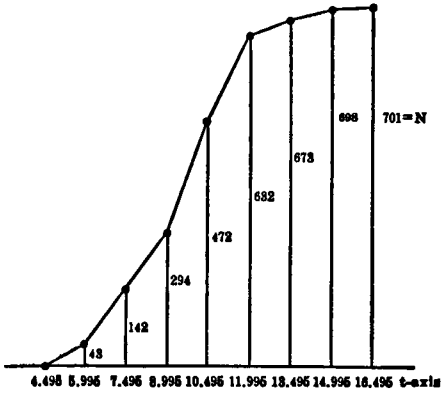
Example 1. Obtain a cumulative frequency table from the data of Example 8, Chapter I, page 14.

CUMULATIVE TABLE

DATA		COMPUTATION		
f	*Mid-t*	*End-t*	*Cum f*	*Cum f/N*
		\$4.495	0	.000
$f_1 =$ 43	\$5.245			
		5.995	$43 = f_1$	.061
$f_2 =$ 99	6.745			
		7.495	$142 = f_1+f_2$	.203
152	8.245			
		8.995	$294 = f_1+f_2+f_3$	.419
178	9.745			
		10.495	472	.673
160	11.245			
		11.995	632	.902
41	12.745			
		13.495	673	.960
25	14.245			
		14.995	698	.996
3	15.745			
		16.495	$701 = N$	1.000

This is sometimes called a "less than" distribution, and this description indicates the reasons for using the end-*t*'s and for placing

the cumulative f figures on the lines between the f figures. What is meant is that 43 individuals received less than \$5.995, 142 less than \$7.495, 294 less than \$8.995, etc. One could begin adding at the bottom of the table, and, in a similar manner, construct a "more than" cumulative distribution. Sometimes the t's are given in the reverse order, the larger ones coming at the top. Then the cumulative f, as we have found it, would be a "more than" instead of a "less than" distribution. The diagram is constructed by plotting the points (4.495, 0), (5.995, 43), etc., as in analytic geometry, and joining them with straight lines. Each of the ordinates is one of the cumulative frequencies, and the difference between any two successive ordinates is an ordinary frequency. The column "cum f/N" gives the ratio of the cumulative frequency to the total frequency. This is not necessary but it is often instructive: thus, from it we can easily see that about 20% of the wages are less than \$7.495, and 90% are less than \$11.995. Of course the graph of the cumulative f/N diagram is exactly like that of the cumulative f diagram, with a proper adjustment of the scale used for the ordinates. The cumulative frequency table is of value in finding another type of average, of measure of dispersion, and of skewness.



Cumulative Diagram

2. Cumulative Frequency Function. It is to be noticed that cumulative f (or cumulative f/N) is a function of t, but its values are known only at the "end-t" points indicated. In fact, it is defined only at these points. These are indicated by the dots in the figure. The dots have been joined by straight lines to aid the eye in locating them. If we want to

know what the value of cumulative f is at some intermediate point, say at $t = 7$, we have to admit that we have no sure means of obtaining it; it depends on how many of the individual wages that lay in the interval \$6 to \$7.49 inclusive were less than \$7. Perhaps all were. Perhaps none were. As a first approximation, however, we shall assume that the wages were evenly distributed throughout the interval. This was in fact the implicit assumption made when, in Chapter I, we represented the frequency over this interval by a rectangle. This same assumption would now require us to represent the cumulative frequency over this interval by a straight line, as we have. From now on, therefore, the straight lines in this diagram shall be thought of as more than a device to aid the eye in locating the points; they are to become a definition (arbitrary but fairly reasonable) of the function at the intermediate points. A slightly better definition might be constructed had we drawn a smooth curve through the points instead of a broken line, but this would be a refinement which is usually not worth while making. We postpone it for the present.

3. Median. *Definition: The median is that value of t for which cumulative $f/N = \frac{1}{2}$.* It will be denoted in this text by M; the mean value being denoted by $M.V.$, or $\bar{t}$. It will sometimes happen that this value of t is given exactly in the table. It may be one of the end-t's. In that case the median is the value of t such that 50% of the data have smaller t's and 50% have larger t's. If there are an odd number of measurements, the median becomes approximately[1] the middle measurement. In a company of soldiers lined up in order of height, the median height is approximately[1] the height of the middle soldier if the number is odd. If the number is even, there is no middle soldier. The median then becomes approximately[1] the average height of the two middle

[1] The approximation is exact if, at least near the median point, the successive t's differ by equal amounts. Where the total frequency is,

soldiers.[1] To compute the median, we should compute the value of that abscissa in the cumulative f/N diagram which would correspond to an ordinate of length $\frac{1}{2}$, or, what is the same thing, compute the value of that abscissa in the cumulative f diagram which would correspond to an ordinate of $N/2$. Numerically, this is the same as interpolating in the end-t table to find the value of t which would correspond to cum $f = N/2$, thus: In our example, $N/2 = 350.5$, and we interpolate to find the t opposite cum $f = 350.5$:

	End-t	*Cum f*	
	8.995	294	
INTERPOLATE →	*M*		← 350.5
	10.495	472	

The result is $M = 9.471$. To aid the student who is not used to interpolating in tables where neither of the differences is unity, the following rule for interpolation is suggested:

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

small, and the spacings uneven, it is a little better to introduce a slight modification of our definition so that these approximations will be exact in every case. (See Chapter IV.) This modification, though slight, would be an undesirable complication in the present more general case.

[1] Authors differ as to what is the best definition of median. Some object to the one we have given because it sometimes introduces fractional t's that cannot exist in the data. For example, in Example 1 (p. 36) the median is \$9.471, and it is clear that no individual was paid this exact wage. Therefore, it is objected that it is a poor definition which compels us to call it the median wage. The natural answer to this objection is that it is as reasonable to call \$9.471 the median wage as it is to call \$9.424 the mean wage. However, all the various definitions yield substantially the same practical results, and the uses to which the median is actually put are not sufficiently exacting to make small differences worth quarreling over. The reason the above definition is adopted in the present text is that it is simple to understand and to use, and it makes the definition of the median a special case of the definition of a percentile, and the semi-interquartile range a special case of the median.

This simply states in usable form the assumption that the partial differences are proportional to the total differences. In the example above, this rule becomes:

$$\frac{M - 8.995}{10.495 - 8.995} = \frac{350.5 - 294}{472 - 294}.$$

Hence, $$M = 8.995 + 0.476 = 9.471.$$

4. Use of Median. There are at least three sorts of frequency distributions in which the median is preferable to the mean as an "average," or single number by which the whole distribution may be described;

(*a*) When occasional and unexpected items near the end of a distribution would unduly affect the mean. Consider, for example, these two artificial distributions:

(I)

t	f
1	4
2	10
3	35
4	10
10	1
N	60

(II)

t	f
1	4
2	10
3	35
4	11
10	0
N	60

In (I) the mean is 2.983 and the median is 2.957. In (II) the mean is 2.883, but the median is 2.957, as in (I). The only difference between the two tables is that the item opposite $t = 10$ in (I) has been put opposite $t = 4$ in (II). Its position does not affect the median, provided only that its t coördinate exceeds the median in both cases, but its position does affect the mean considerably. Now if we are measuring distributions of this sort, in which sporadic items like this occur, the median might often seem to be a fairer value to accept as the "average" than the mean. Examples constantly occur

in investigations in economic theory. The median index number is often chosen instead of the mean. A particular case which the student can understand without special knowledge of economics is a method sometimes used in estimating what is called seasonal fluctuation. The prices of many commodities fluctuate with the seasons as well as from year to year, and it is important to know what this average or normal fluctuation is, so that allowance can be made for it. The interest rates in Canadian banks are a well-recognized example. Because of the importance of the grain industry in Canada, larger borrowings are made at these banks at some seasons than at others, producing higher interest rates. Now one might think that a good way of finding out what the seasonal fluctuation of interest rates was, would be to average them over a period of say 25 years as follows: first, eliminate any long time trend, then get the average for the 25 Januarys, the average for the 25 Februarys, etc. This is in fact essentially [1] the way it is done, but one would do better to use the median for the average rather than the mean, because occasionally some widespread upheaval in finance, due to extraneous causes, might produce very high or very low interest rates, abnormal both as to amount and as to time of occurrence. The mean would be considerably affected by such upheavals, but the median only slightly, if at all.

(*b*) When the only data obtainable are such that the table is left "open" at one or at both ends, as in Example 3 of Chapter II. The table is then a "less than" table at one end, or a "more than" table at the other, or both. Here we sometimes have the option of making assumptions with regard to the limits of the end intervals, as we did in that example, but we may avoid such assumptions by using the median instead of the mean.

(*c*) When we have what is called an "ordered" rather

[1] Actually it is the median of the "link relatives" (Chapter VII) which is used, instead of the median of the absolute values of interest rates.

than a "measured" series of frequencies. This is often the case in psychological and educational studies, in measuring intelligence, for example. There is no foot rule by which one can actually measure that sort of thing, no psychic watt in terms of which one can compute mental power. About the best we can do is to place individuals in order. This may be done by means of scores made on tests, but the scores are not really the measured things they seem to be. If A, B, and C have scores 25, 50, and 75, respectively, we do not think therefore that B's intellectual power is two times and that C's intellectual power is three times as great as A's. We think that their mental abilities can be clearly differentiated the one from the other and that they ought to be placed in that order. If, for example, A answered 25 questions correctly, and B 50, and C 75, we do not know that the increase in mental ability needed in going from 25 to 50 was the same as the increase in going from 50 to 75. It is quite proper to conclude that 50 is the score of the middle individual (median) of the group, but it is not proper to say that the intelligence of B is midway (mean) between that of the other two. In cases of this sort, the median score has a meaning, but the mean score is partly an arbitrary number depending on the necessarily arbitrary nature of the tests and of the grading. The fractional part of the median score made by interpolating between two actual scores does not, however, have any special significance.

5. Percentiles. *The first quartile, to be denoted by* Q_1, *is defined as that value of* t *for which cumulative* $f/N = \frac{1}{4}$. *The second quartile* Q_2 *is that value of* t *for which cumulative* $f/N = \frac{1}{2}$, *and is the same as the median* M. *The third quartile* Q_3 *is that value of* t *for which cumulative* $f/N = \frac{3}{4}$. *Similarly, for any percentile: the* p *percentile is that value of* t *for which cumulative* $f/N = p$ *per cent.* The 10, 20, 30, $\cdots$ percentiles are called deciles, and are denoted by D_1, D_2, D_3, etc. Percentiles, like the median, are found by interpolation in the cumulative tables.

Example 2. Find Q_1, Q_3, and D_1 in Example 1.

End-t	*Cum f/N*
4.495	0
5.995	.061
$D_1 \rightarrow$ 7.495	.203
$Q_1 \rightarrow$ 8.995	.419
10.495	.673
$Q_3 \rightarrow$ 11.995	.902
13.495	.960
14.995	.996
16.495	1.000

$$\frac{D_1 - 5.995}{1.5} = \frac{.10 - .061}{.203 - .061}, \quad D_1 = 6.41.$$

$$\frac{Q_1 - 7.495}{1.5} = \frac{.25 - .203}{.419 - .203}, \quad Q_1 = 7.82.$$

$$\frac{Q_3 - 10.495}{1.5} = \frac{.75 - .673}{.902 - .673}, \quad Q_3 = 11.01.$$

In this example we have used the cumulative f/N table, but in Example 1 we chose the cumulative f table. When a large number of percentiles is to be found, it is often more convenient to use the cumulative f/N table. When a small number only is desired, it is always easier to use the cumulative f table, avoiding the necessity of forming the second one. Thus, the quartiles can be found very quickly by the use of cumulative f. *E.g.*, suppose we want Q_1. Since $N/4 = 175\frac{1}{4}$, we add the frequencies (form cumulative f) until we get the cumulations next below and next above $175\frac{1}{4}$. Then we interpolate between these two. They are 142 and 294, and they lie opposite end-t = 7.495 and 8.995, respectively. So

$$\frac{Q_1 - 7.495}{1.5} = \frac{175.25 - 142}{294 - 142}; \quad Q_1 = 7.82.$$

EXERCISES § 5

1. Find Q_1, Q_2, Q_3 in each of the following sets:

(*a*)

t	*f*
1	12
3	28
5	36
7	16

(*b*)

t	*f*
5	2
4	6
3	10
2	7
1	3

(*c*)

t	*f*
2	100
4	350
6	400
8	200
10	100

Ans., (*a*) 2.787, 4.333, 5.611; (*b*) 3.667, 2.900, 2.072.

2. Find all the deciles in each of the preceding sets.

Ans., (*c*) 3.086, 3.743, 4.400, 5.050, etc.

6. Semi-interquartile Range. The semi-interquartile range will be denoted by s and is defined by the formula

$$s = \frac{|Q_3 - Q_1|}{2}. \tag{1}$$

It is half the absolute distance between the first quartile and the third. It is also called, for a reason to be considered later, the probable deviation, or, more loosely, the probable error. It is a measure of dispersion like σ, but in general not equal to σ. It is to be preferred to σ as a measure of dispersion in those cases (§ 4 *a*, *b*, *c*) where the median is to be preferred to the mean as an average. The value of s depends on the unit used, and the unit should be mentioned when the value is given.

Example 3. For Example 2, $s = \frac{11.01 - 7.82}{2} = 1.60$ dollars.

In Chapter II, Example 5, we saw that in this case σ was equal to 2.19 dollars.

7. Quartile Coefficient of Skewness. In cases where the mean and standard deviation cannot be used, of course a similar difficulty appears with regard to the use of the moment coefficient of skewness, *viz.*, $\alpha_3/2$. The following definition uses only percentiles:

$$\textit{Skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{s} = \frac{Q_3 - 2Q_2 + Q_1}{s}. \tag{2}$$

According to this, the skewness will be positive if $Q_3 - Q_2$ is greater than $Q_2 - Q_1$. This would generally mean that the material on the right of M would be spread out more than the material on the left. The unit in which this excess spread of material is measured is naturally s, just as σ was the unit[1] used in computing α_3.

[1] Some authors use $2s$ instead of s. This cannot be called incorrect because the definition is arbitrary, but it is unfortunate to have two

Example 4. For Example 2, the quartile coefficient of skewness is $\frac{11.01 - 2(9.472) + 7.82}{1.60} = -0.071$. Previously we had found $\alpha_3/2 = 0.051$. The two coefficients often differ by as much as in this case.

8. Mean Deviation.[1] A third measure of dispersion is called the mean deviation. It is the mean of the numerical values of the differences between the several measurements and the mean value, and is therefore given by the formula:

$$\textit{Mean deviation} = \frac{1}{N}\sum_{i=1}^{n} f_i\,|t_i - \bar{t}|, \tag{3}$$

if the t unit is used. If the class interval c is the unit, we would have:

$$\textit{Mean deviation} = \frac{1}{N}\Sigma f|u - \bar{u}|, \tag{4}$$

and, as with σ, the relation:

Mean deviation (t *unit*)

$$= c \textit{ times mean deviation } (u \textit{ unit}). \tag{5}$$

Equation (5) is to be proved in Problem 7. Being based on numerical values of differences, rather than on their squares, the mean deviation is probably a more natural measure of dispersion than σ. It is less common than σ chiefly because it is not so easy to use in mathematical discussions. Also, it cannot be used in place of s because, wherever σ should not

---

different expressions for the same thing. The definition given here is the one given by Yule in his well-known text.

[1] Mean deviation from the mean is meant. There is also a mean deviation from any point t' defined by the equation:

$$\textit{Mean deviation} = \frac{1}{N}\Sigma f|t - t'|.$$

The mean deviation from the *mean* is sometimes used by physicists and occasionally by statisticians. The mean deviation from the *median* is theoretically more attractive because of the following theorem: The mean deviation is a minimum when computed with reference to the median. This theorem requires a slight change in our definition of the median at points where interpolation is necessary.

be used, neither should this measure. It is possible to devise a slightly shorter method of computing the mean deviation than by the use of formula (4), but this will not be done here because the gain is not great and the measure is infrequently needed.[1]

Example 5. Find the mean deviation in Example 1.

Mid-t	u	f	$\lvert u - \bar{u} \rvert$	$f \lvert u - \bar{u} \rvert$
5.245	− 3	43	2.786	119.798
6.745	− 2	99	1.786	176.814
8.245	− 1	152	.786	119.472
9.745	0	178	.214	38.092
11.245	1	160	1.214	194.240
12.745	2	41	2.214	90.774
14.245	3	25	3.214	80.350
15.745	4	3	4.214	12.642
Sums		701		832.182

$$\bar{u} = -.214; \text{ mean deviation} = \frac{832.182}{701}$$
$$= 1.187 \ (u \text{ unit})$$
$$= 1.780 \ (t \text{ unit}).$$

EXERCISE. Find the mean deviation of each set in Exercise 1 of § 5. *Ans.*, (*a*) 1.579, (*b*) .852.

9. Mode. A third type of average of considerable theoretical importance is the mode. This is commonly defined as the value of that measurement which occurs most frequently. It is *la mode*, the fashion, or the typical one. This definition is not exact enough to lend itself to mathematical treatment. Sometimes there is no measurement which occurs more frequently than any other. The position of the mode as thus defined might be seriously affected by making the

[1] Of course, as in the formula for σ, a "grouping error" is involved, since all the data are not truly at the mid-points of the intervals. See Chapter IV.

grouping finer, and with a very fine grouping its position might have little practical meaning. Often we wish to distinguish more than one mode in a distribution; this is obviously impossible if we are loyal to the definition given. Another possible definition is this: Form the histogram and fit it with a simple curve; the abscissa of any maximum [1] point of this curve is a mode. This definition also has serious objections, the most obvious of which is that no two persons might agree on how the curve was to be drawn. It does, however, yield a simple and fairly satisfactory working rule for determining the position of the mode, when the distribution is of a simple type. There is a curve, called Pearson's Type III curve, which fits distributions of this sort. Its equation is of the form

$$y = ae^{-bt}(b + t)^c, \tag{6}$$

where $e = 2.718$, nearly, and the other constants can be found if the first three moments of the distribution are given. Fortunately, in most cases it is not necessary to find these constants or to plot the curve, for it turns out that when this is done, the position of the mode can be found by the following formula:[2]

$$\frac{\textit{Mean minus mode}}{\sigma} = \frac{\alpha_3}{2}. \tag{7}$$

It will be recalled that $\alpha_3/2$ was our moment definition of skewness. We might also now say, therefore, that the skewness is the distance from the mode to the mean, in terms of σ as the unit of measurement.

[1] A curve has a maximum at a point P if its ordinate is greater at P than at all other points in some interval surrounding P.

[2] A more satisfactory formula can be given by the use of a more general curve and four moments. It is

$$\frac{mean - mode}{\sigma} = \frac{\alpha_3(\alpha_4 + 3)}{2(5\alpha_4 - 6\alpha_3^2 - 9)}.$$

Equation (7) can also be found by using a so-called "Charlier A" curve (see Part II).

Example 6. Find the mode in Example 8 of Chapter I.

Mean = 9.424 dollars, $\quad \frac{9.424 - \text{mode}}{2.190} = \frac{.100}{2}$.

$\sigma = 2.190$ " $\quad$ Mode = 9.31 dollars.

$\alpha_3 = 0.100$.

PROBLEMS CHAPTER III

1. Construct a cum f/N table and the corresponding diagram for each of the following cases: (*a*) Problem 5 (*b*) of Chapter I, page 16; (*b*) Problem 5 (*c*) of Chapter I, page 16.

2. Find M, Q_1, Q_3, and s in Problems 1 (*a*) and 1 (*b*).

3. Obtain the quartile coefficient of skewness in each of the following cases:

(*a*) Problem 1 (*a*).

(*b*) Manufacturing establishments in 1921, *Statistical Abstract of the U. S. Department of Commerce*, 1929, page 786.

Value of Product	*Number of Establishments*
Less than \$5000	53 999
\$5000 and less than \$20,000....	71 075
\$20,000 and less than \$100,000..	72 251
\$100,000 and less than \$1,000,000	45 608
\$1,000,000 and over.....	7 333

4. Obtain all the deciles in Example 1 of the text, page 36.

5. A dean reports the following distribution of marks as given by all departments in a certain year. Find the quartiles.

Grades	90–100	80–89	70–79	60–69	Below 60
Frequencies	618	1728	2388	1085	377

6. In the same report, the distributions in mathematics and in physics are presented. Compare Q_1 and Q_3 for these departments.

Grades	90–100	80–89	70–79	60–69	Below 60
Mathematics	60	101	164	103	32
Physics	26	46	43	19	4

7. Prove equation (5).

8. The following problem illustrates a method of utilizing measures of dispersion which will be displayed more fully in Part II. Data from *U. S. Census Bureau, Mortality Statistics:* Deaths by ages from tuberculosis and from cancer.

Ages	Tuberculosis		Cancer	
	Males	*Females*	*Males*	*Females*
Under 5	2237	1913	180	165
5– 9	609	658	129	81
10–14	650	1043	92	93
15–19	2343	4277	147	122
20–24	4559	6562	207	187
25–34	9721	9786	796	1564
35–44	8804	5899	2330	5238
45–54	6854	3688	5787	9887
55–64	4668	2455	10156	12009
65–74	2671	1798	11124	11365
75 and over	872	888	6349	7171
Totals	43988	38967	37297	47882

(*a*) Compute the medians and show that for each of these diseases the median age of death is greater for males than for females; and that the difference is about twice as great in the case of tuberculosis as in the case of cancer.

(*b*) Whether this larger difference is really that much more significant depends on the dispersions. Compute the four semi-interquartile ranges, and then the average s for each disease. Then show that the ratio, difference between medians, divided by average s, is only about 1.7 as great in the case of tuberculosis as in the case of cancer. (*Note.* This is only an approximate answer to the question: Is the difference for tuberculosis the more significant? A thoroughly satisfactory answer cannot be given until after we have studied the theory of probability.)

9. Find the modes in Problems 5 (*a*), (*b*), and (*c*) of Chapter I, page 16.

CHAPTER IV

GROUPING ERRORS. SMALL TOTAL FREQUENCIES

1. Grouping Error. When a frequency distribution is divided into frequency groups, and within each group all the data are supposedly concentrated at the middle point of the interval, as has been our practice, an error is of course introduced, because in fact all the data are usually not truly at the middle point. This is called a grouping error. It affects all the constants we have been using to a greater or less degree. By the theorem of Chapter II, page 22, if the middle of the interval happened to be also the mean of the group — as would be the case if the material were spread out evenly over the interval — the value of the mean would not be affected; but the values of the second and higher moments would be affected even then. These errors could be minimized by making our grouping very fine, but this is often undesirable and sometimes impossible.

2. Sheppard's Corrections. N Large. If the total frequency N is large enough to permit of ten or twenty groups, the grouping errors in the moments can usually be reduced by the following formulae.[1] These formulae may be applied *only* where the class interval is unity, *i.e.*, they are to be applied to the moments in what we have called the u unit.

$$\left.\begin{aligned} &\textit{Corrected } \mu_2 = \textit{uncorrected } \mu_2 - \tfrac{1}{12} \textit{ in the } u \textit{ unit},\\ &\textit{Corrected } \mu_3 = \textit{uncorrected } \mu_3 \textit{ in the } u \textit{ unit},\\ &\textit{Corrected } \mu_4 = \textit{uncorrected } \mu_4 - \tfrac{1}{2}(\textit{uncorrected } \mu_2) + \tfrac{7}{240}, \textit{ in the } u \textit{ unit}. \end{aligned}\right\} \quad (1)$$

$$(\tfrac{1}{12} = 0.083333,\ \tfrac{7}{240} = 0.029167.)$$

Example 1. Find the corrected σ and μ's in the t unit in Example 8, Chapter I, page 14 ($c = 1.5$).

[1] The proof is not appropriate to this text. These formulae are valid only for the simpler types of distributions.

	u unit		*t* unit
	Uncorrected	*Corrected*	*Corrected*
μ_2	2.139	2.056	
σ		1.434	\$ 2.15
μ_4	12.274	11.234	56.87

3. N Small. Moments. When the total frequency N is as small as 50, it is seldom possible to arrange the material in satisfactory equi-spaced groups. Then it is best to treat each item separately, even though there may be several cases where the values of the measurements are the same. We have to abandon certain of our short cuts in computing the moments, but if an adding machine is used with a table of squares and cubes, the labor is not great.

Example 2. Find the mean, σ, α_3, and α_4, of the following grades obtained in an examination in statistics. The grades are given in the usual percentage system; $N = 22$.

Grade (t)	t^2	t^3	t^4
57	3249	1852×10^2	1056×10^4
75	5625	4219	3164
69	4761	3285	2267
67	4489	3008	2015
71	5041	3579	2541
93	8649	8044	7481
53	2809	1489	789
59	3481	2054	1212
76	5776	4390	3336
41	1681	689	283
98	9604	9412	9224
62	3844	2383	1478
74	5476	4052	2999
50	2500	1250	625
85	7225	6141	5220
36	1296	467	168
47	2209	1038	488
47	2209	1038	488
61	3721	2270	1385
64	4096	2621	1678
41	1681	689	283
96	9216	8847	8493
1422	98638	72817×10^2	56673×10^4

Since in this case $f_i = 1$ for every value of i, $\Sigma ft = \Sigma t$, $\Sigma ft^2 = \Sigma t^2$, etc.

$$\bar{t} = \frac{1422}{22} = 64.636,$$

$$\nu_2 = \frac{98638}{22} = 4483.5, \qquad \mu_2 = 305.7, \qquad \sigma = 17.49,$$

$$\nu_3 = \frac{72817 \times 10^2}{22} = 3.3099 \times 10^5, \quad \mu_3 = 1.699 \times 10^3, \quad \alpha_3 = .32,$$

$$\nu_4 = \frac{5667 \times 10^5}{22} = 2.576 \times 10^7, \quad \mu_4 = 2.100 \times 10^5, \quad \alpha_4 = 2.25.$$

We might have shortened the labor in this example by using 40 instead of 0 as the origin, but, with tables of squares and cubes available, this shifting of the origin is often not worth while.

Example 3. Find the mean deviation in Example 2, and also the other constants asked for there, using $\bar{t}$ as the origin (an alternative method).

t	$x = t - \bar{t}$	x^3	x^3	x^4
36	− 28.64	820.25	− 23 492	672 810
41	− 23.64	558.85	− 13 211	312 313
41	− 23.64	558.85	− 13 211	312 313
47	− 17.64	311.17	− 5 489	96 827
47	− 17.64	311.17	− 5 489	96 827
50	− 14.64	214.33	− 3 138	45 937
53	− 11.64	135.49	− 1 577	18 357
57	− 7.64	58.37	− . 446	3 407
59	− 5.64	31.81	− 179	1 012
61	− 3.64	13.25	− 48	176
62	− 2.64	6.97	− 18	49
64	− .64	.41	0	0
67	2.36	5.57	13	31
69	4.36	19.01	83	361
71	6.36	40.45	257	1 636
74	9.36	87.61	820	7 676
75	10.36	107.33	1 112	11 520
76	11.36	129.05	1 466	16 654
85	20.36	414.53	8 440	171 835
93	28.36	804.29	22 810	646 882
96	31.36	983.45	30 841	967 174
98	33.36	1112.89	37 126	1 238 524
Sums	$\Sigma\|x\| = 315.38$	6725.10	36 670	4 622 321
$\frac{1}{N}$ (Sums)	14.33	306	1 667	210 105

$\bar{t}$ = 64.64, found as before.
$\mu_2 = 306$,
$\mu_3 = 1667$,
$\mu_4 = 210105$.
$\sigma = 17.49$,
$\alpha_3 = .31$,
$\alpha_4 = 2.24$,
Mean deviation = 14.33.

The formulae used are:

$$\mu_2 = \frac{\Sigma x^2}{N},$$

$$\mu_3 = \frac{\Sigma x^3}{N},$$

$$\mu_4 = \frac{\Sigma x^4}{N},$$

$$\textit{Mean deviation} = \frac{\Sigma |x|}{N}.$$

When N is small, one can often find the moments more easily by the method just used, in which the origin is placed at the mean. Thus the moments about the mean are computed directly, without the intervention of the u's. When the mean deviation (from the mean) is also desired, this is clearly the best origin to choose. It is better in such cases to arrange the observations in the order of their t's, as shown in the table on page 52.

EXERCISES §§ 1-3

1. Apply Sheppard's corrections, and find the corrected constants in each case, giving the answers in t units. The constants are given in u units, and the class interval c represents the ratio between the length of the u unit and the length of the t unit.

(*a*)	(*b*)	(*c*)	(*d*)
$\mu_3 = .5$	$\sigma = 2$	$\sigma = 1.5$	$\mu_2 = 4$
$\mu_5 = .1$	$\mu_3 = -1$	$\alpha_3 = 0.5$	$\mu_3 = -2$
$\mu_6 = .8$	$\mu_4 = 40$	$\alpha_4 = 2$	$\alpha_5 = 3$
$c = 4$	$c = 3$	$c = 2$	$c = 1$
Ans., $\mu_4 = 148.3$			*Ans.*, $\alpha_4 = 3.0005$

2. Find the mean and uncorrected σ, α_3, and α_4 by the method of Example 2 in Problems (*a*), (*b*), (*c*), and (*d*) at the top of page 54.

Problem	*Values of t*	*Ans.* σ_t
(*a*)	1, 2, 3, 4, 5	1.41
(*b*)	40, 35, 27, 52, 69, 25	
(*c*)	− 6, − 2, − 2, 1, 3, 0, 6, 6	3.90
(*d*)	0, 1, 1, 1, 2, 2, 2, 3	.87

3. Do Exercise 2 by the method of Example 3. Also find the mean deviations. *Ans.*, *M.D.* = 1.2, 12.8, 3.25, .75.

4. *N* Small. Percentiles. When N is so small that a set of equal class intervals cannot be used without introducing serious error, our method of computing the median and other percentiles fails, because we have no end-t's to use.[1] Our definitions still hold, however, provided we construct our cumulative f function by reference to the mid-t's instead of the end-t's. This might have been done — and is done by some authors — initially, but it takes a little longer, for it requires us to split each frequency into two halves, the first half being a part of cumulative f up to the mid-point of the interval, and the second half a part of cumulative f up to the mid-point of the next interval. The method when each $f = 1$ is illustrated in the next example.[2] It is supposed that half of each individual is on one side of the mid-point, and half on the other.

[1] The difficulty of using the average of two consecutive mid-t's arbitrarily as end-t's is illustrated in Problem 4, page 57.

[2] This example also illustrates the case where a few of the f's are more than 1; *e.g.*, when $t = 41$, $f = 2$. It might seem better in such cases to write: mid-$t = 41$, $f = 2$, cum $f = 2.0$, instead of both cum $f = 1.5$ and cum $f = 2.5$ at $t = 41$ as on p. 55. So far as the numerical values of the percentiles are concerned it is immaterial which procedure is followed; the one actually used makes the formation of the cum f table slightly easier.

Example 4. Find the median, Q_1 and D_2, using the data of Example 3.

Mid-t	*f*	*Cum f to mid-t*
36%	1	.5
41	1	1.5
41	1	2.5
47	1	3.5
47	1	4.5 $\leftarrow D_2$
50	1	5.5 $\leftarrow Q_1$
53	1	6.5
57	1	7.5
59	1	8.5
61	1	9.5
62	1	10.5
64	1	11.5 $\leftarrow M$
etc.		

$N = 22$,
$\frac{N}{4} = 5.5$,
$.2N = 4.4$,
$\frac{N}{2} = 11$,
$Q_1 = 50\%$,
$D_2 = 47\%$,
$M = 63\%$.

Care must be taken not to confuse the percentages, in which the grades are (arbitrarily) given, with the percentiles. This difficulty is avoided if one thinks of the *t*'s as being grades, not as percentages. As was indicated in earlier examples, it is not necessary to form the whole cumulative *f* table, only such part as is needed to obtain the percentiles desired.

In the example above, more detail is given than is usually necessary. Suppose all one wanted was the median. *This is the t of the middle observation if N is odd; halfway between the t's of the two middle observations if N is even.* This statement follows directly from the given definition that the median point is that one at which the cumulation of the frequency is $N/2$, or cumulative $f/N = \frac{1}{2}$, for the cumulative function up to any mid-point is simply the sum of the whole number of observations up to that point, plus $\frac{1}{2}$. So, to obtain the median, one does not need to form the whole table of cumulative *f*. Similarly, to find the first quartile, Q_1, it is only necessary to find that *t* at which the cumulation to that point is $N/4$, or where cumulative $f/N = \frac{1}{4}$. Here it is usually necessary to interpolate between two values of *t*, one

for which the cumulation is a little less than $N/4$, and one for which the cumulation is a little more than $N/4$. These two points between which the interpolation is to be made can be readily picked out, once the observations have been arranged in order, because they are close to the observations which are about one fourth of the way from the beginning to the end of the series.

EXERCISES § 4

1. Find the medians and first quartiles, computing only such portions of the cumulative f table as are necessary.

(*a*) $t = 1, 3, 5, 8, 10, 12.$ *Ans.*, 6.5, 3.
(*b*) $t = 1, 3, 5, 8, 10, 12, \cdots (N = 10).$ *Ans.*, 11, 5.
(*c*) $t = -7, -4, -1, 0, 1, 2.$ *Ans.*, $-.5, -4.$
(*d*) $t = 5, 4, 3, -3, -6.$ *Ans.*, 3, 4.25.
(*e*) $t = 47, 43, 43, 42, 42, 41, 41, 41, 40, \cdots (N = 14).$ *Ans.*, 41, 42.

2. Compute the cumulative f tables in 1 (*a*), (*c*), (*d*), and find all the deciles. *Ans.*, (*a*) $D_1 = 1.2$, $D_3 = 3.6$, $D_7 = 9.4$, $D_9 = 11.8$.

3. In each of the Exercises 1 (*b*) and 1 (*e*), reverse the series and obtain Q_3, showing that the value is the same as Q_1 before. This property of reversibility is essential to a good definition of percentiles.

5. N Small. Mode. Unless N is so large that the data can be grouped, the mode has little meaning. For example, in the data just used, the percentage grades that occur most frequently are 41 and 47, but in no sense might it have been said that the typical, usual, or even fashionable grade was either 41 or 47. A modal grade might have been distinguished if the method of grouping had been coarser. Suppose grade A corresponded to percentages 90–100, B to 80–89, etc. The frequencies of these letters were:

Grade	*A*	*B*	*C*	*D*	*E*	*F*	*G*
f	2	1	4	5	4	4	1

The modal grade was D, 60–69.

PROBLEMS CHAPTER IV

1. From the results in Problem 4, Chapter II, page 34, find the corrected values of the constants in each case, and express them in the t units.

2. On a certain examination in training in physics the following scores were made. Find: (*a*) the mean score; (*b*) σ in the same unit as the scores; (*c*) the skewness by the moment formula (*Ans.*, .23); (*d*) the median score; (*e*) the skewness by the quartile formula (*Ans.*, .72); (*f*) that one of the following intervals which contained the modal score: 20–, 50–, 80–, 110–, 140–. The scores were 29, 36, 38, 41, 43, 48, 50, 51, 54, 54, 56, 60, 60, 63, 65, 69, 80, 85, 85, 87, 90, 99, 101, 111, 112, 114, 121, 122, 124, 149; $N = 30$.

3. The number of *millions of tons of unfilled orders* of the United States Steel Corporation for ten years is given for each quarter year below. Compute the median and semi-interquartile range for each quarter.

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
1920	9.89	10.98	10.37	8.15
1921	6.28	5.12	4.56	4.27
1922	4.49	5.64	6.69	6.75
1923	7.40	6.39	5.03	4.44
1924	4.78	3.26	3.47	4.82
1925	4.86	3.71	3.72	5.03
1926	4.38	3.48	3.59	3.96
1927	3.55	3.05	3.15	3.97
1928	4.34	3.64	3.70	3.98
1929	4.41	4.26	3.90	4.42

Ans., 1st quarter, $M = 4.635$, $s = .95$.

4. (*Theory*) Use the following five measurements, each occurring but once: $t = 1, 4, 7, 8, 10$. (*a*) Find M and P_{40}. (*b*) Reverse the series and find P_{60}: this should equal the P_{40} found before. (*c*) Repeat (*a*) and (*b*), using end-t points, as we did when N was large, instead of mid-t points, and assume that these end-t's are halfway

between the given t's. The effect of this method is to produce a median which is not always the value of the middle observation. (*d*) Repeat (*c*), choosing the end-t's in accordance with another hypothesis: insert zero frequencies so as to make a set of equal intervals and choose the end-t's at the ends of these intervals, thus:

f	1	0	0	1	etc.
t	1	2	3	4	"
End-t	1.5	2.5	3.5		"

The effect of this method is to make the value of P_{40} ambiguous. If it be defined as the *smallest* value of t for which cumulative $f/N = 40\%$, then P_{40} will be different from P_{60} when the order is reversed. Cumulative diagrams are helpful in (*c*) and (*d*).

CHAPTER V

THE NORMAL LAW

1. Equation and Graph. When a frequency diagram is nearly symmetrical, it may commonly be fitted approximately by a curve whose equation is

$$y = ae^{-h^2t^2}, \tag{1}$$

and this is called the "normal law." [1] Here a and h^2 represent positive numbers, and e is the base of the so-called Naperian system of logarithms and may be represented by the sum of the infinite series:

$$1 + \frac{1}{1} + \frac{1}{1\cdot 2} + \frac{1}{1\cdot 2\cdot 3} + \frac{1}{1\cdot 2\cdot 3\cdot 4} + \cdots. \tag{2}$$

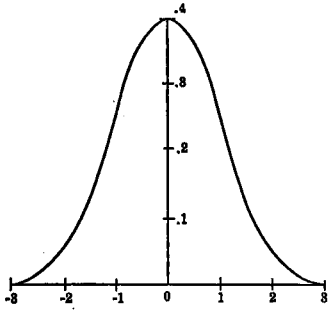
Approximately, $e = 2.718$, and $\log e = 0.4343$. In Problem 1, the student is asked to plot the normal curve when $a = h = 1$. Since the values assigned to a and h merely determine the scales of y and t, this picture will be similar to that obtained by choosing any other values of a and h. When $a = \frac{1}{\sqrt{2\pi}}$, and $h = \frac{1}{\sqrt{2}}$, the function has certain simple properties, and it is customary to represent it by the Greek letter phi,[2] ϕ. We shall also use x in this case in place of t, thus:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{3}$$

[1] So called by K. Pearson and other English writers. It is frequently called the Gaussian Law, sometimes Laplacean. It may be attributed more justly to De Moivre than to either Gauss or Laplace.

[2] Pearson uses z instead of ϕ.

The values of $\phi(x)$ are tabulated in Table I (a). The graph is given in the accompanying figure, and may be quickly drawn by plotting a few well-chosen points from the table, such as the following:



The Normal Curve

x	$\phi(x)$
0	.3989
0.5	.3521
1.0	.2420
1.5	.1295
2.0	.0540
2.5	.0175
3.0	.0004

Since the curve must be symmetrical with respect to the y-axis because x occurs only to an even power, it is unnecessary to tabulate negative values of x: $\phi(-x) = \phi(x)$. (Cf. Problem 7 (a), Chapter I, page 17.) The curve extends over the whole interval, $-\infty$ to $+\infty$, and might therefore seem unsuitable to the representation of finite frequency distributions. This is, however, only a very slight disadvantage, for beyond $x = 3$ it rapidly flattens down extremely close to the x-axis. The combined area of the two "tails" beyond $x = -4$ and $x = +4$ is only 0.000,064, relative to the whole area under the curve, and the combined area beyond $x = -5$ and $x = +5$ is only 0.000,000,78, relative to the whole.

EXERCISES § 1

1. Interpolate in Table I (a), using the auxiliary table of tenths of the mean tabular differences, so as to obtain:

$\phi(1.722)$, $\phi(1.728)$, $\phi(-1.754)$, $\phi(-1.797)$, $\phi(1.794)$.

Ans., .0906, .0896, .0857, .0793, .0798.

2. Find x if $\phi(x)$ has the following values:

0.2180, 0.1600, 0.0318, 0.1720, 0.3943.

Ans., $\pm$ 1.099, $\pm$ 1.352, $\pm$ 2.249, $\pm$ 1.297, $\pm$ 0.153.

3. Plot a cumulative ϕ diagram from the table of $\phi(x)$ given in § 1, joining the plotted points with a smooth curve instead of with straight lines, as heretofore. It will be seen later that this is also a graph of Table I. The curve is called an ogive.

2. Properties. The normal curve has many simple and interesting geometrical properties. We shall list some of them here, but shall not prove them all. To obtain the geometrical properties of most curves, the calculus is indispensable, and so the demonstrations of many of these properties, though simple by the calculus, are beyond the scope of this book. In this list we shall mention the area "under the curve" or "of the curve," and the "higher moments of the curve." The exact meaning of these expressions can be stated as follows. Form a histogram by plotting equi-spaced ordinates of the curve by the use of Table I (*a*). Let the interval between the ordinates be c (cf. Figure IV, p. 67). This may be done when c is large and also when c is small. Obviously, the smaller c, the more closely will the form of the histogram approximate that of the curve. So we define the area under the curve as the limit of the area of this histogram as c approaches zero. In like manner the moments, mean deviation, etc., of the curve are defined as the limits of the corresponding constants of the histogram as c approaches zero.

(*a*) **The area under y in the units used in (1) is $\frac{a}{h}\sqrt{\pi}$. The area under ϕ in the units used in (3) is 1.**

(*b*) **The mean of y and of ϕ is 0; also the median, mode, and skewness are zero.**

(*c*) **For y, $\mu_2 = \frac{1}{2h^2}$ so that $\sigma = \frac{1}{h\sqrt{2}}$, in the unit of t. For ϕ, in the unit of x, $\mu_2 = 1$, $\sigma = 1$. Thus σ is the unit of x.**

(*d*) The mean deviation $= \sigma\sqrt{\frac{2}{\pi}} = 0.794\sigma$, approximately, in both cases.

(*e*) The semi-interquartile range $s = 0.6745\sigma$, approximately, in both cases, and this equals 0.845 times the mean deviation, by (*d*).

(*f*) The abscissae of the points where the curve crosses its own tangent [1] are $\pm\ \sigma$, in both cases.

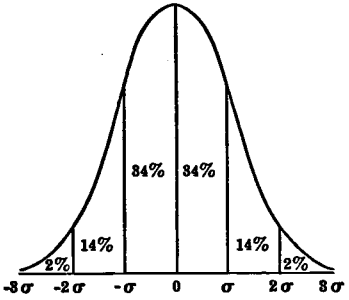


FIGURE I

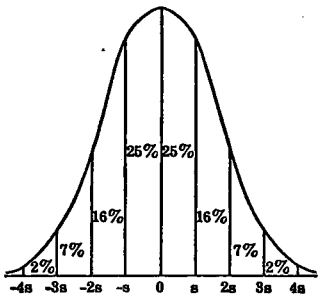


FIGURE II

(*g*) The relative distribution of area, in both cases, is indicated approximately in Figure I, when σ is chosen as the width of one interval; and in Figure II, when s is so chosen.

(*h*) By (*c*), $h = \frac{1}{\sigma\sqrt{2}}$; h has been called the "measure of precision" by physicists and astronomers. Obviously, the smaller the standard deviation, the greater is the measure of precision.

3. Table I. We have seen that the area under $\phi(x)$ from $-\infty$ to $+\infty$ is unity. The partial area from $-\infty$ to x

[1] These are called points of inflection. Between these points the curve is convex, like the rounded top of a hill; beyond them it is concave, like the sides of a bowl.

corresponds to what we have termed cumulative frequency. This is called by Pearson $\frac{1}{2}(1 + \alpha)$ or $(\frac{1}{2} + \frac{\alpha}{2})$, and is tabulated in Table I. Pearson's α is the area from $-x$ to x (see the diagram). In the calculus, a handy notation is introduced to indicate the area under a curve, and we shall now make use of it. The area under $\phi(x)$ from $x = a$ to $x = b$ is called the "*integral of* $\phi(x)$ *from* a *to* b" and is denoted thus:[1]

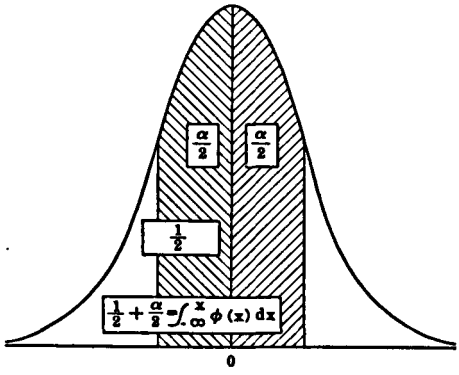


FIGURE III

$$\int_a^b \phi(x)dx.$$

So we may now write:

$$\alpha = \int_{-x}^{x} \phi(x)dx,$$

$$\tfrac{1}{2}(1 + \alpha) = \int_{-\infty}^{x} \phi(x)dx,$$

and by (a) $1 = \int_{-\infty}^{\infty} \phi(x)dx$, and therefore $\int_{-\infty}^{0} \phi(x)dx = \tfrac{1}{2}$.

[1] The reason for inserting the "dx" will not be obvious, and it is sometimes carelessly omitted by students of the calculus. Although it has no significance for this course, we shall leave it in to avoid starting a bad habit. We are merely interested in a symbol to use for the words: "area of $\phi(x)$ from a to b," and this could as well be given by $\int_a^b \phi(x)$. We shall, however, omit both $\phi(x)$ and dx and use simply the sign $\int_a^b$ when this can be done without ambiguity.

Example 1. Find the following areas under $\phi(x)$ and use the integral notation:

(*a*) Area from $x = 0$ to $x = 1$.
(*b*) " " $x = 1$ to $x = 2$.
(*c*) " " $x = -3$ to $x = -2$.
(*d*) " " $x = 4$ to $x = \infty$.

Solutions:

(*a*) By the table, the area from $-\infty$ to 1 is 0.8413, and the area from $-\infty$ to 0 is 0.5000. It is obvious from the figure (Figure I, set $\sigma = 1$) that the difference is the area from 0 to 1, *viz.*, 0.3413. In the integral notation, we would write:

$$\int_0^1 = \int_{-\infty}^1 - \int_{-\infty}^0 = 0.8413 - 0.5000 = 0.3413.$$

(*b*) $$\int_1^2 = \int_{-\infty}^2 - \int_{-\infty}^1 = 0.9772 - 0.8413 = 0.1359.$$

(*c*) Since the curve is symmetrical, the area from -3 to -2 is the same as the area from 2 to 3. So

$$\int_{-3}^{-2} = \int_2^3 = \int_{-\infty}^3 - \int_{-\infty}^2 = 0.99865 - 0.97720 = 0.02145.$$

(*d*) $$\int_4^\infty = \int_{-\infty}^\infty - \int_{-\infty}^4 = 1 - 0.999968 = 0.000032.$$

Example 2. Using Table I, prove (*g*), Figure I. Formula (1) gives the same curve as formula (3) except for scales. A change in the scale of x or of y or both will not alter the *relative* areas. Therefore, to establish Figure I for both cases, it is sufficient to establish it for one case only, and we choose formula (3). In this case $\sigma = 1$, $2\sigma = 2$, and the total area equals 1. Thus the relative areas are given by Table I, as in Example 1:

$$\int_0^1 = 0.3413, \text{ that is about } 34\%,$$

$$\int_1^2 = 0.1359, \text{ that is about } 14\%,$$

$\int_2^\infty = 1 - \int_{-\infty}^2 = 1 - 0.9772 = 0.0228$, that is about 2%.

The rest of the figure follows from symmetry.

The student is asked to prove (*g*) for Figure II in Problem 7.

EXERCISES § 2

1. Find the portions of the area under $\phi(x)$ indicated, and draw a figure in each case.

(*a*) $\int_{-\infty}^{2.3} \phi(x)dx$, $\int_{2.3}^{\infty}$, $\int_{-\infty}^{-2.3}$, $\int_{-2.3}^{\infty}$.

(*b*) $\int_{-2.3}^{2.3}$, $\int_{0}^{2.3}$, $\int_{-2.3}^{0}$, $\int_{-2.3}^{2.6}$.

(*c*) $\int_{-\infty}^{2.333}$, $\int_{-\infty}^{2.338}$, $\int_{-\infty}^{4.333}$, $\int_{5.333}^{\infty}$.

(*d*) $\int_{3.42}^{4.32}$, $\int_{4.32}^{4.82}$, $\int_{-3.478}^{-2.072}$, $1 - \alpha$ for $x = 2.078$.

2. Find x, given the following partial areas:

(*a*) $\int_{-\infty}^{x} \phi(x)dx = 0.9954$, $\int_{x}^{\infty} = 0.0027$, $\int_{-x}^{\infty} = 0.9954$.

(*b*) $\int_{-\infty}^{x} = 0.999\,9954$, $\int_{x}^{\infty} = 0.000\,0002$, $\int_{-x}^{x} = 0.4376$.

(*c*) $\alpha/2 = 0.2789$, $1/2 + \alpha/2 = 0.7843$, $1 - \alpha = 0.2788$.

Ans., 1 (*a*) .9893, .0107, .0107, .9893;
(*b*) .9786, .4893, .4893, .9846;
(*c*) .9902, .9903, .999 9926, .000 000 048;
(*d*) .0003, .000 0071, .0188, .0378.
2 (*a*) 2.605, 2.780, 2.605;
(*b*) 4.435, 5.070, .579;
(*c*) .768, .787, 1.083.

3. *Proofs.* *Proof of* (*a*). There is no difficulty about the area under $\phi(x)$: by Table I, the half-area,

$$\int_{-\infty}^{0} = 0.5000,$$

and, therefore, by symmetry, the total area is 1. To get the area under y, we have to investigate more closely the changes in the scales. The units used in (1) are the units of t and y, and the units used in (3) are the units of x and ϕ. Comparison of the equations (1) and (3) shows that $t = x/h\sqrt{2}$, and $y = \phi a\sqrt{2\pi}$. The interpretation of this is that in a given distance there are $1/h\sqrt{2}$ times as many t units as x units, $a\sqrt{2\pi}$ times as many y units as ϕ units, and that in a given area there are $(a\sqrt{2\pi}/h\sqrt{2} = a\sqrt{\pi}/h)$ times as many ty units as $x\phi$ units. But in the area under the normal curve we have just proved that there is exactly one $x\phi$ unit; hence the number of ty units is $a\sqrt{\pi}/h$. The proof of (b) is left for Problem 3.

Proof of (c). The student is asked to derive the first part from the second part in Problem 6. Without the aid of the calculus, we can prove the second part only approximately. Let us form an approximate histogram, using the areas over the following intervals: $(0, \frac{1}{2})$, $(\frac{1}{2}, 1)$, $(1, \frac{3}{2})$, $(\frac{3}{2}, 2)$, $(2, \frac{5}{2})$, $(\frac{5}{2}, 3)$. The mid-points are to be denoted by x_1, x_2, x_3, x_4, x_5, x_6. By symmetry we know also the corresponding areas at $-x_1$, $-x_2$, etc. So our table of frequencies to three decimal places and our computation would be as follows:

APPROXIMATELY NORMAL FREQUENCY DISTRIBUTION

Mid-points		f	u	fu^2
$-x_6$	−2.75	.005	−5	.125
$-x_5$	−2.25	.017	−4	.272
$-x_4$	−1.75	.044	−3	.396
$-x_3$	−1.25	.092	−2	.368
$-x_2$	− .75	.150	−1	.150
$-x_1$	− .25	.191	0	.000
x_1	.25	.191	1	.191
x_2	.75	.150	2	.600
x_3	1.25	.092	3	.828
x_4	1.75	.044	4	.704
x_5	2.25	.017	5	.425
x_6	2.75	.005	6	.180
Totals		.998		4.239

$\Sigma fu^2 = 4.239$, $N = \Sigma f = 0.998$,
$\nu_2 = 4.239/0.998 = 4.247$,
$\mu_2 = 4.247 - 0.25 = 3.997$.
Using Sheppard's corrections,
$\sigma_u^2 = 3.914$, $\sigma_u = 1.956$; $\sigma_x = 0.978$,
since the class interval is 0.5.

Thus, this approximate histogram yields $\sigma = 0.98$ instead of $\sigma = 1.0$, which is the value for the exact curve.

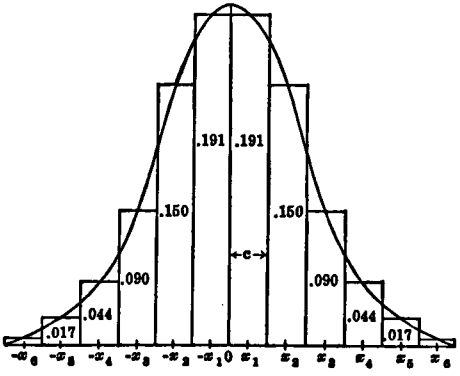


FIGURE IV

The student is asked to prove (*d*) and (*e*) in Problems 4 and 5. The proof of (*f*) requires the calculus. It may be noticed, however, that in Table I (*a*) the tabular differences increase until $x = 1$, and then decrease. The slopes of the tangents at the several points are approximately proportional to these differences since the intervals for x are all equal. Hence, as we go to the right from the origin, the slope of the tangent to the curve begins at zero and increases (in numerical value, being negative in sign) until we reach the point $x = 1$, and then decreases to zero again. Thus, up to the point $x = 1$, the curve becomes steeper and steeper, and is concave downward, like the rounded top of a hill. At this point it begins to grow less and less steep and is concave upward, like the side of a bowl. It is helpful, in drawing free-hand graphs of the normal curve, to bear these facts in mind.

We can now state and prove approximately two properties of the normal curve in addition to those (a to h) given in § 2.

(i) The mean value of that portion of the area under the curve which lies over the interval, a to ∞, is equal to the ordinate at the first point a, divided by the area, *i.e.*,

$$\text{Mean } (a, \infty) = \frac{\phi(a)}{\int_a^\infty \phi(x)dx}.$$

To understand what is meant by this theorem, let us turn back to Figure III in § 3, and look at the unshaded partial area to the right of x. This partial area has, of course, a mean value. It could be found as accurately as desired by inscribing within it a histogram with small class intervals and finding the mean value of the frequency distribution so represented. Our theorem says it may be found much more easily: let $x = a$, look up $\phi(a)$ in Table I (a); $\int_a^\infty \phi(x)dx$ in Table I; and divide the first result by the second. This quotient is also the reciprocal of the quantity R_a, whose logarithm is given by Table VII, *i.e.*, $1 - R_a = \log \text{mean } (a, \infty)$. Thus, by the use of Table VII and an ordinary table of logarithms, one can easily obtain this mean value.

(j) The mean value of that portion of the area under the curve which lies over the interval, a to b, is found by subtracting the ordinate at b from the ordinate at a, and dividing the result by the area, thus:

$$\text{Mean } (a, b) = \frac{\phi(a) - \phi(b)}{\int_a^b \phi(x)dx}.$$

Proof of (i) *and* (j). Since the value of ϕ at ∞ is zero, it is obvious that (i) is merely a special case of (j) in which b is ∞. So it is necessary to prove (j) only. Without the use of the

calculus [1] this cannot be done exactly. But, somewhat as in the proof of (*c*), we can for any desired special case prove the property approximately. Let us in fact use the same approximately normal distribution as before, and let us choose $a = 1$ and $b = 3$. We ought to obtain as the mean:

$$\frac{\phi(1) - \phi(3)}{\int_1^3 \phi(x)dx} = \frac{.2420 - .0044}{.99865 - .8413} = \frac{.2376}{.1573} = 1.51.$$

To verify this we have the following computation:

Mid-x	*f*	*u*	*fu*
1.25	.092	−1	−.092
1.75	.044	0	.000
2.25	.017	1	.017
2.75	.005	2	.010
Totals	.158		−.065

$$\bar{u} = -.411, \; c\bar{u} = -.21,$$
$$\bar{x} = 1.75 - .21 = 1.54.$$

The agreement would have been better had shorter intervals been chosen (see Problem 19).

EXERCISES § 3

1. Using Tables I and I (*a*), find the means of the partial areas under $\phi(x)$ indicated:

(*a*) To the right of $x = 1$, of $x = 1.5$, of $x = 3$, of $x = -1$, of $x = -1.5$.

(*b*) To the left of $x = -2$, of $x = 0$, of $x = 2.5$.

Ans., (*a*) 1.525, 1.939, 3.259, .2877, .1388; (*b*) − 2.368, − .7978, − .01761.

(*c*) Between $x = 1$ and $x = 1.5$, $x = 1.5$ and $x = 3$, $x = 0$ and $x = 2.5$.

[1] By the calculus it is easier to prove (*i*) first, and then obtain (*j*) by a repeated use of (*i*).

(*d*) Between $x = -1$ and $x = 0$, $x = 0$ and $x = -1$, $x = -2.5$ and $x = -3.5$.

Ans., (*c*) 1.224, 1.911, .7724; (*d*) − .4597.

2. Do the first three parts of Exercise 1 (*a*) and the first two parts of 1 (*b*) by means of Table VII, obtaining thus a higher degree of accuracy in the extreme cases.

Ans., for $x = 1$, $\log R = \bar{1}.8167 = 9.8167 - 10$, $\operatorname{colog} R = 0.1833$, mean = 1.525; for $x = 1.5$, 3, etc., the means are 1.9385, 3.283, − 2.373, − .7978.

4. Curve Fitting. It was stated on page 59 that the normal curve would fit approximately most of those frequency distributions which are nearly symmetrical. But of course the precise form of the curve fluctuates with the choice of the constants, as the student has seen in plotting equations (1) and (3). Moreover, by choosing different origins for t, the position of the curve can be shifted to the right or left. We still have, therefore, a certain problem in curve fitting to solve here, *viz.*: Given a frequency distribution in the usual t unit and with the t origin, what is the proper form in which one should write the equation of the normal curve in order that when plotted it may fit the histogram? In the first place we must now use $t - \bar{t}$ in place of t in equation (1), because in general the origin of t will not be the mean of the frequency distribution. Our equation (1) becomes

$$y = ae^{-h^2(t-\bar{t})^2}, \tag{4}$$

and we have thus used one of the first principles of curve fitting:

(*i*) *The mean of the theoretical curve should equal the mean of the frequency histogram.*

We shall now use two more principles:

(*ii*) *The area of the curve should equal the area of the histogram.*

(*iii*) *The standard deviation of the curve should equal the standard deviation of the histogram.*

To satisfy (*ii*) and (*iii*), let N be the area and σ the (corrected) standard deviation of the histogram.

By (*a*) and (*ii*), $\qquad \frac{a}{h}\sqrt{\pi} = N.$

By (*c*) and (*iii*), $\qquad \frac{1}{h\sqrt{2}} = \sigma_t.$

Solving these two equations for a and h, we obtain:

$$h = \frac{1}{\sigma\sqrt{2}}, \quad a = \frac{N}{\sigma\sqrt{2\pi}},$$

so that equation (4) becomes

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\bar{t})^2}{2\sigma^2}} \tag{5}$$

This is the required equation in the t unit and with the t origin.

Example 3. Find and graph the equation of the normal curve which fits as nearly as possible the data of Example 8 of Chapter I, page 14. For this group of data we have found already the following constants:

$$\bar{t} = \$9.424,$$
$$\sigma_t = \$2.151 \text{ (corrected)},$$
$$N = 701.$$

Hence, the equation is

$$y = \frac{701}{2.151\sqrt{2\pi}} e^{-\frac{(t-9.424)^2}{2(2.151)^2}}.$$

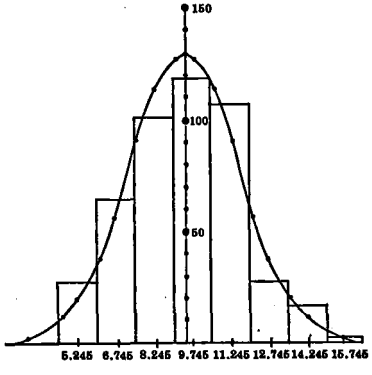
To plot this, we set $x = \frac{t-\bar{t}}{\sigma}$ and write the equation in the form

$$y = \frac{N}{\sigma_t}\phi(x), \tag{6}$$

and obtain $\phi(x)$ from Table I (*a*), as indicated on page 72.

t	x	$\phi(x)$	y	f/c
5.245	−1.943	.0605	20	29
6.745	−1.245	.1838	60	66
8.245	− .548	.3433	112	101
9.745	.149	.3946	129	119
11.245	.847	.2787	91	107
12.745	1.544	.1212	39	27
14.245	2.241	.0324	11	17
15.745	2.939	.0053	2	2
$\bar{t}=9.424$	0	.3989	130	

The histogram is also shown. The heights of the rectangles are in the column f/c; $c = 1.5$. When both the curve and the histogram are to be drawn, it is better to plot the curve first so that it may be drawn smoothly through the points indicated without the presence of the rectangles to prejudice one. In plotting the curve, notice that for every point plotted there is another point that also lies on the curve. This is symmetrical to the first with respect to the central ordinate.



5. Graduation. Sometimes a normal curve fits a given distribution so well that we feel that the deviations are accidental, due in part to grouping, in part to the fact that the actual data are only a sample of a larger group or "population" whose characteristics we are studying. We then feel that we would get a better picture of the total population studied if we used the smooth curve rather than the observed histogram. In such a case, we need to determine what the several frequencies *would have been* had the distribution followed the curve

exactly. This determination is called "graduation by means of the normal curve." It is a process of smoothing out the data to fit the curve. We find the equation of the curve, (5) or (6), of § 4, and then from Table I find the areas over each of the several intervals. These intervals are located by means of their end points.

Example 4. Graduate the material of Example 3 by means of the normal curve. This material is not very symmetrical, and, since the fit is not very good, we should not be truly justified in this instance in using a normal graduation. We use the data only for a numerical illustration. In a later chapter we shall see how to graduate these data more perfectly by means of another type of curve.

Observed f	*Mid-t*	*End-t*	*End-z*	$\int_{-\infty}^{z}\phi(x)dx$	*Theoretical f/N*	*Theoretical f*
		4.495	−2.291	.0110		
43	5.245				.0444	30.8
		5.995	−1.594	.0554		
99	6.745				.1295	90.8
		7.495	− .897	.1849		
152	8.245				.2362	165.6
		8.995	− .199	.4211		
178	9.745				.2697	189.1
		10.495	.498	.6908		
160	11.245				.1932	135.4
		11.995	1.195	.8840		
41	12.745				.0868	60.8
		13.495	1.893	.9708		
25	14.245				.0244	17.1
		14.995	2.500	.9952		
3	15.745				.0043	3.0
		16.495	3.287	.99949		

$\bar{t} = \$9.424,\ \sigma = \$2.151,\ N = 701.$

The column labeled "theoretical f/N" is found by subtracting the successive numbers in the preceding column. It indicates the *proportions* to be expected in the several intervals on the hypothesis of a normal distribution. The *numbers* to be expected are given in the last column.

EXERCISES § 5

(A cumulative review of §§ 1–5)

1. Use Tables I and I (*a*) to obtain: (*a*) $\phi(1.236)$, $\phi(0.231)$, $\phi(-0.239)$, $\phi(-3.107)$, *Ans.*, .1858, .3884, .3877, .0032; (*b*) x if $\phi(x) = 0.0170, 0.3827, 0.0003, 0.2000$, *Ans.*, ± 2.512, $\pm .288$, ± 3.8, ± 1.175; (*c*) $\frac{1}{2}(1+\alpha)$ at $x = 0.502$, at $x = 0.508$, $\int_{-\infty}^{.583}\phi(x)dx$, $\int_{-\infty}^{1.583}\phi(x)dx$, $\int_{-\infty}^{1.587}\phi(x)dx$, *Ans.*, .6922, .6943, .7200, .9433, .9437; (*d*) $\int_{0}^{1.587}\phi(x)dx$, $\int_{1.2}^{2.3}\phi(x)dx$, $\int_{1.2}^{\infty}\phi(x)dx$, $\int_{-1.2}^{\infty}\phi(x)dx$, $\int_{-2.0}^{-1.2}\phi(x)dx$, $\int_{-1.2}^{2.3}\phi(x)dx$, $\int_{-.782}^{.782}\phi(x)dx$, *Ans.*, .4437, .1044, .1151, .8849, .0923, .8742, .5658; (*e*) x if $\frac{1+\alpha}{2} = .5089$, if $\int_{-\infty}^{x}\phi(x)dx = 0.8925$, if $\int_{0}^{x}\phi(x)dx = 0.3827$, if $\int_{-x}^{\infty}\phi(x)dx = 0.8420$, if $\int_{-x}^{x}\phi(x)dx = 0.2428$, if $\int_{x}^{\infty}\phi(x)dx = 0.0032$, if $\int_{x}^{\infty}\phi(x)dx = 0.000{,}032$, *Ans.*, .022, 1.240, 1.189, 1.003, .309, 2.730, 4.00; (*f*) D_7 of $\phi(x)$, D_2 of $\phi(x)$, median of last 25% of a normal distribution, median of the second 25%, median x of each of the following intervals, ($x = .27$ to $x = .37$), ($x = 1.37$ to $x = 1.47$), ($x = -2.1$ to $x = -1.2$), *Ans.*, .524, $-.842$, 1.150, $-.319$, .32, 1.418, -1.502.

2. The following distributions are nearly normal. In each case establish approximately the properties of the normal curve given in § 2 (*a*), (*d*), omitting Sheppard's correction.

t	f	*Ans.*
0	.125	$N = 1$
1	.375	$\sigma_t = .87$
2	.375	$\frac{\text{Mean dev.}}{\sigma} = .87$
3	.125	

t	f	*Ans.*
0	.0625	$N = \sigma = 1$
1	.250	
2	.375	$\frac{\text{Mean dev.}}{\sigma} = .75$
3	.250	
4	.0625	

3. (*a*) By the use of § 3 (*i*), (*j*), compute the means of the following portions of the normal curve $\phi(x)$: portions over the x-intervals, (2, 3), (− 3, − 2), ($\frac{1}{2}$, $\frac{3}{2}$), (− $\frac{1}{2}$, $\frac{1}{2}$), (3, ∞). *Ans.*, 2.31, − 2.31, 1.11, .356, 3.26. (*b*) Obtain approximate answers to (*a*), except in the last case, by the use of the data used in the proof of (*c*), page 66. *Ans.*, 2.36, − 2.36, 1.13, .362.

4. Graduate each of the following distributions by means of the normal curve. Also find the ordinates of the best-fitting normal curve. Plot the histogram and the curve:

t	f	*Ans.*	
		Theoretical f	y
0	1	.9	.8
1	3	3.0	3.1
2	3	3.0	3.1
3	1	.9	.8

t	f	*Ans.*	
		Theoretical f	y
0	10	10	9
1	40	39	39
2	60	61	64
3	40	39	39
4	10	10	9

PROBLEMS CHAPTER V

1. Plot equation (1) when $a = h = 1$, using a table of logarithms to obtain the values of y.

2. On the same diagram with Problem 1, plot the curves:

$$y = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/2\sigma^2},$$

(*a*) when $\sigma = 1$; (*b*) when $\sigma = 2$, using Table I (*a*) in each case (cf. equations (5) and (6)).

3. Prove (*b*) of § 2.

4. Using the approximately normal distribution of § 3, prove (*d*) of § 2.

5. Using the approximately normal distribution of § 3, prove (*e*) of § 2.

6. Derive the first part of (*c*), § 2, from the second part.

7. Prove (*g*), Figure II, of § 2 (p. 62).

8. Use Table I (*a*) and find the slope of the secant to the curve $\phi(x)$ which goes through the points where $x = .99$ and $x = 1.01$. *Ans.*, $-\phi(1)$.

9. From the approximately normal distribution in § 3, find the mean of that part of the area under the curve $\phi(x)$ which lies to the right of $x = 3/2$. The answer should be $1/R_x$ of Table VII.

10. Find the equation which best fits the following data (*Gavett*) and plot the curve and histogram:

HEIGHTS OF FRESHMEN

Inches	*f*	*Inches*	*f*	*Inches*	*f*	*Inches*	*f*
60.0–61.0	3	64.5–65.5	51	69.0–70.0	148	73.5–74.5	5
61.5–62.5	8	66.0–67.0	115	70.5–71.5	64	75.0–76.0	3
63.0–64.0	33	67.5–68.5	156	72.0–73.0	43	Total	629

11. Do the same for the data of Chapter I, Problem 5 (*b*).

12. Graduate the data of Problem 10.

13. Graduate the data below:

SALARIES OF PROFESSORS IN PUBLICLY SUPPORTED INSTITUTIONS, 1919–1920 (*Burgess*)

Salary	*Number*	*Salary*	*Number*
\$ 250– 750	2	\$4250– 4750	127
750–1250	4	4750– 5250	94
1250–1750	50	5250– 5750	45
1750–2250	302	5750– 6250	10
2250–2750	628	6250– 6750	1
2750–3250	552	7750– 8250[1]	1
3250–3750	372	9750–10250[1]	1
3750–4250	271	Total	2460

14. Plot cumulative f for the normal law $\phi(x)$, using the data in § 1, and from this diagram find graphically the value of x cor-

[1] Note the intervals carefully. Fill the gaps in the table with zero frequencies before finding the moments.

responding to the first 15%. Compare the result with that found by coming out of Table I at the point where $\frac{1}{2}(1 + \alpha) = 0.1500$.

15. How far from the mean of a normal curve is the 7th decile? That is: find x if $\frac{1}{2}(1 + \alpha) = 0.7000$. How far is the 1st decile? the 9th?

16. (*a*) If $l = 9.053$, and $\sigma_l = 1.789$ for a normal curve, what is the value of t corresponding to the 3rd decile? (*b*) What is s?

17. The standard deviation of a certain set of 100,000 high school grades was 11%, and the mean grade was 78%. Assume the distribution to have been normal, and, being careful not to confuse percentage in the sense of grade with a percentage of frequency, answer the following questions: How many grades were (*a*) above 90%? (*b*) above 100%? (*c*) What was the highest grade of the lowest 1000? (*d*) What was the semi-interquartile range? (*e*) Within what limits did the middle 90,000 lie? *Ans.*, 59.9 to 96.1%.

18. Answer all the questions of Problem 17 with reference to a set of 100,000 grades in which the median was 83% and Q_3 was 90%. Also find σ.

19. Repeat the special proof of (*j*) in § 3, choosing the intervals half as broad. (Do not divide the given frequencies by 2. Derive them anew from Table I.)

CHAPTER VI

APPLICATIONS

1. Gunnery. We shall consider in this chapter certain standard uses of the normal law. One of the oldest and simplest of these is the application to artillery fire. Suppose shots are fired at a target which is in the same horizontal plane as the gun. Let the target be at the point T and the direction of fire as indicated in Figure I. Now, even if there is no real mistake on the part of the gunner, the shots will not all fall on T, but will be distributed about T at varying distances. This is due to the chance combinations of a number of small errors that cannot be avoided. They are due to many causes, such as slightly imperfect adjustments by the gunner, imperfect knowledge of the direction and velocity of the wind, variations in the amount of powder used and in its temperature, variations in the form and the condition of the surface of the projectiles, and changes in the form of the gun itself due to continuous firing. The theory of gunnery supposes that this dispersion of the shots about T obeys the normal law approximately. Consider first the dispersion in the direction of fire. This is called longitudinal dispersion. The distribution of shots in this direction is supposed to be given by a "ladder of dispersion" similar to Figure I.

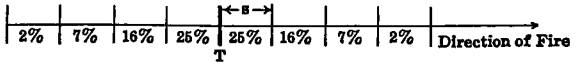


FIGURE I. Ladder of Dispersion

The distance between the rungs is the semi-interquartile range, s, which in gunnery is called the probable error of the gun in this direction and at the range in question. By

reference to Figure II of § 2, Chapter V, page 62, it will be seen that this "ladder" is an approximate mapping of the normal law. The idea may be put picturesquely, though crudely, as follows: Suppose the shells as they fell had their velocities suddenly reduced so that, instead of being buried in the ground, they simply piled up on the surface. Viewing the pile of shells from one side, one would see a mound in the shape of a normal curve. Similarly, the distribution of the shots in "latitude" or "azimuth," that is, at right angles to the line of fire, is given by a similar ladder of dispersion, but in this case the distance between the rungs is the probable error of the gun in azimuth, and this is usually less than the probable error in the direction of fire. The combination of these two ladders gives a "rectangle of dispersion" (Figure II).

	.50	.32	.14	.04	
	1.75	1.12	.49	.14	
	4.00	2.56	1.12	.32	
T	6.25	4.00	1.75	.50	Direction of Fire →
Totals:	12.50	8.00	3.50	1.00%	

FIGURE II. One Quarter of the Rectangle of Dispersion — beyond and to the left of *T*

This indicates the percentages of hits expected within each small rectangle about *T*. Thus the mound of shells would appear in the shape of a normal curve whether viewed from one side or from a point in the line of fire, but the second curve would be narrower than the first.

Example 1. The probable error of a gun in longitude is 20 yards. What proportion of the shots will fall (*a*) at least 40 yards short? (*b*) at least 10 yards short? (*c*) within 10 yards (longitudinally) of the target? Let us think of a ladder of dispersion of the right size laid down on the ground with its center at the target. The

size of the ladder is governed by the fact that the distance between the rungs in yards must be equal to the probable error, 20. Thus we have Figure III.

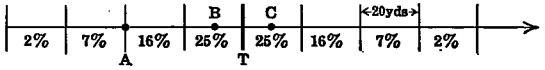
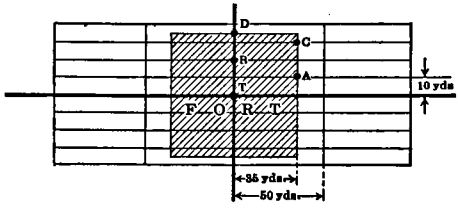


FIGURE III

On the diagram, the shots that fall to the left of A fall at least 40 yards short of T. The total number is 9%, and so this is the answer to (a). Those that fall to the left of B are at least 10 yards short. If we assume that 12.5% fall between B and T, we have as the proportion which fall to the left of B, 37.5%, the answer to (b). The student may object to this assumption, for by Table I he can find that the percentage of shots which fall between B and T is really 13.2% instead of 12.5%, but this is the assumption commonly made in gunnery. Moreover, it is justifiable, because, although it is true that the dispersion [1] is normal to the degree of approximation indicated by our ladder of dispersion, it is not true that it would be normal if the division were much finer than that. Making a similar assumption now to find the answer to (c), we note that we require the proportion of shots between B and C. This is 25%.

Example 2. The probable error of a gun is 50 yards in longitude and 10 yards in latitude. How many shots out of 150 will be expected to fall within a square fort 70 yards on each side?

We think of a rectangle of dispersion placed over the fort in the manner indicated in the figure. (Only the central portion is actually drawn.)



[1] At least for machine guns. Some authors do, in fact, use a more finely divided table, but it is of doubtful advantage.

The percentage expected in the rectangle TA is $35/50 \times 6.25$.
The percentage expected in the rectangle BA is $35/50 \times 4.00$.
The percentage expected in the rectangle BC is $35/50 \times 1.75$.
The percentage expected in the rectangle CD is $35/50 \times .50 \times \frac{1}{2}$.

The total percentage in the fort is four times the sum of these, and equals $4 \times 35/50 \times 12.25 = 34.3$. The number of shots out of 150 which may be expected to fall within the fort is therefore $.343 \times 150 = 51.45$.

The determination of the probable error of a gun for a given setting is found by actually firing a number of shots on the proving ground. The mean distance of these shots from some arbitrary origin in the direction of fire is found, and then the mean deviation, and thence the probable error, by multiplying the mean deviation by .845 (Chap. V (*e*), page 62). Then we find the mean distance to one side of the line to that origin, and in a similar manner obtain the probable error in latitude. The point which is at both these mean distances is called the center of impact. In Examples 1 and 2 it was supposed that the gun was so pointed that the center of impact was also the center of the target T. Of course, this might not be the case in practice. The problem of placing the center of impact on the target is evidently the major problem of ballistics, but it is not the one with which we are concerned here. We are concerned here only with the theory of the dispersion about that point. From this theory, simple approximate rules are derivable which may be used by the officer who is to have charge of the gun.

Example[1] 3. Find the probable error in longitude and latitude for a gun from the following six shots fired at a target (not the center of impact) at a range of 9000 meters.

[1] Taken from a War Department text on gunnery for heavy artillery issued during the World War. The small number of observations would mean a poor determination of the probable error, but heavy guns wear out so quickly that it is not desirable to use up many firings on the proving ground.

(1)	9275	meters,	right 20	meters
(2)	9410	"	right 10	"
(3)	9450	"	left 5	"
(4)	9370	"	right 15	"
(5)	9290	"	left 10	"
(6)	9360	"	right 5	"

By our usual methods, the mean of the numbers in the first column is 9359, the mean deviation is 51.16, and so the probable error in longitude is 43.23. Similarly, from the numbers in the second column, the mean is, right, 5.83, the mean deviation is 9.16, and s in latitude is 7.74. Incidentally, we have found the center of impact to be located at a distance of 9359 meters from the gun, and 5.83 meters to the right.

EXERCISES § 1

1. Consider latitudinal dispersion only in the following problems. Probable error is 8 meters. How many shots out of 200 would be expected to strike in the following places?

(*a*) More than 24 meters to the right of the line of fire. *Ans.*, 4.

(*b*) More than 20 meters to the right. *Ans.*, 11.

(*c*) More than 15 meters to the left. *Ans.*, 22.

(*d*) Within 7 meters of this line. *Ans.*, $87\frac{1}{2}$.

(*e*) Between 10 and 20 meters to the left. *Ans.*, 31.

(*f*) Between 30 and 35 meters to the left. *Ans.*, 1.

2. Find the center of impact and the two probable errors from the following observed shots: Right 3, over 12. Right 2, over 13. Right 1, short 10. Left 2, short 5. *Ans.*, (1, 2.5), 1.27, 8.45, if s is found from the mean deviation.

2. Physical Observations. There are good reasons for believing that, when many observations of the same quantity are made, their values are grouped about their mean value in an approximately normal fashion. This sort of observation is common in physics, engineering, and astronomy, and was illustrated in Example 6, Chapter I, page 9, where the length of a hall was supposed measured 50 times. The errors considered in § 1, under the head of gunnery, were also, strictly

speaking, of this type; but they were considered separately because the measurements there were less precise, and it is customary to use coarser approximations in dealing with them. The normal distribution of physical observations will generally occur, to a fair degree of approximation, but not unless all truly avoidable errors, called mistakes, and certain progressive errors due to progressive changes in the measuring machine have been eliminated. The little unavoidable errors which remain are called accidental errors and are said to be due to chance. The statistical discussion of physical observations containing only accidental errors is commonly called the "theory of errors." All that we have learned about normal frequency distributions applies to this discussion. The language and current practice of the physicist are, however, in some respects slightly different from those of the statistician and require special explanations.

As before, if t is one of N measurements, and $\bar{t}$ the mean, and σ the standard deviation of these measurements, the form of the distribution is approximately

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\bar{t})^2}{2\sigma^2}}. \tag{1}$$

The physicist calls $(t - \bar{t})$ the "deviation of a single observation," and σ the "root mean square deviation." Moreover, he is careful not to confound the terms "deviation" and "error." By deviation, he means a difference from a *mean* value, and by error, he means a difference from a *true* value. So, if V stands for the true value, $(t - V)$ is the "error of a single observation," while, since $\bar{t}$ stands for the mean value, $(t - \bar{t})$ is the "deviation of a single observation." The "root mean square error," to be designated by ϵ, is defined with respect to V in a manner analogous to "the root mean square deviation," σ, with respect to $\bar{t}$, thus:

$$\epsilon = \sqrt{\frac{\Sigma f(t - V)^2}{N}}, \text{ as } \sigma = \sqrt{\frac{\Sigma f(t - \bar{t})^2}{N}}. \tag{2}$$

At first, this distinction between error and deviation might seem futile, because in general we do not know what the true value V is, and how then can we know s, $t - V$, and ϵ, all of which depend on V? Nevertheless, it can be proved, if one be willing to grant certain reasonable hypotheses, that

$$\epsilon = \sqrt{\frac{N}{N-1}}\,\sigma, \tag{3}$$

and so the formula for ϵ could have been written

$$\epsilon = \sqrt{\frac{\Sigma f(t-\bar{t})^2}{N-1}}. \tag{4}$$

In this expression everything is known. Similarly, the probable deviation is $.6745\sigma$ as before (if the distribution is normal), and the probable error is defined as $.6745\epsilon$, and, therefore, in the usual language of the physicist:

Probable error of a single observation

$$= .6745\epsilon = .6745\sqrt{\frac{\Sigma f(t-\bar{t})^2}{N-1}}. \tag{5}$$

Since the physicist usually deals with ungrouped material, his observations being few in number, he commonly omits the f in all these formulae (see footnote, page 8). In the United States, the probable error is a more common measure of dispersion than the root mean square error ϵ, and so formula (5) is the important one in practice. From the theoretical point of view, it is of course quite immaterial whether one uses the probable error $.6745\epsilon$ or simply ϵ as a measure of dispersion, provided the definition just given is strictly adhered to. If, however, one attempts to define probable error in a manner analogous to probable deviation, *i.e.*, as a semi-interquartile range, then one must make use of the assumption that the distribution of errors is normal in order to know that this probable error will equal $.6745\epsilon$. Although, as with artillery fire, this is nearly the case under

certain ideal conditions, it is an ideal which often fails of realization; and so, to avoid confusion, it is better to stick to the quantity $.6745\epsilon$ as the definition of probable error.[1]

Example 4. Find the probable error of a single observation in Example 6, Chapter I, page 9. We have already found that

$$\bar{t} = 200.492,\ \Sigma f(t - \bar{t})^2 = .7568,\ N = 50.$$

So probable error $= .6745\sqrt{\dfrac{.7568}{49}} = 0.0838.$

Probable Error of the Mean. By this time the reader may have associated the idea of probability in some way with the term probable error, but it is better that this should not be done until after the true relationship can be appreciated, and this will come only after a study of the chapter on Probability. Moreover, the term is not a well-chosen one, and the relation to probability is not very important. For example, the probable error is not, as is sometimes supposed, the most probable error. In normal distributions we have learned rather that it is a sort of average error, the semi-interquartile range, or the median of all the errors, if their signs are disregarded. It might better have been called, therefore, the median error, and the student should insist on thinking of it as a median or average, rather than as something connected with the theory of probability. This is essential if he is to understand what is meant by the "probable error of the mean." More properly this should be called the median error of the mean, and it is in fact defined as the median of certain errors

[1] A few physicists do as we did with artillery fire. They define probable error with reference to a mean error, using the normal hypothesis. This gives

$$\textit{Probable error} = .845\sqrt{\frac{\Sigma f|t - \bar{t}|}{N(N-1)}}.$$

This is easier to compute than (5), but it is not very commonly used, and it will not give the same values for the probable error as (5) except when the hypothesis of normality is satisfied.

now to be described. Suppose the group of 50 observations of the length of the hall made in Example 6 of Chapter I were to be repeated, say, 100 times. We should then have 100 means, one for each group. They might not be all different, but there would be 100 numbers in all. Each of these means would be in error, the error being the difference between it and the true value. We would not know the values of these errors. (Some might in fact be zero.) Now the probable error of the mean is defined as the median of the absolute values of all these unknown errors. The formula for this probable error is as follows:

Probable error of mean [1]

$$= \frac{\text{prob. error of single observation}}{\sqrt{N}} = .6745\sqrt{\frac{\Sigma f(t-\bar{t})^2}{N(N-1)}}. \quad (6)$$

Example 5. Find the probable error of the mean in Example 4. *Ans.*, $0.0838/\sqrt{50} = 0.0119$.

It is customary to write the mean value in a case like this as follows: *mean length* equals $200.49 \pm .01$. The double signs do not indicate the *extreme* range of error of the result; they indicate the interquartile range; for, as stated above, the probable error of the mean does not signify the extreme error to which the mean is liable, only the median size of all the errors to which it is liable; it is a sort of average error. On the average, then, we may expect

[1] This formula may well seem mysterious to the student; indeed, if he stops to consider the matter, he will be inclined to be very skeptical about it. How is it possible, given only one real set of fifty observations, to learn anything at all about the other ninety-nine sets which are not real? Unfortunately, this simple and important formula in the theory of errors cannot be proved, except to the advanced student, and because it cannot be proved, it is not possible to explain the exact conditions under which it is true. It is perhaps sufficient at this point to remark that it is only true if certain assumptions regarding the way errors most commonly occur are made. A more detailed discussion is given in Part II. See also *Biometrika*, vol. 2, pp. 273–275.

The f is commonly omitted in this formula as well as in formulae (2), (4), and (5).

the mean of a group of observations like this one to be in error by as much as 0.01, numerically. How much in error the mean of this particular group is, we do not know.

Deviation and Error. In this section we have been distinguishing between deviation and error. The effect on the formulae has been to replace N by $N - 1$ in (5) and (6). Of how much importance is this distinction? The proof of (5) was not given, for, like (6), it depends on the advanced theory of sampling. Had the proofs been given, it would have been observed that the validity of the formulae depends on the assumption that N is fairly large.[1] But if N is fairly large, then the difference between N and $N - 1$ is relatively small. It would seem, therefore, that the exact scientists were a bit meticulous at this point: if N is large, the distinction does not matter, and if N is small, the formulae are not valid. But, since these formulae, as they stand, are the familiar ones of almost all books on engineering, physics, astronomy, etc., we also shall use them when dealing with problems in these fields. Elsewhere, we shall disregard the distinction made here and shall use the words *error* and *deviation* interchangeably; and then formula (6) will be replaced by (6*a*):

$$\text{Probable deviation of the mean}^2 = .6745\sqrt{\frac{\Sigma f(t-\bar{t})^2}{N^2}}. \quad (6a)$$

EXERCISES § 2

In each of the following sets of measurements find the standard deviation of a single observation, the probable error of a single observation, and the probable error of the mean.

[1] A proof commonly given in standard texts on the theory of errors is not rigorous. It *can* be rigorously proved that the "mean" value of ε^2 equals $N/(N-1)$ times the mean value of σ^2; but it is not true that the mean value of ε^2 usually equals $N/(N-1)$ times the given value of σ^2, and it would be necessary to prove this in order to justify the formula above. One may not replace the mean value of σ^2 by the given value without risk of serious error unless N is large.

[2] The f is commonly omitted.

1. 4, 3, 4, 6, 3. *Ans.*, 1.095, .826, .37.

2. 5 occurs 3 times, 4 occurs 2 times, 3 occurs 2 times, 2 occurs 3 times. *Ans.*, 1.2, .85, .27.

3. 670.2, 671.2, 671.3, 671.0, 671.1, 671.1. *Ans.*, .36, .27, .11.

4. 43′ 21″, 43′ 18″, 43′ 17″, 43′ 12″. *Ans.*, 3.2, 2.5, 1.3.

3. Psychological Measurements. In psychology and in certain related fields one has to do with measurements of mental, moral, and emotional characters. Here, and occasionally elsewhere, the data consist very largely of what are called (page 41) ordered rather than measured series. The measurements which we seem to have of these characters are not like physical measurements. They do little more, sometimes nothing more, than arrange the individuals examined in order: the boy who gets a score of 90 in honesty is not known to be as much more honest than the boy who gets 80 as that boy is more honest than he who gets 70; the scores 90, 85, and 70 would have meant as much. This fact has been noted before, and we were led to use the median and semi-interquartile range in such cases rather than the mean and σ. We may now go much further than this, provided we are willing to make just one more assumption. We hesitate to make it, for the supporting argument is feeble, but it is necessary if further progress is to be made. The argument runs thus: We have certain "yardsticks" by means of which one can measure biological quantities like lengths and weights. By their use we learn that some of these quantities, especially lengths, are distributed approximately in accordance with the normal law. Now, if we only did have a yardstick by which we might measure psychic characters, would these measurements also obey the same law? We do not know, but by analogy we assume that they would.[1]

[1] Unfortunately, even the analogy is rather poor, for biological weights are *not* normally distributed. Would it not be just as reasonable to choose as our psychic measure something which would give distributions analogous to weights as to lengths? However, the distribution of

To Normalize an Ordered Series. One can easily assign to the several groups of an ordered series their proper spacings so that the whole will fit a normal curve. This means, of course, that some groups will be squeezed into shorter intervals, others spread out over longer ones. This process is commonly called normalizing a series, and we shall obtain thus three different results, all of which are of value: (*a*) the relative positions of the end points of the several groups; (*b*) the relative positions of the median points of the several groups; and (*c*) the relative positions of the mean points of the several groups. We are now thinking of those cases where the total frequency is large enough to make several groupings feasible. The case where the total frequency is small will be taken up later.

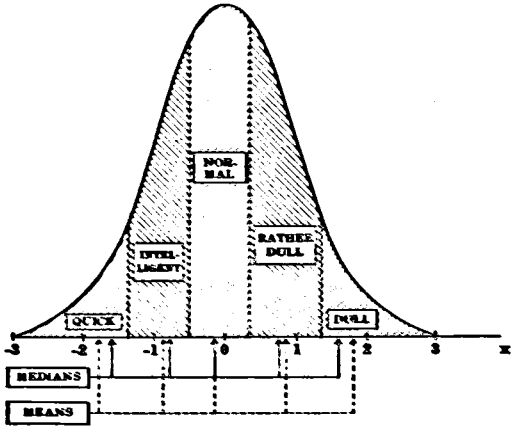
(*a*) *The end points.* We cannot obtain the *absolute* positions of any of the points (*a*), (*b*), and (*c*), only the *relative* positions. These relative positions are given by the "deviates" $x = \frac{t - \bar{t}}{\sigma}$ in our previous notation. The method of obtaining them is really only a formulation of what has already been suggested in certain problems of the last chapter. It is illustrated by the following example, in which fictitious data are used. The first two columns are supposed given. The last column is found by coming out of Table I at the places indicated by the values of cumulative f/N. These points mark the limits of the various groups in the figure.

---

weights is not far from normal. There is another argument for the normal law, based on the distribution of measurements made in terms of least discoverable differences. Cf. E. L. Thorndike, *The Measurement of Intelligence*, Appendix iii.

Example 6. INTELLIGENCE RATINGS OF 50 BOYS.

Rating	*f*	*Cum f to End Points*	*Cum f/N*	*End Points, z*
Quick.....	5	0	0	$-\infty$
Intelligent.....	10	5	.10	−1.282
Normal.....	15	15	.30	−0.524
Rather Dull.....	15	30	.60	0.253
Dull.....	5	45	.90	1.282
Total (N).....	50	50	1.00	∞



A Normalized Distribution

(*b*) *The median points.* These are found in a similar manner, except that one finds the cumulative frequencies to the median points instead of to the end points. These are determined by the condition that, in each group, the same amount of frequency lies on one side of the median as on the other.

Rating	*f*	*Cum f to Median Points*	*Cum f/N*	*Median x's*
Quick.....	5	2.5	.05	−1.645
Intelligent.....	10	10.0	.20	−0.842
Normal.....	15	22.5	.45	−0.125
Rather Dull.....	15	37.5	.75	0.674
Dull.....	5	47.5	.95	1.645

(*c*) *To find the mean points.* We may make use of the properties (*i*) and (*j*) stated in the last chapter. We use the end points already found. By (*i*), the mean of the last (*dull*) group will be [1]

$$\frac{\phi(1.282)}{\int_{1.282}^{\infty}\phi(x)dx} = \frac{0.1753}{0.10} = 1.753.$$

The denominator of the fraction may be found from Table I, but this is not necessary, for its value must be the relative frequency in the last interval, *viz.*, 5/50 = 0.10. By (*j*) the mean of the next preceding (*rather dull*) group is

$$\frac{\phi(.253) - \phi(1.282)}{\frac{1}{5}\frac{5}{0}} = \frac{0.2111}{0.30} = 0.704.$$

Next consider the groups at the beginning, where the x's are negative. On account of the symmetry of the normal curve, we may first find the means as if the x's were positive and then change their signs from plus to minus. Thus, for the first (*quick*) group,

$$\frac{\phi(1.282)}{\frac{5}{5}\frac{}{0}} = 1.753,$$

and so the mean is − 1.753. Or, we may apply (*j*) rigorously, paying attention to the signs, thus:

[1] The mean of this group is given more accurately by the use of Table VII, *especially where x is as large as 3.* By Table VII, $\log 1/R_x = -\log R_x = -\log R_{1.282} = 0.2444$. So by (*i*) the mean, $1/R_x = 1.755$, or perhaps 1.756.

$$\frac{\phi(-\infty) - \phi(-1.282)}{0.1} = \frac{0 - .1753}{0.1} = -1.753.$$

The middle (*normal*) group may cause trouble because x_1 is negative ($x_1 = -0.524$) and x_2 is positive ($x_2 = 0.253$), but again we need only to apply our formula rigorously, paying attention to signs:

$$Mean = \frac{\phi(-0.524) - \phi(0.253)}{\frac{15}{50}} = \frac{0.3477 - 0.3864}{0.3}$$

$$= -0.129.$$

The entire work may be organized in a table thus:

Rating	f	*Cum* f/N *to End Points*	*End* x*'s*	$\phi(x)$	Δ *Differences in* $\phi(x)$ *Column*	*Mean* x $\frac{-\Delta}{f/N}$
		0	$-\infty$	0		
Quick.....	5	.10	−1.282	.1753	.1753	−1.753
Intelligent....	10	.30	−0.524	.3477	.1724	−0.862
Normal.....	15	.60	0.253	.3864	.0387	−0.129
Rather Dull..	15	.90	1.282	.1753	−.2111	0.704
Dull.....	5	1.00	∞	0	−.1753	1.753
Total.....	50					

It is interesting to compare the means given here with the medians given under (*b*).

EXERCISES § 3

1. Normalize each series, p. 93, and find the x's of the end points.

2. Normalize each series so as to find the x's of the median points.

3. Normalize each series so as to find the x's of the mean points.

(a)

Data		Answers		
Order	f	(1)	(2)	(3)
A	3	−1.56	−1.9	−2.0
B	15	− .36	− .8	− .8
C	25	1.08	.3	.3
D	7		1.5	1.6

(b)

Data		Answers		
Order	f	(1)	(2)	(3)
D	2	−1.28	−1.6	−1.8
C	7	− .13	− .6	− .6
B	8	1.04	.4	.4
A	3		1.4	1.6

(c)

Data		Answers		
Order	f	(1)	(2)	(3)
A	2	−2.05	−2.3	−2.4
B	40	− .20	− .8	− .8
C	12	.10	−.05	−.05
D	10	.36	.2	.2
E	36		.9	1.0

4. Transfer to Arbitrary Scales. After the points just considered under the headings (*a*), (*b*), and (*c*) have been determined, relative to the mean of the normal curve and in terms of the σ of this curve as unit, we may, if we wish, determine them relative to any other origin and in terms of any other unit desired, by means of the relation,

$$x = \frac{t - \bar{t}}{\sigma_t}, \text{ i.e., } t = \sigma_t x + \bar{t}. \tag{1}$$

Here, t is the coördinate in the new system of the point designated by x in the system just used, $\bar{t}$ is the new coördinate of the mean of the normal curve, and σ_t the length in the new unit of the standard deviation. Equation (1) is valid only if t and x increase together, but if t increases when x increases, we should [1] replace (1) by equation (1*a*), *viz.*,

[1] But if we are willing to let σ become negative in the second case we may use (1) there also.

$$x = \frac{t - \bar{t}}{-\sigma_t}, \quad i.e., \quad t = -\sigma_t x + \bar{t}. \tag{1a}$$

The t system, and therefore the values of σ and $\bar{t}$ in that system, are quite arbitrary. In grading intellectual achievement, a traditional system is the percentage system, *i.e.*, a set of scores running from 0 to 100, called percentages. As before remarked, it is important that the reader shall not confuse percentage in the sense of a grade with percentile.

Example 7. Replace the ratings given initially in the preceding example by "percentage" grades so chosen that 90% shall separate the *Quick* from the *Intelligent*, and 50% shall separate the *Dull* from the *Rather Dull*.

First we must find σ and $\bar{t}$, and we have two equations or conditions with which we may do this. When $x = -1.282$, $t = 90$. Hence, using equation (1*a*):

$$(i) \qquad 90 = -\sigma(-1.282) + \bar{t}.$$

When $x = +1.282$, $t = 50$; hence,

$$(ii) \qquad 50 = -\sigma(1.282) + \bar{t}.$$

Solving these two equations (*i*) and (*ii*) simultaneously, we find σ and $\bar{t}$:

$$\sigma = +15.6, \quad \bar{t} = 70.$$

Therefore our general equation of transformation (1*a*) becomes

$$t = -15.6x + 70,$$

and this may be applied to any of the x points previously determined. Thus we may now find that the other end points, median points, and mean points are as follows:

End Points	
x	t, %
$-\infty$	∞
-1.282	90.0
-0.524	78.2
0.253	66.1
1.282	50.0
∞	$-\infty$

Median Points	
x	t, %
-1.645	95.66
-0.842	83.13
-0.125	71.95
0.674	59.49
1.645	44.34

Mean Points	
x	t, %
-1.753	97.35
-0.862	83.45
-0.129	72.01
0.704	59.02
1.753	42.65

Of course, since the normal scale extends over an infinite range, it is not possible to associate with it a proportional scale running from 0% to 100%. This scale also will run from $-\infty\%$ to $+\infty\%$; but it is usually possible to choose $\bar{t}$ and σ in such a fashion that all the individuals actually observed shall have grades between 0% and 100%. In this example it is clear that we have chosen σ and $\bar{t}$ subject to two quite arbitrary conditions (*i*) and (*ii*). This means that we have chosen σ and $\bar{t}$ arbitrarily. The reason for our choice might have been convenience merely, or custom. It may have been customary to think of those who had 90% or more as belonging to the best group, or the test may have been devised so that all the individuals of the best group would obtain these grades. The following example illustrates the case of another arbitrary choice.

Example 8. Replace the median ratings of Example 6 by "percentage" grades so chosen that the modal grade is 65% and the probable error is 10%.

Since in the normal system the mean and mode are the same, the first condition is that

(*a*) $$\bar{t} = 65.$$

The second is that

(*b*) $.6745\sigma = 10$; hence, $\sigma = 14.85$; and our equation of transformation is, by (1*a*):

$$t = -14.85x + 65.$$

This yields the following results for the median points:

Median Points	
x	t, %
−1.645	89.43
−0.842	77.50
−0.125	66.86
0.674	54.99
1.645	40.57

5. The Case Where N is Small. This case may be handled by the same methods as those used for the case where N was large, and if the median points are to be found, it is necessary,

as on page 55, to suppose each isolated frequency to be divisible into two halves, one half lying on one side of its median point and one half on the other. However, in this case it is possible to make out a simple table which will give the deviates immediately, and since the means have a slight advantage over the medians, the table (Table VI) has been constructed so as to give the mean points rather than the median points. By hypothesis, now, the only part of the given data that will be used is the rank list. This may always be supposed given by the numbers 1, 2, 3, etc., whether the given scores are equi-spaced numbers or not. The only possible exception arises where among the given series there are ties; but we may treat ties initially exactly as if they did not occur. For example, if three individuals are tied for second place, assign them initially, in any order, the ranks 2, 3, and 4, as if it were really possible to distinguish between them; but these ranks should be bracketed together so that we shall remember to give them special treatment later. Now for each rank, R, Table VI gives the corresponding normalized position, x, for all total frequencies N from 1 up to 50.

An alternative explanation of precisely what is accomplished by this table may be given as follows: Suppose $N = 10$; by means of the table the area under the normal curve is divided into 10 equal partial areas and the *mean* x of each part is indicated. This explanation tells us how to treat ties; for suppose the ranks 2, 3, and 4 are truly ties. Then we should like to lump together the partial areas corresponding to $R = 2$, 3, and 4, and find their general mean. Their general mean is the weighted mean of their several x's, but, since the several frequencies are all equal, the weights are also equal, and so the weighted mean is also the simple mean of their several x's. Hence, opposite ranks 2, 3, and 4 we should now put an x equal to the mean of the three x's which are given by the table. This will not be quite the same (usually) as the x opposite 3. Hence, it would not have been quite so

well to have averaged 2, 3, and 4 to begin with. After the x's have been found, a shift may be made to an arbitrary scale if desired, just as in the case where N was large.

Example 9. The following percentage grades were given in a test. Revise them so that the new percentage grades will be normally distributed, and so that the highest quartile will be 70% and the lowest 30%. (The data are also in Example 3, page 52, except that the sixth number has been altered.) $N = 22$.

Given t, %	*Rank R*	x_R *(Table VI)*		*New t, %*
36	1	−2.102		− 12.3
41	2	−1.497	−1.353	+ 10.0
41	3	−1.210		
47	4	−1.000		
47	5	−0.826	−0.834	+ 25.3
47	6	− .675		
53	7	− .538		34.0
57	8	− .410		37.8
59	9	− .289		41.4
61	10	− .172		44.9
62	11	− .057		48.3
64	12	+ .057		51.7
67	13	.172		55.1
69	14	.289		58.6
71	15	.410		62.2
74	16	.538		65.9
75	17	.675		70.0
76	18	.826		74.5
85	19	1.000		79.6
93	20	1.210		85.9
96	21	1.497		94.4
98	22	2.102		112.3

The new t is found from the equation of transformation (1), which becomes

$$t = 29.65x + 50.$$

This is obtained easily from the given conditions, thus: Half the difference between the highest and lowest quartile is $s = 20$, and $\sigma = \frac{s}{.6745} = 29.65$. Half the sum of the quartiles is the median, which is the same as the mean in the normal curve; so $\bar{t} = 50$.

EXERCISES §§ 4–5

1. Use the given answers to the exercises in the preceding set, and choose new t scores for the sets (*a*), (*b*), and (*c*) as follows:

(*a*)
- *End points:* make median $t = 100$, $s = 15$.
- *Medians:* make mean $t = 50$, $\sigma = 12$.
- *Means:* make mean point of D at $t = 20$, mean point of B at $t = 10$.

(*b*)
- *End points:* make median $t = 100$, make $t = 10$ divide A from B.
- *Medians:* make median $D = 100$, mode of whole $= 75$.
- *Means:* make 1st decile $= 90$, $Q_3 = 10$.

(*c*) *All points:* make a percentage scale such that 98% of the entire group shall fall in the interval 2% to 98%, inclusive. *Ans.*, *End points*, 7.7, 45.9, 52.1, 57.4; *medians*, 2.6, 33.5, 49.0, 54.1, 68.6; *means*, .5, 33.5, 49.0, 54.1, 70.6.

2. Find the mean x points in each case:

(*a*) *Ranks:* 1, 2, 3, 4, 5, 6.

(*b*) *Ranks:* 1, 2, 3 and 4 tied, 5, 6.

(*c*) *Scores:* 30, 35, 38, 41, 41, 41, 52, 76, 78, 78. *Ans.* to (*c*), $-1.76, -1.04, -.68, -.13, -.13, -.13, .39, .68, 1.4, 1.4$.

3. (*a*) Replace the scores in Exercise 2 (*c*) by a normal set such that 30 indicates the first individual and 78 the mean position of the last two.

(*b*) Same as (*a*), but make the scores run in the opposite direction, 30 indicating the last two, and 78 the first.

4. Use Table VI to solve Exercises 3 (*a*), (*b*) in the preceding set of exercises.

PROBLEMS CHAPTER VI

1. At a certain range the probable error of a gun is 100 yards in longitude and 30 yards in latitude. How many shots on the average will fall:

(*a*) 150 yards beyond *T*? (*b*) 75 yards short of *T*? (*c*) 100 yards to the right of *T*? (*d*) in a square whose sides are each 80 yards from *T* and are either parallel to or perpendicular to the line of fire?

2.[1] "You are ordered to make breaches in 2 wire entanglements, the entanglements being 10 *m.* in depth and 50 *m.* apart, center to center, lying perpendicular to the direction of fire. You are using the 10" gun, with reduced charge, and the range to the wire is 6000 *m.* The probable error at this range is 20 *m.*

"(*a*) Show . . . whether it will be more advantageous to keep the center of impact halfway between the entanglements or to adjust first on the center of one entanglement, and then on the center of the other.

"(*b*) If 45 hits on each entanglement will accomplish the desired destruction, how many rounds will be required in each case, *i.e.*, adjusting on mid-point or on each entanglement?"

3. A gun was fired 10 times in a fixed position on the proving grounds and the shots were noted with reference to a fixed mark as indicated below. Find the probable errors in longitude and in latitude. This gun is brought into action against a long rectangular network of fortifications (100 yards by 10 yards) at the average range used on the proving grounds. Choose the most advantageous position for the gun and find thus the maximum number of hits that could be expected out of every 100 shots.

Over 150 yds., Right 15 yds.	Over 90 yds., Left 30 yds.
" 120 " " 10 "	" 80 " " 15 "
" 100 " Left 20 "	" 60 " Right 20 "
" 90 " " 10 "	Short 20 " " 10 "
" 70 " " 5 "	" 10 " " 10 "

4. A battleship is 10 times as long as it is wide. The probable error of a gun is 5 times as great in longitude as in latitude. The

[1] Textbook on gunnery. *U. S. War Department, 1917.*

probable error in latitude equals half the width of the battleship. Compare the number of hits on this ship when it lies in the direction of fire and when it lies broadside to it. (Use a rectangular form in place of the battleship.)

5. Using the following set of micrometer measurements, find:

(*a*) the probable error of a single observation,
(*b*) the probable deviation of a single observation,
(*c*) the probable error of the mean,
(*d*) the probable deviation of the mean. *Ans.*, .0025.

56.199, 56.182, 56.180, 56.178, 56.193.

6. The same for Problem 1 (*d*), Chapter II, page 34. (Cf. also Problem 8, Chapter II.)

7. In what sense is it true that the average of 200 observations is 5 times as reliable a result as the average of 8? Would the reliability of an average of 1,000,000 observations be increased in the same fashion? Explain.

8. Show that, in any set of measurements, normal or not, the mean deviation from the mean is less than or equal to the standard deviation.

9. Replace (*a*) the end points, (*b*) the median points, (*c*) the mean points given in Example 6, § 3, by percentage grades on a normal scale so chosen that the median boy in the Quick group shall have 90%, and the median boy in the Dull group 50%.

10. Replace the scores given in Example 9, page 97, by percentage grades on a normal scale so chosen that the limits shall be 36% and 98%, as in the given data.

11. (*Thorndike*) "On the hypothesis that the distribution of darkness of eyes is normal . . . transmute into terms of units of amount (our x) the following relative positions."

Eye Color	*Light Blue*	*Blue Dark Blue*	*Gray Blue-green*	*Dark Gray Hazel*	*Light Brown Brown*	*Dark Brown*	*Very Dark Brown Black*
Per Cents of Englishmen	2.9	29.3	30.2	12.3	11.0	10.8	3.6

(*a*) That is: find the mean value of each group in σ units.[1]

(*b*) Find also the median points.

12. If two groups are normally distributed, the standard deviations being the same, and if the mean of the first exceeds the mean of the second by 2σ, what per cent of the first group exceeds the smallest 90% of the second? (It will be helpful to draw the figures.)

13. Answer the question in Problem 12 if the standard deviation of the first group is σ and of the second group $\sigma/2$.

[1] In Thorndike's book, *Mental and Social Measurements*, pp. 91–94, there is a short table which gives these results directly. In our notation this table is constructed thus:

$\frac{100\,cum\,f}{N}$ / $\frac{100\,f}{N}$	0	1	2
1			
2			
3			$-100x = 182$

Thus, if $\frac{cum\,f}{N} = .02$, and $\frac{f}{N} = .03$, the cumulation being to the first end point of the group in question, then the x of the mean of that group is -1.82, approximately. The *Kelley-Wood* Table is also specially constructed so as to be useful in problems of this sort.

CHAPTER VII

TIME SERIES: TREND AND RATIO CHARTS

1. Time Series. Before progressing further with frequency distributions, it is desirable to develop certain elementary notions of a different sort. What we are to do in this chapter is particularly important in the study of time series, and so, although the theory has other applications, we shall use the time series as the model. When a variable is tabulated as a function of the time, the set of values which results is called a time series. We have had already several illustrations of time series (*e.g.*, Example 2, § 2, Chapter I, page 4). In this chapter we shall consider only those cases in which the function is single valued; that is, at any given time there is not more than one corresponding value of the function. Later we shall, by an easy transition, make the application to multiple-valued functions. The weight of a child is an example of a single-valued function of the time. Child weight is an example of a multiple-valued function of the time, being different for different children. Average child weight, however, would be a single-valued function.

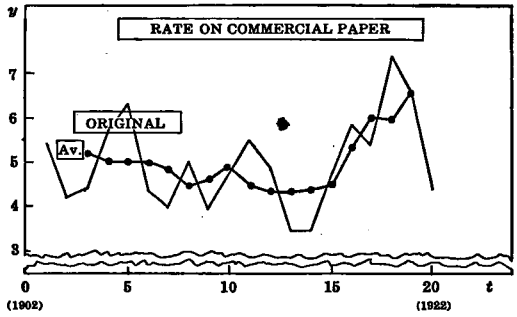
2. Moving Average. In studying a time series, one is often confused by its irregularity. One method of smoothing out such a series is by the use of the moving average. This is defined as the average of a group of a fixed number of successive terms of the series. More precisely, let the table be $(t_1, y_1), \ldots, (t_n, y_n)$, the t's denoting the times and the y's the corresponding values of the function. Let k be any positive integer (usually chosen between 3 and 20) less than n. Let

$$\bar{y}_1 = \frac{y_1 + \cdots + y_k}{k}, \quad \bar{y}_2 = \frac{y_2 + \cdots + y_{k+1}}{k}, \text{ etc.}$$

This group of $\bar{y}$'s is called the moving average of the y's. Similarly, the moving average of the t's is $\bar{t}_1 = \dfrac{t_1 + \ldots + t_k}{k}$, $\bar{t}_2 = \dfrac{t_2 + \ldots + t_{k+1}}{k}$, etc. Usually, the t's are equi-spaced. Then, of course, $\bar{t}_1 = \dfrac{t_1 + t_k}{2}$, $\bar{t}_2 = \dfrac{t_2 + t_{k+1}}{2}$, etc. In graphing the moving average, each $\bar{y}$ is regarded as the ordinate of a point and the corresponding $\bar{t}$ as the abscissa.

Example 1. Find the moving average of the rate on 60–90-day commercial paper in the years 1903 to 1922, taking $k = 5$. For convenience in writing let $t = 0$ at 1902. Then $t_1 = 1$, $t_2 = 2$, etc.

t	y	$\bar{y}k$	$\bar{y}$
1902+ 1	5.47		
2	4.20		
$\bar{t}_1$= 3	4.41	26.14	$5.228 = \bar{y}_1$
$\bar{t}_2$= 4	5.69	25.02	$5.004 = \bar{y}_2$
$\bar{t}_3$= 5	6.37	24.80	$4.960 = \bar{y}_3$
6	4.35	25.40	5.080
7	3.98	23.74	4.748
8	5.01	22.11	4.420
9	4.03	23.33	4.666
10	4.74	24.14	4.828
11	4.57	22.58	4.516
12	4.79	21.98	4.396
13	3.45	21.98	4.396
14	3.43	22.28	4.456
15	4.74	22.91	4.582
16	5.87	26.83	5.366
17	5.42	29.93	5.986
18	7.37	29.61	5.920
19	6.53		
20	4.42		



The third column contains the successive sums:

$$y_1 + y_2 + y_3 + y_4 + y_5 = 5.47 + 4.20 + 4.41 + 5.69 + 6.37 = 26.14;$$
$$y_2 + \cdots + y_6 = 4.20 + \cdots + 4.35 = 25.02.$$

The first two points are: $\bar{t}_1 = \dfrac{1+5}{2} = 3$, $\bar{y}_1 = \dfrac{26.14}{5} = 5.228$;

$$\text{and } \bar{t}_2 = \frac{2+6}{2} = 4,\ \bar{y}_2 = \frac{25.02}{5} = 5.004.$$

Note that the successive sums may be found easily with a computing machine, *e.g.*, to get 25.02 after having gotten 26.14, take out 5.47 and put in 4.35. As a check, compute the last sum independently also.

3. Trend. The straight line which best approximates the graph of a time series is called the trend line, or trend. If its equation is desired, it may be found either (*a*) graphically, or (*b*) numerically.

(*a*) *Graphically.* One simply draws the graph of the time series, then a straight line which seems to approximate it as closely as possible. The only problem is to find the equation of this straight line; and it is easily solved by the methods of analytics. Choose two points on the line, preferably far apart, and estimate their coördinates (t', y') and (t'', y'').

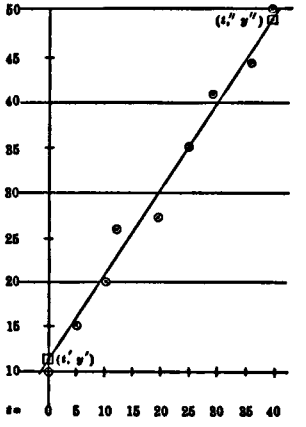
The equation of the line through these two points is then, by analytics,

$$y - y' = \frac{y'' - y'}{t'' - t'}(t - t'). \qquad (1)$$

The numerical work of simplifying this equation is slightly lessened if the first point is chosen so that $t' = 0$.

Example 2. Find the trend line graphically in the following series:

t	y
0	10
5	15
10	20
15	27
20	28
25	35
30	41
35	44
40	50



From the figure, $(t' = 0, y' = 10.8)$ and $(t'' = 40, y'' = 49.4)$.

Hence, by (1) $$y - 10.8 = \frac{49.4 - 10.8}{40} t = 0.965t.$$

So $$y = .965t + 10.8$$

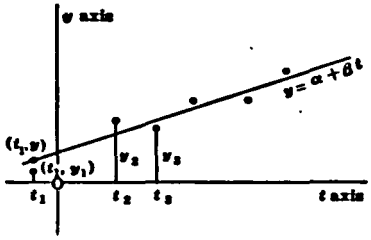
is the required equation.

This graphical method is open to the obvious objection that different workers will not agree on the position of the line, since it is a matter of visual judgment. The numerical method is almost as simple (in the usual case) and rather more satisfactory.

(*b*) *Numerically.* We use what is called the principle of moments. Let the equation to be determined be in the form

$$y = \alpha + \beta t. \qquad (2)$$

It is required to find the numerical values of α and β that will give the best-fitting line, but what is to be meant by "best fitting" requires explanation. Suppose for a moment that the equation of the desired line were known and graphed, as in the figure. Consider first the given point (t_1, y_1) of our table and also the corresponding point (t_1, y) of the line. By y here we mean the value obtained from the equation, *viz.*, $y = \alpha + \beta t_1$.



There are all together n given points, and their ordinates are $y_1, \ldots, y_n$; there will be also n corresponding values of y. The principle of moments says that we shall get a good fit if the zero*th* moment about the origin of t of these n y_i's equals the zero*th* moment of the n y's, and if also the first moment of the y_i's (about the origin of t) equals the first moment of the y's. These requirements yield two equations:

$$\left.\begin{array}{ll} \text{(zero}th\text{ moments)} & \Sigma y_i = \Sigma(\alpha + \beta t_i) \\ \text{(first moments)} & \Sigma t_i y_i = \Sigma t_i(\alpha + \beta t_i) \end{array}\right\} \qquad (3)$$

These equations can be solved for α and β. In Problem 4 the student is asked to show that the solution is

$$\left.\begin{array}{l} \alpha = \dfrac{\Sigma y_i \Sigma t_i^2 - \Sigma t_i \Sigma t_i y_i}{D} \\ \beta = \dfrac{n\Sigma t_i y_i - \Sigma t_i \Sigma y_i}{D} \\ D = n\Sigma t_i^2 - (\Sigma t_i)^2 \end{array}\right\} \qquad (4)$$

Substituting these values of α and β in (2), we get the required numerical equation.

If the t's are *equi-spaced*, as they usually are, this work can be much simplified: Let c be the common distance between successive t's. (Frequently $c = 1$.) Shift the origin to the mean of the t's and change the unit of measurement to c. That is, make the following substitutions:

$$\left.\begin{array}{ll} \text{Let} & \bar{t} = \frac{1}{n}\Sigma t_i = \frac{t_i + t_n}{2}. \\ \text{Substitute} & x = \frac{t - \bar{t}}{c},\ x_i = \frac{t_i - \bar{t}}{c}. \end{array}\right\} \tag{5}$$

These substitutions could be made in (4) and the resulting α and β determined, but it is easier now to think of the equation of the line in a new form,

$$y = A + Bx, \tag{6}$$

and to find A and B anew, as we found α and β before:

$$\left.\begin{array}{ll} \text{(zero}th\text{ moments)} & \Sigma y_i = \Sigma A + B\Sigma x_i \\ \text{(first moments)} & \Sigma x_i y_i = A\Sigma x_i + B\Sigma x_i^2 \end{array}\right\} \tag{7}$$

Since by (5) $\Sigma x_i = \frac{1}{c}(\Sigma t_i - n\bar{t}) = 0$, and since $\Sigma A = nA$, equations (7) can be solved separately as they stand, and

$$A = \frac{1}{n}\Sigma y_i,\quad B = \frac{\Sigma x_i y_i}{\Sigma x_i^2}. \tag{8}$$

Substituting these values of A and B in (6), we have a numerical equation which can be plotted as easily as if the variable t and the constants α and β had been used. However, if we really wish to find α and β we can do so, for the two equations,

$$y = A + Bx, \text{ and } y = \alpha + \beta t,$$

will relate to the same function of t if, in the first, one puts $x = \frac{t - \bar{t}}{c}$. So

$$A + B\left(\frac{t - \bar{t}}{c}\right) = \alpha + \beta t, \text{ identically, } i.e.,$$

$$A - \frac{B\bar{t}}{c} + \frac{Bt}{c} \equiv \alpha + \beta t, \text{ so that}$$

$$\alpha = A - \frac{B\bar{t}}{c}, \beta = \frac{B}{c}. \tag{9}$$

The origin and unit of t were, it will be remembered, quite arbitrary. The numerical work of finding A and B, indicated in (8), may be shortened a little more by using a relation which will be proved in the chapter on Finite Differences:

$$\Sigma x_i^2 = \frac{(n-1)(n)(n+1)}{12}. \tag{10}$$

Its truth will be illustrated in the problems [1] which follow. In the simple case which occurs most commonly, then, (6), (8), and (10) are the important equations. This simple case, where the t's are equi-spaced, will be illustrated first:

Example 3. *Find the trend numerically for the data of Example* 2. Here $c = 5$ and the mean t is 20, so that the x column is written down immediately, beginning with $x = 0$ opposite $t = 20$, and is exactly like the u column we had in finding moments in Chapter II.

t	y	x	xy
0	10	−4	− 40
5	15	−3	− 45
10	20	−2	− 40
15	27	−1	− 27
20	28	0	0
25	35	1	35
30	41	2	82
35	44	3	132
40	50	4	200
Sums	270		+297

[1] Cf. Problem 15 at the close of the chapter.

Since $n = 9$, by equation (10), $\Sigma x_i^2 = \frac{8 \cdot 9 \cdot 10}{12} = 60$. So, by (8), $A = \frac{270}{9} = 30$, $B = \frac{297}{60} = 4.95$; and, by (6), $y = 30 + 4.95x$ is our required equation. If we want it in terms of t: by (9),

$$\alpha = 30 - \frac{(4.95)(20)}{5} = 10.2,$$

$$\beta = \frac{4.95}{5} = 0.99,$$

so that $y = 10.2 + .99t$.

From this last equation we can observe the discrepancy between this numerical solution and the graphical one of Example 2.

Example 4. *Same as Example* 3, *omitting the last observation.* This example is introduced because, when n is even, the values of x are different from what they are when n is odd, but again a rigid application of the formulae brings a correct result.

t	y	x	$2xy$
0	10	−7/2	− 70
5	15	−5/2	− 75
10	20	−3/2	− 60
15	27	−1/2	− 27
20	28	1/2	28
25	35	3/2	105
30	41	5/2	205
35	44	7/2	308
Sums	220		414

$$\bar{t} = 17.5,$$

$$A = \frac{220}{8} = 27.5,$$

$$B = \frac{207}{42} = 4.929;$$

$$y = 27.5 + 4.929x.$$

$$y = 27.5 + 4.929\left(\frac{t - 17.5}{5}\right)$$

$$= 27.5 + .986t - 17.25;$$

$$y = 10.25 + .986t.$$

Note that it is convenient to use $2xy$ rather than xy in the last column of this table.

Example 5. *Same as Example* 3, *omitting the second observation.* This example illustrates the case where the t's are not equi-spaced. We have to use formula (4) and the work is much longer. Since the unit and origin of t are quite arbitrary, we could, if we chose, shorten the numerical work a little by taking the origin near the middle and the unit equal to 5 of the given units, but this is hardly worth while in a short set of data, and in some applications it is not feasible.

t	v	tv	t^2
0	10	0	0
10	20	200	100
15	27	405	225
20	28	560	400
25	35	875	625
30	41	1230	900
35	44	1540	1225
40	50	2000	1600
175	255	6810	5075

Hence,

$$D = (8)(5075) - (175)^2 = 9975,$$

$$\alpha = \frac{(255)(5075) - (175)(6810)}{D} = 10.26,$$

$$\beta = \frac{(8)(6810) - (175)(255)}{D} = 0.988,$$

and the equation desired is

$$y = 10.26 + 0.988t.$$

EXERCISES § 3

1. Plot Example 3 and the moving average for $k = 3$.

2. Same as Example 3, omitting the first observation.

3. Same as Example 3, omitting the first two observations.

4. Same as Example 3, omitting the observations at $t = 5$, $t = 15$, $t = 25$, and $t = 35$.

5. Same as Example 3, omitting the observations at $t = 5$ and $t = 10$.

4. Least Squares. The student may have been inclined to doubt whether the principle of moments which we used above really would give the *best* fitting straight line, even though admitting that it might give a line which would fit pretty well. Strictly speaking, there is no such thing as a best fit (except in very special cases). However, the principle of moments does, in this case, produce the same line as would be produced by the so-called principle of least squares. This principle says that the line is so chosen that the sum of the squares of the distances between the line and the given points (measured parallel to the y-axis) is a minimum. In our notation, then, $\sum_i(\alpha + \beta t_i - y_i)^2$ is a minimum if α and β are chosen in accordance with our formula.[1] The distances, $\alpha + \beta t_i - y_i$, are commonly called residuals. They are the differences between the values of y given initially and those obtained by the formula.

Example 6. Compute $\sum_i(\alpha + \beta t_i - y_i)^2$ in Examples 2 and 3, and show that it is less in Example 3 than in Example 2. The two

t	Example 2				Example 3			
	y	$\alpha + \beta t$	$\alpha + \beta t - y$	$(\alpha + \beta t - y)^2$	y	$\alpha + \beta t$	$\alpha + \beta t - y$	$(\alpha + \beta t - y)^2$
0	10	10.800	.800	.6400	10	10.20	.20	.0400
5	15	15.625	.625	.3906	15	15.15	.15	.0225
10	20	20.450	.450	.2025	20	20.10	.10	.0100
15	27	25.275	−1.725	2.9756	27	25.05	−1.95	3.8025
20	28	30.100	2.100	4.4100	28	30.00	2.00	4.0000
25	35	34.925	.075	.0056	35	34.95	− .05	.0025
30	41	39.750	1.250	1.5625	41	39.90	−1.10	1.2100
35	44	44.575	.575	.3306	44	44.85	.85	.7225
40	50	44.400	.600	.3600	50	44.80	.20	.0400
Sums				10.8774				9.8500

[1] The proof will be given in Chapter IX, where the more general case of multiple-valued functions will be investigated.

sums desired are 10.8774 and 9.8500, and, as expected, the second is less than the first.

5. Exponential Trends. If, instead of lying approximately on a straight line, the points of our diagram lie approximately on an exponential curve, we say that we have an exponential trend, and then it is desirable to find the equation of the best fitting exponential curve. An exponential function, generally speaking, is one in which x, or in more complicated cases a polynomial in x, appears as an exponent of a fixed number. The normal curve already considered,

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

is an example in which the fixed number is e. In this section we are to be specially interested in those cases where only the first power of x is used. Also, instead of x we shall use t as the independent variable. Again, our fixed number will be e, so that our trend equation will be in the form

$$y = ke^{mt}, \tag{11}$$

where k and m, like α and β in the straight-line trend, are constants whose numerical values we wish to determine.

It will not be difficult to do this, but let us first pause and see why exponential curves, rather than some other types, are considered important. In statistics of certain biological and economic phenomena, exponential trends are to be expected. If a living cell divides into two cells, and if each of these subdivides into two, and so on, the number of cells at any time is given exactly by an exponential equation like (11). If money is put at interest, and then the interest is also put at interest, and so on, the amount of money accumulated at any time is given exactly by the same sort of equation. It is not necessary that, as with the cells, each dollar shall produce two more. It may be that each twenty will produce one more. Though the rate of growth be slower, it may nevertheless be just as truly exponential. It is only necessary that the condi-

tions of growth that obtain in the first instance shall remain equally potent in the multiplied instances. If an industry can use accretions to its capital as productively as it can use its initial capital, its earnings will increase exponentially. Suppose a farmer had ten acres of land, and that from the first year's produce he could save enough to buy one acre more, and that this process could continue. Each year he would add 10% to what he had and his total holdings would grow at an exponential rate, exactly as if he had invested his money at compound interest, the rate being 10% per year. Indeed, it does often happen in practice that the trend of earnings in the more prosperous industries is exponential for a time. The same is true of trend of mortality from serious epidemics, in support of the theory that, in the early stages of an epidemic, each person who catches the disease infects a group of others, each person thus infected infects another group, and so on until a large portion of the community has been exposed. When the exponential character of these phenomena ceases, it is sometimes said that a saturation point has been reached. The letter m in our equation (11) of the exponential function may stand for a negative number, and then our curve will descend as t increases. If the value of a piece of property is depreciating, losing say 10% the first year, and again 10% of the depreciated value the second year, and so on, its value at any time is given by this sort of formula. Also the tail end of a frequency distribution is very apt to be described well by a decreasing exponential. We saw in Chapter III that often tables of frequency distributions were left open at one end. If in such a case it becomes necessary to distribute the last group given in a table over a set of measured intervals, a pretty good method is to assume that the distribution there is exponential.

6. The Constants k and m. To determine the constants we first take the logarithms of both sides of equation (11):

$$\log y = \log k + mt \log e. \qquad (12)$$

Now, let $Y = \log y$, $\alpha = \log k$, $\beta = m \log e$. (13)

Then equation (12) becomes

$$Y = \alpha + \beta t, \tag{12a}$$

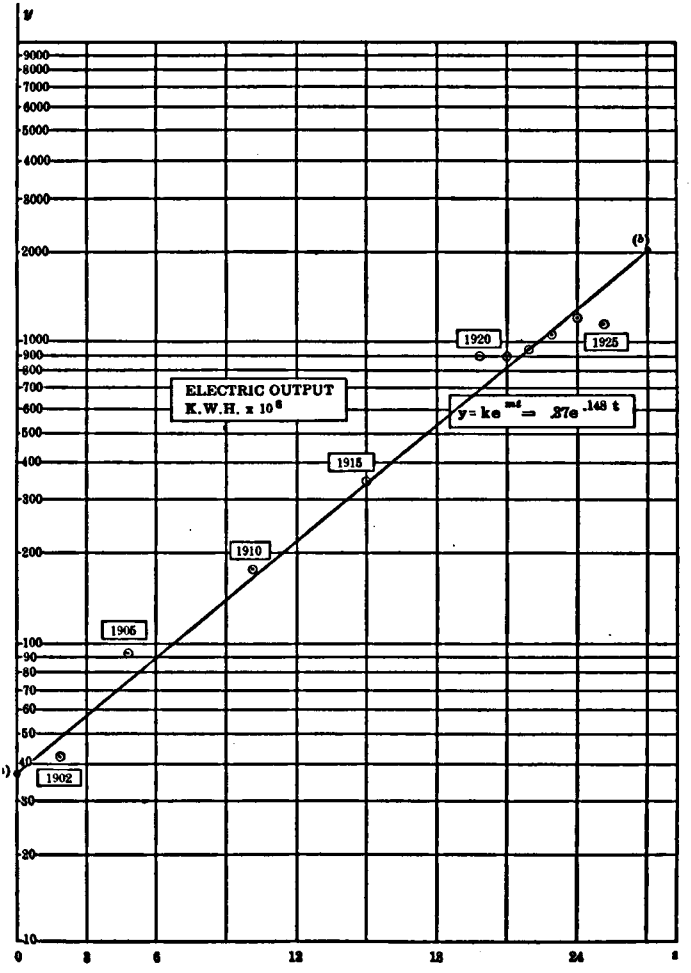
which is just like equation (2) of § 3, page 105. Therefore our problem of fitting (12) to the given (t, y) points is the same [1] as the problem of fitting (12a) to the corresponding (t, Y) points, and this is only the problem of the straight-line trend treated before. It may be solved either (*a*) graphically or (*b*) numerically.

(*a*) *Graphically. Ratio paper.* We have to plot the given (t, Y) points, that is, the given t_i and the logarithm of the given y_i, for each i. Then we are to draw by eye the best fitting straight line that we can. There is no difficulty about doing this, as before, except the slight one that we are required to look up the logarithms. It is a little better, therefore, to use paper so ruled that when one attempts to plot an ordinate y on it, one will really get an ordinate whose geometrical length is $\log y$. Such paper is commonly called ratio paper, or "arith log" paper, or "semi-logarithmic" paper, "semi" because there is no change in the abscissae, only in the ordinates. The rulings are spaced like those on a slide rule; in fact, a slide rule may well be used as a measuring stick in place of the specially ruled paper; and conversely two strips of the paper may be used as a slide rule.

Example 7. Find graphically the exponential trend of the electric output of the Montreal Light and Power Co. (t = number of years after 1900, y = millions of kilowatt-hours.)

t	2	5	10	15	20	21	22	23	24	25
y	43	93	176	350	909	907	945	1089	1200	1176

[1] Very nearly the same. The critical reader will realize that a least square solution of (12a) will not yield quite the same values for k and m as a least square solution of (11) would. The discrepancy can be practically eliminated by a proper choice of a system of weights to be applied to (12a).


y
9000
8000
7000
6000
5000
4000
3000
2000
1000
900
800
700
600
500
400
300
200
100
90
80
70
60
50
40
30
20
10
0
3
6
12
18
24
t
ELECTRIC OUTPUT
K.W.H. x 10⁶
y = ke^mt = .87e^.143 t
1902
1906
1910
1915
1920
1925

In the figure there are three sets of logarithmic rulings. These are to be numbered like those on a slide rule. If we decide that the lowest line reads 1, then the line at the bottom of the next set should read 10, and the next 100, and so on. Or we may begin with 10 and go to 100, then to 1000, etc. This last was the understanding necessary in plotting the points in this example; otherwise they would have run off the page either at the bottom or at the top. Next we choose our (t', y') and (t'', y'') points: (a) and (b) near the ends of our line. They are actually read as (0, 37) and (27, 2020). So far we have paid no attention to the fact that our distances really mean logarithms. We used the numbers at the left of the page exactly as if the lines were equi-spaced. But now to get (12a) we must get Y' from y' and Y'' from y'', that is, $Y' = \log y' = 1.5682$, and $Y'' = \log y'' = 3.3054$. Then we must substitute in the ordinary two-point form of the equation of a line:

$$Y - Y' = \frac{Y'' - Y'}{t'' - t'}(t - t'). \tag{14}$$

We obtain

$$Y = 1.5682 + \frac{3.3054 - 1.5682}{27}t,$$

or

$$Y = 1.5682 + .06434t,$$

as the equation of the straight line pictured. Thus $\alpha = 1.5682$, $\beta = .06434$, and then by (13) we can get also k and m:

$$k = \log^{-1}\alpha = 37, \quad m = \frac{\beta}{\log e} = \frac{.06434}{.4343} = 0.1481.$$

So our exponential trend has the equation

$$y = 37e^{0.1481t}.$$

This is the equation of the curve which on ordinary paper would take the place of the straight line on the ratio paper. The student must not confuse y and Y. Let him remember that the *numbers* at the side of the page are the ordinary numbers y, and that the *geometrical distances* that separate the lines on which these numbers are placed are logarithms. In plotting our points we need only to consider the ordinary numbers, but in order to match them to a straight line we have to consider the geometrical positions involved,

for it is actually the geometrical distances that determine whether we have a straight line or a curve. It will be helpful to repeat the work of this example, using ordinary paper and looking up the logarithms, as suggested at the beginning of the section.

(b) *Numerically.* We look up the value of Y in a table of logarithms. Then the t and Y play the same part here as the t and y played in the problem of the trend line.

Example 8. Find the exponential trend for Example 7 numerically.

t	y	Y	tY	t^2
2	43	1.6335	3.2670	4
5	93	1.9685	9.8425	25
10	176	2.2455	22.4550	100
15	350	2.5441	38.1615	225
20	909	2.9586	59.1720	400
21	907	2.9576	62.1096	441
22	945	2.9754	65.4588	484
23	1089	3.0370	69.8510	529
24	1200	3.0792	73.9008	576
25	1176	3.0704	76.7600	625
167	6888	26.4698	480.9782	3409

$$D = (10)(3409) - 27889 = 6201.$$

$$\alpha = \frac{(26.4698)(3409) - (167)(480.9782)}{D} = 1.5985.$$

$$\beta = \frac{(10)(480.9782) - (167)(26.4698)}{D} = 0.0628.$$

$$k = \log^{-1} 1.5985 = 39.67.$$

$$m = \frac{\beta}{\log e} = \frac{.0628}{.4343} = 0.1446.$$

Therefore the equation is:

$$y = 39.67e^{0.1446t}.$$

EXERCISES § 6

1. Same as Example 7, but use only the points at $t = 5, 10, 15, 20$.

2. Find graphically and numerically the exponential trends for each of the following cases:

(a)

t	0	2	4
y	2	10	100

Ans., $y = 1.78e^{0.978t}$.

(b)

t	1	2	3	3.5
y	3	7	20	33

Ans., $y = 1.09e^{0.967t}$.

(c)

t	0	1	2	4
y	1	.4	.1	.02

7. Properties of Ratio Charts. The usefulness of ratio paper and of the charts made on them does not depend primarily on the slight advantage noticed in § 6. Rather do that advantage and all other advantages depend on the following fundamental property.

(a) *If two numbers, y and y', are plotted as ordinates on ratio paper, these two ordinates having lengths in ordinary units equal to Y and Y', then the ratio y'/y will appear graphically as the difference $Y' - Y$.*

Example 9. Suppose $y = 2$ and $y' = 3$. The actual length of the ordinate Y will be (in centimeters) 0.3010 and the actual length of Y' will be (in centimeters) 0.4771; and the actual length in centimeters of the ordinate corresponding to the point 3/2 will be $-0.3010 + 0.4771 = 0.1761$.

This example makes the proof of (a) obvious, *viz.*, $Y = \log y$, $Y' = \log y'$; therefore $Y' - Y = \log y'/y$. Since on a ratio chart the number y'/y will appear graphically as its logarithm, this equation shows that it will appear also as the

difference, $Y' - Y$. An important corollary of this property is that: *if two time series appear as parallel*[1] *curves on ratio paper, always keeping the same distance apart (this distance being measured perpendicular to the time axis), then the ratio of the one function to the other is constant.* It is very helpful in forecasting economic phenomena when one can discover that two functions customarily bear a fixed ratio the one to the other. Since these functions usually are subject to oscillatory fluctuations it is not always easy to discover an innate proportionality of this sort, even when it exists, but a ratio plotting may make it immediately apparent. Another corollary is that: *if the curves are separating, the ratio of the greater function to the less is increasing, and if they are approaching each other, this ratio is decreasing.* It is easier to discover and to depict changes of this sort on ratio charts than on ordinary charts.

Example 10. MONOGAMY AND THE MOTOR CAR. (Erskine, *North American Review*, August, 1929.) Mr. Erskine says: "There was a greater proportion of married women in 1920 than in 1910 from women of every age . . . and the increase was most marked for the women of younger years." Then he presents the following data:

Age	*Women, Per Cent Married*	
	1910	1920
18	17.0	19.2
20	36.2	38.4
22	50.7	52.9
24	62.0	64.2
25	65.7	67.8

The ratio chart (Figure 1, page 120) does show clearly that the relative increase was most marked for the women of younger years,

[1] The two curves will coincide if one is shifted vertically the appropriate distance. They are not parallel in the sense that two concentric circles are parallel.

for the two curves are wider apart at that end. The ordinary chart (Figure 2) shows merely that the numerical increase was almost uniform throughout.

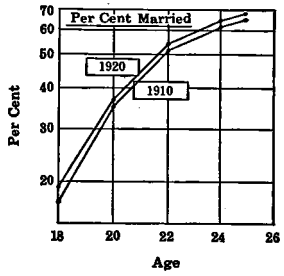


FIGURE 1

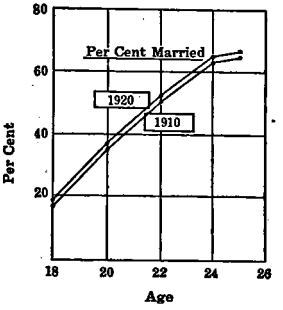


FIGURE 2

A second property of ratio charts has to do with "link relatives":

(*b*) *In an equi-spaced time series,* $(t_1, y_1), \cdots, (t_n, y_n)$, *the ratio of any ordinate to the preceding ordinate is called a link ratio or link relative. These relatives are* $\frac{y_2}{y_1}, \frac{y_3}{y_2}, \cdots, \frac{y_n}{y_{n-1}}$. *If the graph of such a series on ratio paper is a straight line, then the link relatives are all equal. Conversely, if the link relatives are equal, the graph is a straight line.*

Proof:

Since
$$\frac{y_2}{y_1} = \frac{y_3}{y_2} = \cdots = \frac{y_n}{y_{n-1}},$$
$$\log \frac{y_2}{y_1} = \log \frac{y_3}{y_2} = \cdots = \log \frac{y_n}{y_{n-1}},$$
and, therefore,
$$Y_2 - Y_1 = Y_3 - Y_2 = \cdots = Y_n - Y_{n-1}.$$

That is, the differences in the geometrical lengths of successive ordinates are all equal. This is a property of the straight line only. Since the steps are reversible, the converse is also true.

A corollary of this property is that: *if the link relatives*

always increase, the ratio graph is a curve which is concave up. Also, if the link relatives always decrease, the ratio graph is concave down. Let us prove the first part of this corollary.

Since $\frac{y_2}{y_1} < \frac{y_3}{y_2}$, $Y_2 - Y_1 < Y_3 - Y_2$.

But $\frac{Y_2 - Y_1}{t_2 - t_1}$ is the slope of the line from the first point to the second. Also, since $t_3 - t_2 = t_2 - t_1$, $\frac{Y_3 - Y_2}{t_2 - t_1}$ is the slope of the line from the second point to the third. It follows, then, that the first slope is less than the second. Likewise the second will be less than the third, and so on. The curve is such, therefore, that its slope is always increasing. It is therefore concave up, like the inside of a bowl.

EXERCISES § 7

1. Use ratio paper, and plot the ordinates $y = 1, 2, 3, 4, 5$, and $y' = 2, 4, 6, 8, 10$; each group with the abscissae $t = 1, 2, 3, 4, 5$.

(*a*) What property of ratio charts is illustrated by the relation between the broken lines whose ordinates are y and y'?

(*b*) Form the link relatives of y'. What property is illustrated?

(*c*) If y' is the gross business and t is the time, is the business increasing or decreasing? Is the per cent change from year to year increasing, or not?

(*d*) If y is the net business, is the ratio between net and gross increasing, or not?

(*e*) Plot also the ordinates $y'' = 1\frac{1}{2}y'$ by adding ordinates graphically, *i.e.*, without computing y'', except at the first point.

(*f*) What sort of graph would y' have on ordinary paper? Is the absolute change in the gross business from year to year increasing, or not?

2. Given any set of numbers, like (t, y) in Exercise 1, whose graphs lie on a straight line on an ordinary chart, will they always lie on a curve which is concave down on a ratio chart? If so, will it always be true that when y changes by equal absolute amounts, the per cent y will change by diminishing amounts? If not, what other possibilities are there?

3. (*a*) Plot the numbers on ratio paper,

t	1, 2, 3, 4, 5
y'	1, 10

inserting graphically the points not given, so as to make them all lie on a straight line.

(*b*) Answer the questions in Exercise 1 (*b*), (*c*), (*e*), and (*f*), using these values of y'.

4. By taking the logarithms of both sides of the equations, tell what sorts of curves on ratio paper will result from a graphing of the following:

(*a*) $y = e^{-t^2}$.
(*b*) $y = e^{a+bt+ct^2}$.
(*c*) $y^2 = e^{3t}$.
(*d*) $y = 2t$.
(*e*) $y = 2 + t$.

8. Parabolic Trend. When a time series cannot be approximated either by a straight line or by an exponential curve, it occasionally pays to try other types of curves. Of these we shall consider but one, the parabolic. The equation is of the form

$$y = \alpha + \beta t + \gamma t^2, \tag{15}$$

and the letters α, β, γ represent constants whose numerical values are to be found. We shall consider here only the simplest case, *viz.*, when the t's are equi-spaced. Then, as with the straight line (equation 5), it is better to make the substitution:

$$x = \frac{t - \bar{t}}{c}, \text{ where } \bar{t} = \frac{t_1 + t_n}{2}, \; c = t_2 - t_1, \tag{16}$$

and to write the equation in the new form,

$$y = A + Bx + Cx^2, \tag{17}$$

where A, B, C are constants whose numerical values are to be found. We shall use the principle of moments again and in this case [1] also it is equivalent to the principle of least

[1] But the two principles are not equivalent for most types of curves.

squares. Instead of equating the zero*th* and first moments, merely, we now equate the second moments also:

$$\left.\begin{array}{ll} (0th \text{ moments}) & \Sigma y = \Sigma A + B\Sigma x + C\Sigma x^2 \\ (1st \text{ moments}) & \Sigma xy = A\Sigma x + B\Sigma x^2 + C\Sigma x^3 \\ (2nd \text{ moments}) & \Sigma x^2y = A\Sigma x^2 + B\Sigma x^3 + C\Sigma x^4 \end{array}\right\} \tag{18}$$

These equations will simplify. Because of the choice of origin in (16), $\Sigma x = 0$ and also [1] $\Sigma x^3 = 0$. For compactness write $S_2 = \Sigma x^2, S_4 = \Sigma x^4$. In equation (10), page 108, we have already learned that

$$S_2 = \frac{(n-1)(n)(n+1)}{12} = \frac{n(n^2-1)}{12}, \tag{19}$$

and it will be shown, in the chapter on Finite Differences,[2] that

$$S_4 = \frac{(3n^2-7)S_2}{20}. \tag{20}$$

Our equations (18) become:

$$\left.\begin{array}{l} \Sigma y = An + CS_2 \\ \Sigma xy = BS_2 \\ \Sigma x^2y = AS_2 + CS_4 \end{array}\right\} \tag{21}$$

and the solution is immediate:

$$A = \frac{S_4\Sigma y - S_2\Sigma x^2y}{D}, B = \frac{1}{S_2}\Sigma xy, C = \frac{n\Sigma x^2y - S_2\Sigma y}{D}, \tag{22}$$

where $D = nS_4 - S_2^2 = \frac{nS_2(n^2-4)}{15}$. By the use of (19) and (20) the equations (22) can be written in the slightly more convenient forms: [3]

$$\left.\begin{array}{l} A = \dfrac{15}{n(n^2-4)}\left(\dfrac{3n^2-7}{20}\Sigma y - \Sigma x^2y\right) \\ B = \dfrac{12}{n(n^2-1)}\Sigma xy \\ C = \dfrac{15}{n(n^2-4)}\left(\dfrac{12}{n^2-1}\Sigma x^2y - \Sigma y\right) \end{array}\right\} \tag{22a}$$

[1] See Problem 15. [2] Part II, Chapter VII, Exercise 10, page 366.
[3] See Problem 15.

The substitutions of these values for A, B, and C in equation (17) will give us the result sought. Usually we do not need also the equation relative to the arbitrary (t) origin and unit, but if we do we can now get it:[1]

$$\left.\begin{aligned} \alpha &= A - B\frac{\bar{t}}{c} + C\frac{\bar{t}^2}{c^2} \\ \beta &= \frac{1}{c}(B - 2\frac{\bar{t}}{c}C) \\ \gamma &= \frac{C}{c^2} \end{aligned}\right\} \qquad (23)$$

Example 11. The following data give the approximate mean vital capacity of males at various ages (*Biometrika*, vol. 16, p. 141). Find the equation of parabolic trend.

Data		Computation		
Age t	*Vital Capacity* y.	$z = (t - \bar{t})/c$	$2xy$	$4x^2y$
19.5	227	$-\frac{7}{2}$	− 3859	65 603
22.5	230	$-\frac{5}{2}$	− 3450	51 750
25.5	230	$-\frac{3}{2}$	− 2990	38 870
28.5	237	$-\frac{1}{2}$	− 2607	28 677
31.5	227	$-\frac{9}{2}$	− 2043	18 387
34.5	229	$-\frac{7}{2}$	− 1603	11 221
37.5	222	$-\frac{5}{2}$	− 1110	5 550
40.5	218	$-\frac{3}{2}$	− 654	1 962
43.5	216	$-\frac{1}{2}$	− 216	216
46.5	210	$\frac{1}{2}$	210	210
49.5	205	$\frac{3}{2}$	615	1 845
52.5	193	$\frac{5}{2}$	965	4 825
55.5	201	$\frac{7}{2}$	1407	9 849
58.5	185	$\frac{9}{2}$	1665	14 985
61.5	200	$\frac{11}{2}$	2200	24 200
64.5	169	$\frac{13}{2}$	2197	28 561
67.5	160	$\frac{15}{2}$	2400	36 000
70.5	163	$\frac{17}{2}$	2771	47 107
Sums	3722		− 4102	389 818

[1] See Problem 15(d).

Obviously, $\bar{t} = 45.0$. Since $c = 3$, $x = \dfrac{t - 45.0}{3}$. The x's proceed by unit differences, so that only one need be computed. Instead of writing the denominator of x each time, it is better to compute the columns, $2xy$, $4x^2y$, not xy and x^2y. Hence,

$$\Sigma xy = -\frac{4102}{2} = -2051, \quad \Sigma x^2y = \frac{389{,}818}{4} = 97{,}454.5;$$

$$n = 18, \quad n^2 - 1 = 323, \quad n^2 - 4 = 320.$$

By (22a) then,

$$\left.\begin{aligned} A &= \frac{15}{18 \cdot 320}\left(\frac{965 \cdot 3722}{20} - 97454.3\right) = 213.89, \\ B &= \frac{12}{18 \cdot 323}(-2051) = -4.23, \\ C &= \left(\frac{15}{18 \cdot 320}\,\frac{12 \cdot 97{,}454.3}{323} - 3722\right) = -.264; \end{aligned}\right\}$$

so that equation (17) is:

$$y = 213.89 - 4.23x - .264x^2.$$

By (23), $\quad \alpha = 217.94, \ \beta = 6.51, \ \gamma = -.0293;$

so that equation (15) is:

$$y = 217.94 + 6.51t - .0293t^2.$$

The work could have been simplified a little if we had used, in place of y, the difference $(y - 160)$ throughout. If N is odd, instead of even, as here, the x's will be integers, as in Example 3, page 108. Then, of course, in the next two columns we shall have xy and x^2y, instead of $2xy$ and $4x^2y$.

EXERCISES § 8

1. Fit a parabola to the points of Example 2, considering only the points at $t = 20, 25, 30, 35$.

2. Same as Exercise 1, but add the point at $t = 40$.

3. Fit a parabola to each set and plot the data and the curve.

(a)

t	0	1	2	3
y	−3	0	10	20

Ans., $y = 4.56 + 7.90x + 1.75x^2$, $x = t - 1.5$.

(b)

t	-2	-1	0	1	2
y	-20	-5	2	-1	-15

Ans., $y = 1.91 + 1.40t - 4.86t^2$.

PROBLEMS CHAPTER VII

1. Compute and exhibit graphically the moving average of unfilled orders of the United States Steel Corporation for the years 1919–1922, taking $k = 4$.

Millions of Tons				
	1919	1920	1921	1922
January....	6.68	9.29	7.57	4.24
February...	6.01	9.50	6.93	4.14
March.....	5.43	9.89	6.28	4.49
April	4.80	10.36	5.85	5.10
May	4.28	10.94	5.48	5.25
June	4.89	10.98	5.12	5.64
July	5.58	11.12	4.83	5.78
August	6.11	10.81	4.53	5.95
September..	6.28	10.37	4.56	6.69
October.....	6.47	9.84	4.29	6.90
November..	7.13	9.02	4.25	6.84
December..	8.27	8.15	4.27	6.75

2. Do the same for Brokers' Loans for 1928–1929, taking $k = 3$.

Month	*J.*	*F.*	*M.*	*A.*	*M.*	*J.*	*J.*	*A.*	*S.*	*O.*	*N.*	*D.*	
1928	3.81	3.82	3.70	3.98	4.28	4.56	4.31	4.26	4.29	4.57	4.97	5.18	Billions of Dollars
1929	5.33	5.67	5.65	5.56	5.53	5.28	5.77	6.02	6.35	6.80	4.88	3.45	

3. Do the same for *The Annalist's* index of the prices of 50 rail and industrial stocks, taking $k = 6$.

Month	*J.*	*F.*	*M.*	*A.*	*M.*	*J.*	*J.*	*A.*	*S.*	*O.*	*N.*	*D.*
1928	180	177	192	195	196	189	190	203	205	209	228	231
1929	248	248	243	249	235	265	282	303	290	230	201	206

4. Prove equations (4).

5. (*a*) Find numerically the straight-line trend in Example 11. (*b*) Show that the sum of the squares of the residuals is smaller with the parabolic trend (Example 11) than with the straight-line trend.

6. Find (*a*) graphically and (*b*) numerically the straight-line trend of the death rate from cancer in the United States for the years 1900–1924.

Year	1900	1910	1920	1924
Rate	9.4	12.1	14.	14.9

7. Find graphically the equation of the trend line of Brokers' Loans, using the data of Problem 2.

8. Do Problem 7 numerically.

9. Find graphically the exponential trend from the data:

t	0	1	2	2.5	3
y	1	2.7	7.4	12.2	20.1

10. Do Problem 9 numerically.

11. Find graphically the exponential trend of gross earnings of all Bell telephone companies in the United States and estimate the earnings for 1930, on the hypothesis (usually very uncertain) that the apparent law of growth will not change.

Year	1921	1922	1923	1924	1925	1926	1927	1928
Earnings (Millions of Dollars)	521	564	623	678	761	845	917	1003

12. Find analytically the exponential trend of net earnings of the General Electric Company and estimate the earnings for 1929. Use the same hypothesis as in Problem 11.

Year	1924	1925	1926	1927	1928
Millions of Dollars	45	43	49.6	51.5	57.3

13. United States Steel Corporation. Plot on the same ratio chart (*a*), (*b*), and (*c*), where (*a*) is the amount of unfilled orders [1] on June 30 of each year, (*b*) equals the gross earnings reported at the end of each year, and $c = a/b$. Draw a smooth curve through the points of (*c*) and project it into 1929. Hence estimate (*b*) for 1929 (cf. Chapter I, Problem 1 (*a*)). What does the lack of parallelism of curves (*a*) and (*b*) indicate?

Year	1921	1922	1923	1924	1925	1926	1927	1928	1929
Millions of Tons (*a*)	29.92	21.09	25.58	15.96	17.11	16.48	13.71	15.10	16.35
Millions of Dollars (*b*)	986.75	1092.7	1571.4	1263.7	1406.5	1508.1	1310.4	1374.4	

14. The figures for 1920 in Example 10 are such that their logarithms lie nearly on a parabola. Using only the first four, equi-spaced, numbers, find the equation of this parabola. Now what is the equation of the exponential curve which nearly fits the given numbers?

15. (*a*) In Example 11, show by actual computation that

$$S_2 = \frac{n(n^2-1)}{12}, \quad S_4 = \frac{(3n^2-7)S_2}{20}. \quad \text{(19) and (20)}$$

(*b*) Prove the statement of (16) that $\Sigma x^3 = 0$. (*c*) Derive (22*a*) from (19), (20), and (22). (*d*) Derive (23), noting the proof of (9). (*e*) Same as (*a*), using Example 2.

16. Fit a parabolic curve to the figures for the net sales of the General Motors Company:

Year	1924	1925	1926	1927	1928	1929
Millions of Dollars	568	735	1058	1270	1460	1504

17. Same as Problem 16, omitting the year 1924.

[1] More precisely, (*a*) is the sum of the four quarterly reports of unfilled orders up to June 30 of each year, and is therefore proportional to the average of these reports.

CHAPTER VIII

CORRELATION, THE SURFACE AND THE COEFFICIENT

1. The Frequency Surface. We now return to the subject of frequency distributions already begun in Chapters I–VI. We are about to consider the interrelation of three variables, X, Y, and Z, instead of two as heretofore. Usually we have chosen t and y as the two variables, but we might have used instead X and Y. We said that y was a function of t, if for every value of t (of a given set) there was a corresponding value of y; y was then a function of a single variable t, and could be graphed as a curve. We now say that Z is a function of *two* variables, X and Y, if to every *pair* of values of X and Y (in a given set) there is a corresponding value of Z; this function may be graphed as a surface in three-dimensional space. An example of a function of one variable was

$$y = A + Bt.$$

This was graphically a straight line. An example of a function of two variables is

$$Z = A + BX + CY \text{ (Figure 1).}$$

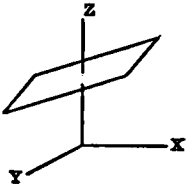


FIGURE 1

This is graphically a plane. Another example is

$$Z = \sqrt{C^2 - X^2 - Y^2} \text{ (Figure 2).}$$

This is half of a spherical surface. In general, any algebraic or other mathematical expression involving X and Y is an example of a function of X and Y, and may be graphed as a surface.

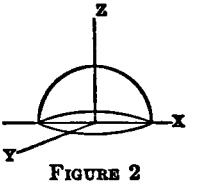


FIGURE 2

Example 1. We may, of course, have *frequency* functions of two variables. A good example is the rectangle of dispersion which we

saw was used in gunnery. A rectangle was supposed drawn over the field where the shots were to fall. Within each little cell of this rectangle was a number. This told us the number of shots to be expected in that cell. Now the position of the cell was located by two variables, X, the longitudinal distance from the center of impact, and Y, the latitudinal distance. The relative frequency was the number in the cell, and was the third variable Z. The value of Z differed for different cells, its value depending on the position of the cell, that is, upon the values assigned to X and Y. Z was then a frequency function of these two variables. We might now think of solid rectangular columns (parallelepipeds) erected on these cells as bases, so that the volume of each would be proportional to the frequency in the cell. Since the cells all had equal areas, these volumes would in turn be proportional to the altitudes, and so the altitudes of these columns would also represent the frequencies. Such a figure is a solid histogram. A smooth surface drawn through the tops of the columns would be an example of a frequency *surface;* in this case also a *normal* surface. This was pictured crudely in our earlier discussion as the surface of a mound of shells such as would result if the various shells piled up as they fell (Figure 3). The rectangle of dispersion is, then, an example of what we shall call a two-dimensional frequency *table.* Any frequency table of two (or more) dimensions is also commonly called a *correlation* table; another name for a frequency surface is a correlation surface.

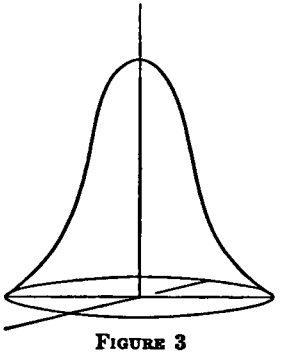


FIGURE 3

The rectangle of dispersion was a symmetrical table in certain respects, but usually correlation tables are not symmetrical. Consider now another example (*Example* 2), observations on the lengths and breadths, in centimeters, of 900 books chosen in a nearly random fashion from a large library.

900 Books

Length X / Breadth Y	12	16	20	24	28	32	$f(Y)$
10	5	61	24				90
14		21	368	202			591
18			8	164	15		187
22				1	21	7	29
26					1	2	3
$f(X)$	5	82	400	367	37	9	900 (N)

The correlation table proper lies within the double lines. Here are placed the various frequencies with which the several observations indicated in the top and left-hand margins occurred. Thus, 5 books were found for which the length was about 12 and the breadth about 10 centimeters, 61 for which the length was 16 and the breadth 10. The frequency may be denoted by Z, or, more commonly, by $f(X, Y)$. If we add the frequencies in the several vertical columns of this table, we obtain the numbers in the margin at the bottom. These are called marginal totals, and are represented by $f(X)$. The letter f does not stand for "function," but for "frequency"; $f(X, Y)$ is the frequency in the cell at (X, Y); $f(X)$ is the frequency in the column whose coördinate is X. Similarly, $f(Y)$ is the frequency [1] in the horizontal row whose coördinate is the given Y. One could therefore write the equations:

$$f(X) = \sum_{Y} f(X, Y); \quad f(Y) = \sum_{X} f(X, Y). \qquad (1)$$

[1] It is always true in our notation that f, though it stands for frequency, is actually a function of the variable or variables in the parenthesis following it; but we cannot agree that f shall also stand for *the* f function, because it would then be necessary that the f function of X should be the same expression in X as the f function of Y is in Y, and this is not usually true in a correlation table.

The first of these equations says: consider any cell (X, Y) and its frequency $f(X, Y)$; now, holding X fast, add all such frequencies for all possible values of Y. This means that we add a column of frequencies. The second equation says that we must hold Y fast and add the frequencies for all possible values of X; that is, we add along a row.

Example 3. If, in Example 2, $X = 20$, $f(X) = \sum_Y f(20, Y) = f(20, 10) + f(20, 14) + f(20, 18) = 24 + 368 + 8 = 400$.

2. The Mean. We shall now show, as in the case of one-dimensional frequency distributions, how to find certain important numbers which together describe the character of the two-dimensional distribution. Among the most elementary of these are N, the total frequency, and $(\bar{X}, \bar{Y})$, the position of the mean point.

We get N by adding together all the frequencies of the table, in any order. We may denote such a summation thus:

$$\sum_{X, Y} f(X, Y) = N. \tag{2}$$

The notation $\sum_{X, Y}$ means that we are to find the sum of all expressions of the sort that follow Σ for all values of X and Y, irrespective of the order. If we had desired to add the frequencies, first in the X direction and then in the Y direction, we should have written

$$\sum_Y \sum_X f(X, Y) = N, \tag{3}$$

obtaining, of course, the same result. Since in equation (1) we saw that $f(Y) = \sum_X f(X, Y)$,

equation (3) could have been written:

$$\sum_Y f(Y) = N. \tag{4}$$

It therefore says that the sum of the totals given in the right-hand margin is N. Similarly, for the totals at the bottom:

$$\sum_X f(X) = N. \tag{4a}$$

It should be noticed now that each of these sets of marginal totals, $f(X)$ and $f(Y)$, is an ordinary one-way frequency distribution, whose total is N. Let $\bar{X}$ denote the mean of $f(X)$, and $\bar{Y}$ the mean of $f(Y)$. The point $(\bar{X}, \bar{Y})$ shall then be called the general mean point of the table. By Chapter I,

$$\bar{X} = \frac{1}{N}\sum_X Xf(X), \quad \bar{Y} = \frac{1}{N}\sum_Y Yf(Y). \tag{5}$$

If we had wished, each of the equations (5) might have been written as a double sum; for, insert into them the values of $f(X)$ and $f(Y)$ given by (1), then,

$$\bar{X} = \frac{1}{N}\sum_X\sum_Y Xf(X, Y), \quad \bar{Y} = \frac{1}{N}\sum_Y\sum_X Yf(X, Y); \tag{6}$$

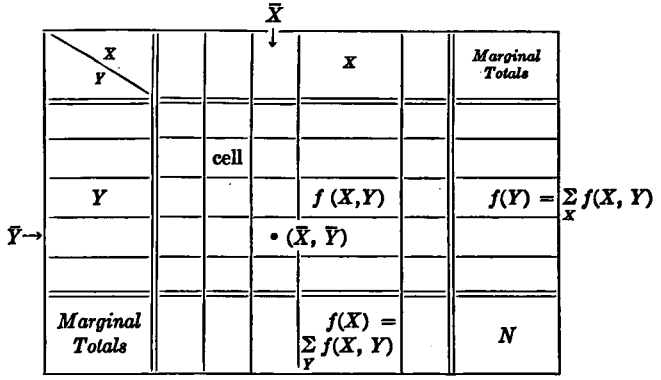
or, again, since the order of summation is immaterial,

$$\bar{X} = \frac{1}{N}\sum_{X,Y} Xf(X, Y), \quad \bar{Y} = \frac{1}{N}\sum_{X,Y} Yf(X, Y). \tag{7}$$

What (7) says in words is this: Multiply each of the tabulated frequencies by its own X, add the products for the whole table and divide by N; the result is $\bar{X}$. Multiply each frequency by its own Y, add over the whole table and divide by N; the result is $\bar{Y}$.

A physical model of a correlation table may be pictured thus. Think of a thin flat piece of metal marked off into small rectangular cells, and suppose the density to vary from cell to cell. The amount of metal in any given cell corresponds with the frequency in a cell of a correlation table. The center of gravity of the whole is the general mean point. Physically, the center of gravity of a flat piece of metal is the place where a vertical pivot should be placed in order that the metal should balance on it, the plane of the metal being horizontal.

The diagram below summarizes the notation of this section.



Notice that when we sum in the X direction we have as a result a function of Y only, and that when we sum in the Y direction we have left a function of X only. Any row or column of a correlation table is also called an array. The rectangles containing the frequencies we have called cells. Some authors use this word for the rectangles containing the marginal totals.

Example 4. Find $\bar{X}$ and $\bar{Y}$ for Example 3. The method was explained in Chapter II. A suitable form for the computation is

v \ u		-2	-1	0	1	2	3	f(v) = f(Y)	vf(v)
	X / Y	12	16	20	24	28	32		
-2	10							90	-180
-1	14		DATA OF					591	-591
0	18		EXAMPLE 2					187	0
1	22							29	29
2	26							3	6
f(u) = f(X)		5	32	400	367	87	9	(900)	(736)
uf(u)		-10	-32	0	367	74	27	(376)	

given above. $N = 900$, $\Sigma vf(v) = -736$, $\Sigma uf(u) = 376$. These numbers are enclosed in circles. The origin of u and the origin of

v are chosen arbitrarily, but preferably near the center of the table. The unit of u is $h = 4$cm., and the unit of v is $k = 4$cm. in this example; they are not always the same; both are class intervals.

$$\bar{u} = \frac{1}{N}\Sigma_u uf(u) = \frac{376}{900} = 0.4178. \quad \text{Hence, } \bar{X} = 20 + h\bar{u} = 21.67.$$

$$\bar{v} = \frac{1}{N}\Sigma_v vf(v) = \frac{-736}{900} = -0.8178. \quad \text{Hence, } \bar{Y} = 18 + k\bar{v} = 14.73.$$

EXERCISES § 2

1. From the following data make a "scatter diagram"; *i.e.*, plot each of the points, using as coördinates: (*length*, *breadth*).

30 Books (Centimeters)

l	23.5	23.8	24.3	24.5	24.3	23.5	24.2	24.2	24.3	24.0	23.4	23.3	23.7	24.5	24.0
b	16.3	16.2	17.3	17.8	17.6	16.3	17.2	17.0	16.4	16.3	17.1	16.6	17.2	17.5	17.1
l	24.5	23.0	24.0	24.7	24.9	22.3	23.4	23.6	22.6	23.7	23.7	22.0	22.9	23.4	23.0
b	17.5	16.2	16.3	16.2	16.0	16.4	16.5	16.1	16.0	17.8	14.9	15.2	14.9	15.1	15.2

2. Form a correlation table, using Exercise 1, made up of 3×4 cells as follows:

X (length): $(22-)$, $(23-)$, $(24.0-24.9)$;

Y (breadth): $(14-)$, $(15-)$, $(16-)(17.0-17.9)$.

3. In Exercise 2, compute each of the following sums, taking $u = 0$ at $X = 23.45$, and $v = 0$ at $Y = 15.45$:

(*a*) $f(u)$, $f(v)$, $\Sigma_v f(v)$, $\Sigma_u f(u) = 30$.

(*b*) $\Sigma_u uf(u)$, $\Sigma_v vf(v)$, $\Sigma_u u^2 f(u) = 17$, $\Sigma_u [f(u)]^2 = 354$.

(*c*) $\Sigma_u f(u, v)$ when $v = 0$, $\Sigma_v f(0, v)$, $\Sigma_u f(u, 1) = 14$, $\Sigma_u uf(u, 1) = 3$.

(*d*) $\Sigma_v \Sigma_u uf(u, v)$, $\Sigma_v \Sigma_u vf(u, v)$, $\Sigma_u \Sigma_v uf(u, v) = 20$.

3. Moments. As with frequency distributions of one variable, the important constants which describe the character of the distribution are moments. But there are rather more than twice as many moments to be considered now as in

the simpler case, for there are moments in the X direction, moments in the Y direction, and also composite moments, obtained by using both directions.

DEFINITIONS. *Moments about an arbitrary or given origin.*

(*a*) *The* r*th moment in the X direction about the origin of X is* $\frac{1}{N}\sum_{X,Y} X^r f(X, Y)$.

(*b*) *The* r*th moment in the Y direction about the origin of Y is* $\frac{1}{N}\sum_{X,Y} Y^r f(X, Y)$.

The first of these definitions may be put into words thus: *Multiply each of the tabulated frequencies by the* r*th power of its own X, sum over the whole table, and divide by N; the result is the* r*th moment in the X direction.*

COROLLARIES:

(*a*) *The* r*th moment in the X direction is the same as the* r*th moment of the marginal totals* $f(X)$.

(*b*) *The* r*th moment in the Y direction is the same as the* r*th moment of the marginal totals* $f(Y)$.

The proof of these corollaries is immediate if one understands the notation as explained in the preceding sections. Thus, since the order of summing is immaterial,

$$\frac{1}{N}\sum_{X,Y} X^r f(X, Y) = \frac{1}{N}\sum_X\sum_Y X^r f(X, Y) = \frac{1}{N}\sum_X X^r \sum_Y f(X, Y)^1$$
$$= \frac{1}{N}\sum_X X^r f(X),$$

which is the definition (Chapter II) of the rth moment of $f(X)$.

Example 5. The zero*th* moment in the X direction is 1, the first moment is $\bar{X}$, and the second moment is the same as ν_2 of $f(X)$.

[1] By Theorem I of Chapter I, for X is constant while the summation is made with respect to Y.

DEFINITION (*c*). *The first product moment about the common origin of X and Y is*

$$\frac{1}{N}\sum_{X,Y} XYf(X, Y).$$

Other product moments are used in the more advanced theory, but are not needed in this course.

Moments about the Mean Point. Let $x = X - \bar{X}$, $y = Y - \bar{Y}$. That is, let x and y be the coördinates of any point referred to the center of gravity or general mean point as origin. The moments about $(\bar{X}, \bar{Y})$ are defined and denoted as follows:

(α) *The* rth *moment in the x direction is* $\frac{1}{N}\sum_{x,y} x^r f(x, y) = \mu_{x^r}$.

(β) *The* rth *moment in the y direction is* $\frac{1}{N}\sum_{x,y} y^r f(x, y) = \mu_{y^r}$.

(γ) *The first product moment is* $\frac{1}{N}\sum_{x,y} xyf(x, y) = p_{xy}$.

Example 6. $\mu_{x^0} = 1$, $\mu_{y^0} = 1$, $\mu_x = 0$, $\mu_y = 0$, $\mu_{x^2} = \sigma_x^2$, $\mu_{y^2} = \sigma_y^2$, where σ_x and σ_y refer to $f(x)$ and $f(y)$ respectively.

Coefficient of Correlation. DEFINITION (*d*). *The coefficient of correlation is denoted by* r, *and is defined as the first product moment about the general mean point in terms of the* σ'*s as units:*

(δ) $$r = \frac{p_{xy}}{\sigma_x \sigma_y}.$$

4. Computation of Moments (N Large). In this section we shall continue to suppose that N is large enough to allow the material to be grouped into equal cells. The case where N is too small for this will be discussed presently.

The simple (one-way) moments, μ_{x^r} and μ_{y^r}, being merely ordinary moments of the one-way frequency distributions, $f(x)$ and $f(y)$, can be computed by the methods of Chapter II.

Let us recapitulate these methods for the first and second moments, using the new notations.

For $f(x)$.[1] Let u be the arbitrary coördinate referred to an arbitrary origin, $X = A$; and let the unit of u equal the class interval h of X. Then,

$$u = \frac{X - A}{h}, \quad X = hu + A, \quad \bar{u} = \frac{1}{N}\Sigma u f(u), \quad \bar{X} = h\bar{u} + A, \tag{1}$$

as already illustrated in § 2.

$$\sigma_u^2 = \mu_{u^2} = \frac{1}{N}\Sigma u^2 f(u) - \bar{u}^2, \quad \sigma_x = h\sigma_u. \tag{2}$$

For $f(y)$. Let v refer to the origin $Y = B$, and have as unit the class interval k of Y. Then,

$$v = \frac{Y - B}{k}, \quad Y = kv + B, \quad \bar{v} = \frac{1}{N}\Sigma v f(v), \quad \bar{Y} = k\bar{v} + B, \tag{3}$$

as in § 2.

$$\sigma_v^2 = \mu_{v^2} = \frac{1}{N}\Sigma v^2 f(v) - \bar{v}^2, \quad \sigma_y = k\sigma_v. \tag{4}$$

To compute p_{xy}. The computation of the product moment is new. We make the same substitutions as in (1) and (3), inserting them in the expression for p_{xy} which might have been written:

$$p_{xy} = \frac{1}{N}\sum_{X,Y}(X - \bar{X})(Y - \bar{Y})f(X, Y). \tag{5}$$

By (1) and (2),

$$X - \bar{X} = h(u - \bar{u}), \; Y - \bar{Y} = k(v - \bar{v}), \text{ so that}$$

$$p_{xy} = \frac{hk}{N}\sum_{u,v}(u - \bar{u})(v - \bar{v})f(u, v) = \frac{hk}{N}\left[\sum_{u,v} uvf - \bar{u}\sum_{u,v} vf - \bar{v}\sum_{u,v} uf + \bar{u}\bar{v}\sum_{u,v} f\right]$$

$$= \frac{hk}{N}\left[\sum_{u,v} uvf - N\bar{u}\bar{v}\right], \text{ by (1) and (3).} \tag{6}$$

[1] The frequency $f(x)$ has the same value as $f(X)$; for, by definition, (x) is the frequency in the column at x, and $f(X)$ is the frequency in the column at X, and x and X are different designations for the same column. Also, $f(u) = f(X)$.

Some authors stop here, performing the computation by means of (6). To do this they place in each cell, in small figures, the value of uv, and then add over the whole table the quantities uvf. It is a little better (except in special cases to be considered later) to proceed as below:[1]

$$\text{Substitute } U = \sum_u uf(u, v), \quad V = \sum_v vf(u, v). \tag{7}$$

Then (6) may be written in either the form (6a) or (6b):

$$p_{xy} = hk\left(\frac{1}{N}\sum_v v\sum_u uf(u, v) - \bar{u}\bar{v}\right) = hk\left(\frac{1}{N}\sum_v vU - \bar{u}\bar{v}\right), \tag{6a}$$

$$p_{xy} = hk\left(\frac{1}{N}\sum_u u\sum_v vf(u, v) - \bar{u}\bar{v}\right) = hk\left(\frac{1}{N}\sum_u uV - \bar{u}\bar{v}\right). \tag{6b}$$

We may compute equally well either (6a) or (6b). Actually in practice we do compute the greater part of both, for we have then a check on the most difficult part of the computation. Because of the equivalence of (6a) and (6b) our check equation reads as follows:

$$\sum_v vU = \sum_u uV. \tag{8}$$

To compute r, we now merely substitute the value of p_{xy} just found in the definition of r:

$$r = \frac{p_{xy}}{\sigma_x\sigma_y} = \frac{\frac{1}{N}\sum_v vU - \bar{u}\bar{v}}{\sigma_u\sigma_v}. \tag{9}$$

Since the expression hk cancels out of (9), it is not necessary[2] to multiply out the product indicated in (6a) or in (6b). All we need, therefore, in order to get r is the following list of numbers: $\bar{u}$, $\bar{v}$, σ_u, σ_v by (1), (2), (3), and (4), $\sum_v vU$ by the use of (7) and (8), then r by (9). This procedure is illustrated in Example 7.

[1] *Handbook of Mathematical Statistics*, Chapter VIII.

[2] Except in rare instances when p_{xy} is needed for some other purpose than to enable one to obtain r.

COROLLARY 1. *By equation* (9), r *is a characteristic of the frequency distribution which is independent of the choice of origin and of the units of measurement.*

COROLLARY 2. *The sign of* r *depends on the directions of the axes which are chosen as the positive directions.*

Proof. The sign of r is the same as the sign of p_{xy}, by (δ). By (γ) the sign of p_{xy} is the same as the sign of Σxyf. Now, let us suppose that, for a certain choice of directions, this quantity is positive, and then suppose that we reverse the direction of the x-axis. The effect is to change the sign of every xy product which follows the summation symbol, and therefore to change the sign of the sum.

Example 7. Find r for the data of Example 2.

v \ u	X / Y	-2	-1	0	1	2	3	f(v)	vf(v)	$v^2f(v)$	U	vU
		12	16	20	24	28	32					
-2	10	5	61	24				90	-180	360	-71	142
-1	14		21	368	202			591	-591	591	181	-181
0	18			8	164	15		187	0	0	omit	0
1	22				1	21	7	29	29	29	64	64
2	26					1	2	3	6	12	8	16
f(u)		5	82	400	367	37	9	(900)	(736)	(992)		(41)
uf(u)		-10	-82	0	367	74	27	(376)				
$u^2f(u)$		20	82	0	367	148	81	(698)				
V		-10	-143	omit	-201	23	11					
uV		20	143	0	-201	46	33	(41)				

Check

$$\bar{u} = \frac{376}{900} = .4178, \quad \bar{v} = \frac{-736}{900} = -.8178; \quad \sigma_u^2 = \frac{698}{900} - \bar{u}^2$$

$$= .6010, \sigma_u = .775; \quad \sigma_v^2 = \frac{992}{900} - \bar{v}^2 = .4335, \sigma_v = 0.658.$$

The figures in the U column are found thus:

$\Sigma_u uf(u, v) = U$; so $(-2)(5) + (-1)(61) + (0)(24) = -71$; $(-1)(21) + (0)(368) + (1)(202) = 181$, etc. Similarly, for the V row, $\Sigma_v vf(u, v) = V$; $(-2)(5) = -10$, $(-2)(61) + (-1)(21) = -143$, etc.

$$r = \frac{\frac{41}{900} - (.4178)(-.8178)}{(.775)(.658)} = 0.581.$$

Grouping Errors. When the material is grouped in a table containing less than 10×10 cells, grouping errors are introduced which ought to be corrected. In general, the smaller the number of cells the larger the grouping errors. These will be discussed further in Part II. They can be corrected in part by the use of Sheppard's corrections applied to σ_u and σ_v. In this text these corrections will not be applied except when specifically indicated. In Example 7, the value of r if obtained from corrected σ's would have been 0.696.

COROLLARY 3. *If, in general, x increases as y increases, r is positive; if x decreases as y increases, r is negative.*

The proof is similar to that of Corollary 2. Examples of cases where r is negative will occur in the problems at the close of the chapter. Negative correlation is sometimes called "inverse" correlation. It will be shown later that, in all cases, $-1 \leq r \leq 1$.

EXERCISE § 4. Compute $\bar{u}$, $\bar{v}$, σ_u, σ_v, r for the data of Exercise 3, § 2. *Ans.*, $r = .559$.

5. Computation of Moments (N Small). In Chapter IV, § 3, we considered the case of simple moments when N was small. It is not necessary to repeat the discussion for $\mu_{r'}$ and $\mu_{r'}$, both of which are simple moments, although the procedure in these cases will be illustrated on page 143. A little further consideration of the product moment is, however, desirable. When N is so small that the data cannot be sepa-

rated into cells without introducing too large a grouping error, it is necessary to abandon the short method of computation just used, but fortunately, since there are not many items in this case, the formula can nevertheless be computed outright without great labor. Each item is now to be treated separately, whether two are alike or not. It is therefore better not to write the $f(u, v)$ now, and to understand by the expression, Σuv, the sum of all pairs, uv, that occur. They do not need to be unequal pairs, as they were before. Moreover, since h and k might be any desired numbers in our formulae, we shall now choose them both to be unity, instead of class intervals, as before. Then formula (6) becomes

$$p_{xy} = \frac{1}{N}(\Sigma uv - N\bar{u}\bar{v}), \tag{10}$$

and, again, $$r = p_{xy} / \sigma_x\sigma_y,$$

where, now,

$$\sigma_x^2 = \frac{1}{N}\Sigma u^2 - \bar{u}^2, \quad \sigma_y^2 = \frac{1}{N}\Sigma v^2 - \bar{v}^2, \quad \bar{u} = \frac{1}{N}\Sigma u, \quad \bar{v} = \frac{1}{N}\Sigma v. \tag{11}$$

These equations are frequently written in the combined form:

$$r = \frac{1}{\sqrt{\frac{1}{N}\Sigma u^2 - \bar{u}^2}\sqrt{\frac{1}{N}\Sigma v^2 - \bar{v}^2}}\left(\frac{1}{N}\Sigma uv - \bar{u}\bar{v}\right). \tag{10a}$$

It should be remembered that (A, B), the origin of the (u, v) system, was arbitrary. Quite commonly, in the case where N is small, this is taken as the given origin. Then $u = X$, $v = Y$, $\bar{u} = \bar{X}$, $\bar{v} = \bar{Y}$, and (10a) may be written:

$$r = \frac{1}{\sqrt{\frac{1}{N}\Sigma X^2 - \bar{X}^2}\sqrt{\frac{1}{N}\Sigma Y^2 - \bar{Y}^2}}\left(\frac{1}{N}\Sigma XY - \bar{X}\bar{Y}\right). \tag{10b}$$

Since we knew before that r was independent both of origin and of units, it was really obvious that the form (10a) could

have been written in the form (10*b*). The arrangement of the computation [1] imitates that in Chapter IV, § 3.

Example 8. Find the correlation between X and Y. X is an index number (*I. Fisher*) of wholesale prices in the United States, Y a corresponding index (*Crump*) for England. The period extends over three and one-half years.

X	Y	u	v	u^2	v^2	uv
150	157	10	17	100	289	170
144	152	4	12	16	144	48
158	167	18	27	324	729	486
160	152	20	12	400	144	240
159	148	19	8	361	64	152
151	142	11	2	121	4	22
147	140	7	0	49	0	0
$N = 7$		89	78	1371	1374	1118

Here we took $A = 140$, $B = 140$, and used formula (10*a*):

$$\bar{u} = \frac{89}{7} = 12.71, \quad \bar{v} = \frac{78}{7} = 11.14, \quad \sigma_u^2 = \frac{1371}{7} - \bar{u}^2 = 34.21,$$

$$\sigma_v^2 = \frac{1374}{7} - \bar{v}^2 = 72.12, \quad p_{uv} = \frac{1118}{7} - (12.71)(11.14) = 18.04,$$

$$r = \frac{18.04}{\sqrt{(34.21)(72.12)}} = 0.36.$$

EXERCISES § 5

Find r in the following cases:

1.

X	4	3	7	6	9	15
Y	30	20	60	50	100	100

Ans., $r = .90$.

2.

A	9	8	6	5	3	2
B	1	3	5	7	9	11

[1] Formula (10*b*) can be computed with very great speed on certain types of multiplying machines; all the following quantities being given as the result of a single set of operations:

$$\Sigma X, \Sigma Y, \Sigma X^2, \Sigma Y^2, \Sigma XY.$$

3.

A	1	2	3	4	5	6
X	1	4	9	16	25	36

Ans., $r = .98$.

4.

x	-3	-2	-1	0	1	2	3
x^3	-27	-8	-1	0	1	8	27

Ans., $r = .93$.

5. Death rates in Connecticut and Massachusetts, per 100,000 population.

Year	1924	1923	1922	1921	1920	1919	1918
Connecticut	11.3	12.0	12.0	11.4	13.6	13.3	20.4
Massachusetts	12.0	13.0	12.8	12.2	13.8	13.6	20.9

6. Would r have been changed if each of the Y's in Exercise 1 had been divided by 10? Why?

PROBLEMS CHAPTER VIII

1. As in Example 3, write out in full the expressions for: (*a*) $f(X)$ when $X = 32$, (*b*) $f(Y)$ when $Y = 22$. (p. 132).

2. Represent in the form of a correlation table the data of Problem 3, Chapter 6, page 99. Let the longitudinal distance from the mark be X, and the latitudinal distance be Y, in each case. What are $\bar{X}$, $\bar{Y}$, $f(X = 100)$, $\sum_Y f(90, Y)$, $\sum_X f(X, 10)$?

3. Find r in the following table. (The data are fictitious and N is really too small to permit of the grouping indicated. Small numbers are chosen in order to produce a simple problem.)

Pounds \ *Feet*	5	9	13	17
1	1	2	1	
8		7	2	
15			2	1

Ans., $r = 0.597$.

4. Find the correlation between the length and breadth of a book, as derived from a group of 900 chosen nearly at random from a large library. The unit is one centimeter. *Ans.*, $r = 0.875$.

B \ L	12 –	14 –	16 –	18 –	20 –	22 –	24 –	26 –	28 –	30 –	32 –
8 –	3	5	6								
10 –	2	13	37	24							
12 –		1	18	186	67	9	1				
14 –			2	21	94	171	21				
16 –				3	4	59	77	4			
18 –					1	4	24	7	4		
20 –								4	6	1	
22 –							1	5	6	5	1
24 –									1	1	1

5. Find the correlation between the length and thickness of the books of Problem 4.

T \ L	12 –	14 –	16 –	18 –	20 –	22 –	24 –	26 –	28 –	30 –	32 –
0 –	1	5	2	5		6					
1 –	1	7	15	41	17	22	5	2			
2 –	3	5	33	109	57	43	10	2	2	1	
3 –		2	7	69	70	77	31	9	5	3	
4 –			2	9	14	55	35	5	6	1	
5 –				1	5	28	24	2	3	1	
6 –			1		2	5	10		1		1
7 –			2		1	2	5				1
8 –						4	3				1
9 –			1			1	1				
Totals	5	19	63	234	166	243	124	20	17	6	3

6. Make a coarser grouping of the data of Problem 4 as indicated below, and find r. Because of the large grouping errors introduced, the (uncorrected) result here will be very different from that in Problem 4.

B \ *L*	10 —	18 —	26 —	*Totals*
8 —				
14 —				
20 —				
Totals				900

7. Select 100 books at random from a library, form a table similar to the one in Problem 6, and find r. A random selection (by the use of *Tippett's* numbers) from the table of Problem 6 is:

5	41	
	48	3
		3

8. Find the correlation between the height of "mid-parent" and the height of adult child.

INCHES

Parent \ *Child*	64	65	66	67	68	69	70	71	72	73	*Totals*
72						1	2	2	2	1	
71				2	4	5	5	4	3	1	
70	1	2	3	5	8	9	9	8	5	3	
69	2	3	6	10	12	12	12	10	6	3	
68	3	7	11	13	14	13	10	7	3	1	
67	3	6	8	11	11	8	6	3	1		
66	2	3	4	6	4	3	2				

9. Find the correlation between the ages of husband and wife. Data in Example 1, Chapter X, § 1 (p. 166).

10. Find the correlation between the statures of fathers and mothers. *Biometrika*, vol. 2, p. 408, slightly altered to avoid fractions. $N = 1079$. The data (F, M) are on page 148.

11. Find the correlation between brokers' loans and stock prices by months for the years 1928 and 1929. Data in Chapter VII, Problems 2 and 3 (p. 126).

12. Find the correlation between weight and stature. *Biometrika*, vol. 20 *A*, p. 306. 733 boys in American private schools, ages 4–20. *Ans.*, $r = .931$. The data (Kilo., Cm.) are on page 148.

REVIEW EXERCISES CHAPTERS I-VIII

1. Find, in the t system, the mean, mode, σ, α_3, and α_4 for (*a*), (*b*), (*c*), (*d*), and (*e*):

(*a*)

t	f
0	1
1	6
2	12
3	8

(*b*)

t	f
0	1
1	8
2	24
3	32
4	16

(*c*)

t	f
135–142	7
127–134	138
etc.	906
	977
	161
	3

(*d*)

t	f
30	5
40	14
50	28
60	37
70	22
80	24
90	6

(*e*)

t	f
4	1
6	5
8	25
10	20
12	19
14	10

Answers

	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Mean	2	$2\frac{1}{2}$	118.27	61.25	10.025
Mode	$2\frac{1}{2}$	$2\frac{1}{2}$		61.44	
σ	$\sqrt{\frac{1}{4}}$	$\frac{1}{2}\sqrt{2}$	5.93	11.8	2.33
α_3	$-\sqrt{\frac{1}{4}}$	$-\sqrt{\frac{1}{4}}$	.084		.0206
α_4	$2\frac{1}{2}$			2.31	

M \ F	59-	60-	61-	62-	63-	64-	65-	66-	67-	68-	69-	70-	71-	72-	73-	74-	75-
52-							1										
53-								1									
54-																	
55-				1				1		1							
56-				1	1		1	2	2								
57-			1	1	2	1	4	3	2	3	1						
58-		1	1	1	2	4	7	5	4	4	2	2					
59-		1	1	1	4	6	7	10	15	17	9	5	3	1	1		
60-	1	1	2	4	5	18	16	25	15	23	13	7	5	4	1		
61-				3	3	16	17	26	21	24	15	14	10	4	1		
62-		1	2	4	9	10	26	22	28	26	23	14	11	4	3	1	
63-			2	3	4	9	21	16	20	19	24	22	11	5	2	2	1
64-		1		2	6	6	10	16	19	24	17	14	7	5	2		1
65-			1	1	2	3	6	9	7	15	13	10	6	5	2		
66-				1	1		3	5	3	7	3	7	3	3	1	1	
67-							1	3	1	3	3	3	2	1	1		
68-									1	2	2	1		1			
69-												2		2			
70-												1					

Kilo. \ Cm.	100.5-	105.5-	110.5-	115.5-	120.5-	125.5-	130.5-	135.5-	140.5-	145.5-	150.5-	155.5-	160.5-	165.5-	170.5-	175.5-	180.5-	185.5-	190.5-
95-																	1		
91-																			
87-																			
83-																		1	1
79-																	1	3	1
75-																1	4	1	
71-															1	4	2		
67-															4	9	6		
63-														1	20	10	5	1	
59-											1	1	4	11	20	12	5	1	
55-												1	3	16	12	6			
51-											2	5	12	19	14	4			
47-									1	1	9	14	17	14	4	1			
43-									1	6	14	30	16	4					
39-								1	2	21	24	18	4						
35-								7	12	24	21	2	2						
31-						1	14	22	22	13	4	2							
27-				1	1	13	18	20	5										
23-				2	20	20	11	2											
19-		6	6	11	6		1												
15-	1	4	5																
11-		1																	

2. (*a*) Find the median and the quartiles in Exercise 1 (*c*). *Ans.*, 118.13, 122.94, 113.64. (*b*) Find the median and P_{20} in Exercise 1 (*d*). *Ans.*, 60.675, 47.93. (*c*) Find Q_1, Q_2, Q_3 for the table: $t = -10, -8, -7.5, -6, -6, -4, 0, 0, 2, 3.5, 6, 8, 8$. *Ans.*, $-6.375, 0, 4.125$. (*d*) Find D_2, D_7, D_8 in Exercise 1 (*e*). *Ans.*, 7.80, $11\frac{1}{2}\frac{8}{5}$, $12\frac{7}{10}$.

3. Graduate by the normal curve the data:

t	10	8	6	4	2
f	3	4	6	5	2
Ans.	1.8	4.8	6.4	4.5	1.6

4. The probable errors of a high angle gun are 40 yards in longitude and 5 yards in latitude. The center of impact is at the center of the farther side of a square fort, and this side is perpendicular to the line of fire and is 30 yards long. What per cent of hits are expected within the fort? *Ans.*, 18.

5. From the following measurements of the radius of a cylinder show that the mean radius is $3.013 \pm .0025$.

3.00, 3.00, 3.00, 3.01, 3.01, 3.01, 3.02, 3.02, 3.03, 3.03.

6. The probable error of a gun in longitude is 50 meters and in latitude 10 meters. How many shots are expected on a rectangle 260 meters long (in the direction of fire) and 40 meters wide? *Ans.*, $\frac{1}{100}$.

7. Derive graphically A and B of the trend equation, $y = A + Bt$, taking $t = 0$ at 1900:

Date	1900	1901	1902	1903	1905	1906
y	7	6	4.5	4.2	2	1.5

The *numerical* solution is: $y = 6.82 - .92t$.

8. Find numerically the trend of the average age of American Rhodes scholars (*Burgess*), taking $t = 0$ at 1900. *Ans.*, $y = 21.56 + .0267t$.

Date of Matriculation	1904	1907	1910	1913	1916
Average Age of Scholars	21.7	21.7	21.9	21.8	22.05

9. Find the correlation between year and age. Year is the year of matriculation, and age is the age at matriculation of American Rhodes scholars (*Burgess*). *Ans.*, .00308.

v \ u / Year \ Age	-1	0	1	
	19–20	21–22	23–24	*Totals*
-1	1904–7	14	51	34	99
0	1908–11	25	45	39	109
1	1912–16	15	46	35	96
	Totals	54	142	108	304

10. In Exercise 9, what are the following sums: $\Sigma_v f(1, v)$, $\Sigma_u f(u, 1)$, $\Sigma_v \Sigma_u u f(u, v)$, $\Sigma_u \sqrt{u+1}\, \Sigma_v f(u, v)$, $\Sigma_{u,v} \frac{v}{u+2} f(u, v)$, $\frac{1}{f(u)} \Sigma_v f(u, v)$ if $u = 0$. Write in Σ form the totals 109 and 142.

Ans., 108, 96, 54, $142 + 108\sqrt{2}$, -1.17, 1, etc.

11. Find graphically the equation of the best-fitting exponential curve for the following data:

t	0	1	.5	.7	1.2	1.5
y	2	.10	.45	.25	.05	.02

Ans., $y = 2e^{-3t}$ (approximately).

12. Find r:

Y \ X	27	29	31	$f(Y)$
62	2	3	6	11
65	1	10	12	23
68		1	5	6

Ans., .216.

13. Using your tables, construct a "ladder of dispersion," putting in twice as many rungs as ordinarily used; *i.e.*, the difference between two consecutive rungs is to equal half the probable error.

14. (*a*) In a normal distribution in which mean $t = 0$ and $\sigma_t = 5$, what proportion of the data will lie where $t > 15$? *Ans.*, .00135. (*b*) If 100 of the data lie between $t = -7$ and $t = -9$, how many data are there in the whole distribution? *Ans.*, 2227.

15. Find the correlation between y and t:

y	2.5	7.5	12.0	20.0	35.0
t	1	2	2.5	3	3.5

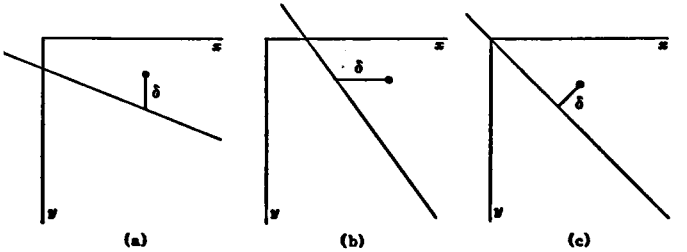
Ans., .93.

CHAPTER IX

REGRESSION, INTERPRETATION OF r

1. Regression Lines. Consider the general case of correlation, where N is large, and the data might be represented by dots spread over the paper. Suppose we wish to draw and to find the equation of that straight line which, on the whole, will come nearest to all these dots. We shall suppose the best-fitting line is that one which fits best in the sense of least squares, but even with this understanding there are at least three different possible points of view. Let δ be the distance between a dot and the line. We wish to make $\Sigma\delta^2$ a minimum. The three cases that arise depend on whether:

Case (*a*) δ is measured parallel to the y-axis,
Case (*b*) δ is measured parallel to the x-axis, or
Case (*c*) δ is measured perpendicular to the line.



In Case (*a*), the line is called the "regression of Y on X"; in Case (*b*) it is called the "regression of X on Y"; in Case (*c*) it has no generally accepted name. We shall call it the "geometrically best-fitting line," because in geometry we usually prefer to think of the distance between a point and a line as measured perpendicular to the line.

Case (*a*). If our material is grouped and at the same time the picture of the dots is kept in mind, one must think of the dots lying many deep at the central points of our cells. The number of dots in a cell whose coördinates are (X, Y) is, of course, $f(X, Y)$, and the number of dots in a vertical column whose coördinate is X is $f(X)$. If we had only *one* point in each column, our problem would be simply to draw the trend line to the several points of our table, as in Chapter VII. Now, although instead of one point in a column we have $f(X)$ such points, we shall nevertheless proceed in a manner analogous to that used for the trend. We shall use the method of moments. It will turn out that the result would not have been different had we first replaced each set of $f(X)$ points with one point at their mean position, and found the trend line to the points, one in each column, thus located, except that in doing so it would have been necessary to have weighted these several mean points proportionally to their $f(X)$'s.

First, let $x = X - \bar{X}$, $y = Y - \bar{Y}$, as in Chapter VIII. To find A and B such that the line

$$y = A + Bx \tag{1}$$

will best fit the data, equate the zero*th* and the first moments in the x direction:

$$\left.\begin{aligned} &(0th \text{ moments}) \quad \frac{1}{N}\Sigma_{x,y} yf(x,y) = \frac{1}{N}\Sigma_{x,y}(A + Bx)f(x,y) \\ &(1st \text{ moments}) \quad \frac{1}{N}\Sigma_{x,y} xyf(x,y) = \frac{1}{N}\Sigma_{x,y} x(A + Bx)f(x,y) \end{aligned}\right\}. \tag{2}$$

Simplifying these,

$$\left.\begin{aligned} 0 &= A + 0 \\ p_{xy} &= 0 + B\sigma_x^2 \end{aligned}\right\}, \tag{3}$$

for

$$\frac{1}{N}\Sigma_{x,y} yf(x,y) = \frac{1}{N}\Sigma_y y\Sigma_x f(x,y) = \frac{1}{N}\Sigma_y yf(y) = \text{mean of } f(y),$$

relative to its mean as origin. This is zero. Similarly,

$$\frac{1}{N}\Sigma\Sigma_{x,y} Bxf(x,y) = \frac{B}{N}\Sigma_x xf(x) = B \cdot 0 = 0.$$

From (3) we obtain

$$A = 0, \quad B = \frac{p_{xy}}{\sigma_x^2} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}. \tag{4}$$

Hence, the equation (1) is:

$$y = r\frac{\sigma_y}{\sigma_x}x, \quad \text{or} \quad \frac{y}{\sigma_y} = r\frac{x}{\sigma_x}. \tag{5}$$

This is the equation of the regression of Y on X in the (x, y) coördinate system. We may now transfer back to the (X, Y) system, and get

$$\frac{Y - \bar{Y}}{\sigma_y} = r\frac{X - \bar{X}}{\sigma_x}. \tag{5a}$$

Here $\sigma_y = \sigma_Y$, $\sigma_x = \sigma_X$, because the units of x and y are the same as the units of X and Y. For the purpose of plotting the regression line, it is more convenient to use the (u, v) system. The equation in this system will, of course, turn out to be of the same form as (5a), because in (5a) both the origin and units were arbitrary. It is

$$\frac{v - \bar{v}}{\sigma_v} = r\frac{u - \bar{u}}{\sigma_u}. \tag{5b}$$

Case (b). An exactly analogous discussion yields the analogous equations for the regression of X on Y:

$$\frac{x}{\sigma_x} = r\frac{y}{\sigma_y}, \tag{6}$$

$$\frac{X - \bar{X}}{\sigma_x} = r\frac{Y - \bar{Y}}{\sigma_y}, \tag{6a}$$

$$\frac{u - \bar{u}}{\sigma_u} = r\frac{v - \bar{v}}{\sigma_v}. \tag{6b}$$

It should be noticed immediately that, unless $r = \pm 1$, (5) is not the same line as (6). If it were, then when (6) is solved for y we should obtain the same function of x as in (5), but (6) yields

$$y = \frac{1}{r}\frac{\sigma_y}{\sigma_x}x,$$

which is not the same as

$$r\frac{\sigma_y}{\sigma_x}x,$$

unless $r = 1/r$, *i.e.*, unless $r = \pm 1$. The relation between (5) and (6) will be exhibited more fully after a consideration of Case (*c*). But first we pause to illustrate the results already obtained, and to prove the statement made earlier that the equation of regression, y on x, might have been found by first replacing our given table by a new one, formed from the first by replacing all the data in each column by the same number of data clustered at the mean of that column. In that case the ordinate of this mean point would have been

$$\bar{y}_x = \frac{1}{f(x)}\sum_y yf(x, y), \tag{7}$$

and the frequency at that point would have been $f(x)$. Instead of finding, as in (2), the moments of the ordinates y of all the points of the table, we should use merely the moments of the $\bar{y}_x$'s. We should have:

$$\left.\begin{aligned} (0th \text{ moments}) \quad & \frac{1}{N}\sum_x \bar{y}_x f(x) = \frac{1}{N}\sum_x (A + Bx)f(x) \\ (1st \text{ moments}) \quad & \frac{1}{N}\sum_x x\bar{y}_x f(x) = \frac{1}{N}\sum_x x(A + Bx)f(x) \end{aligned}\right\}. \tag{8}$$

But, by (7),

$$\frac{1}{N}\sum_x \bar{y}_x f(x) = \frac{1}{N}\sum_x\sum_y yf(x, y) = 0,$$

as at the beginning of (2). Also, in an analogous manner,

$$\frac{1}{N}\sum_x (A + Bx)f(x) = A + 0;$$

and

$$\frac{1}{N}\sum_x x\bar{y}_x f(x) = \frac{1}{N}\sum_x\sum_y xyf(x, y) = p_{xy},$$

and, finally, the last expression in (8),

$$\frac{1}{N}\sum_x x(A + Bx)f(x) = 0 + B\sigma_x^2.$$

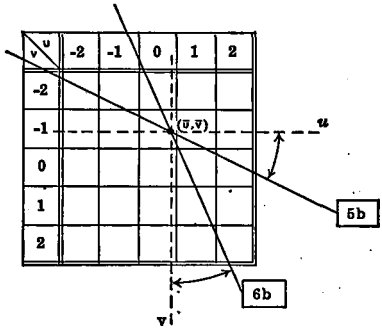
So equations (8) become:

$$\left.\begin{aligned} 0 &= A + 0 \\ p_{xy} &= 0 + B\sigma_x^2 \end{aligned}\right\}$$

which are the same as (3), page 153, and they yield the same values for A and B.

Of course, a similar statement holds for the regression of X on Y.

Example 1. Find the regression lines in Example 7, page 140, of the preceding chapter, and plot them across the table.



By (5*b*) $$\frac{v + .82}{.66} = (.58)\left(\frac{u - .42}{.78}\right).$$

This line goes through $(.42, -.82)$ and has a slope of

$$\frac{(.58)(.66)}{.78} = 0.49.$$

By (6*b*) $$\frac{u - .42}{.78} = (.58)\left(\frac{v + .82}{.66}\right).$$

This goes through the same point and makes with the v-axis an angle whose tangent is 0.69. We must remember here that the positive end of the v-axis is downwards.

EXERCISES § 1

1. Plot across the table of Example 1 the following lines:

(*a*) $$\frac{v-1}{2} = .2\left(\frac{u-2}{3}\right),$$

(*b*) $$\frac{v+2}{1} = .8\left(\frac{u-2}{3}\right),$$

(*c*) $$\frac{u-.3}{.2} = .4\left(\frac{v+.2}{.2}\right),$$

(*d*) $$\frac{u+1}{.5} = .5\left(\frac{v+1}{.4}\right).$$

2. Find both regression lines in the following cases, using the (x, y) coördinates:

(*a*) $\sigma_u = 1$, $\sigma_v = 1.5$, $h = 4$, $k = 5$, $r = .2$;

(*b*) $\sigma_x = 1$, $\sigma_y = 1$, $k = 1.5$, $r = .5$;

(*c*) $\sigma_X = 2$, $\sigma_Y = 3$, $r = 0.3$.

Ans., (*a*) $y = .375x$, $x = .107y$; (*b*) $4y = 3x$, $3x = y$; (*c*) $20y = 9x$, $5x = y$.

2. Least Squares.[1] It was stated that in each of the three cases the quantity $\Sigma\delta^2 f$ was a minimum. Let us prove this statement for the two cases already discussed, beginning with

Case (*a*). Here $\delta = y - (A + Bx)$, and therefore

$$\frac{1}{N}\sum_{x,y}\delta^2 f = \frac{1}{N}\sum_{x,y}(y^2 + A^2 + B^2x^2 - 2Ay - 2Bxy + 2ABx)f(x,y)$$

$$= \sigma_y^2 + A^2 + B^2\sigma_x^2 - 2Br\sigma_y\sigma_x. \tag{9}$$

In order to choose A and B so as to make this essentially positive expression a minimum, it is obvious that first we should take $A = 0$. Then the expression may be written

$$\sigma_y^2 + \sigma_x^2\left(B^2 - 2Br\frac{\sigma_y}{\sigma_x}\right). \tag{10}$$

[1] This section is not needed later and may be omitted if desired.

To make this a minimum, it is merely a question of choosing B so that the quantity in parenthesis

$$B^2 - 2Br\frac{\sigma_y}{\sigma_x}$$

shall be a minimum. Students of the calculus will readily see that for this purpose it is necessary to make $B = r\sigma_y/\sigma_x$. For others, it is sufficient to prove the following simple theorem in analytics:

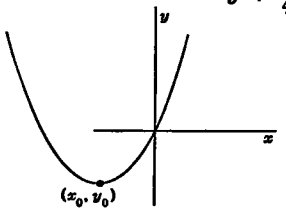
Theorem. *The minimum value of the function,*

$$x^2 + ax,$$

occurs at the point where $x = -a/2$.

Proof. The graph of the equation $y = x^2 + ax$ is a parabola which is concave up, and whose vertex is at the point $(x_0 = -a/2,\ y_0 = -a^2/4)$; for, by adding $a^2/4$ to both sides of the equation, one may write it in the form:

$$y + \frac{a^2}{4} = \left(x + \frac{a}{2}\right)^2,$$



or

$$y - y_0 = (x - x_0)^2.$$

Of course the function has its minimum at this vertex, and here

$$x_0 = -\frac{a}{2}.$$

Using this theorem, we note that in the expression

$$B^2 - 2Br\frac{\sigma_y}{\sigma_x},$$

B plays the rôle of x, and $-2r\frac{\sigma_y}{\sigma_x}$ the rôle of a. Hence the rôle of $-a/2$ is played by $r\frac{\sigma_y}{\sigma_x}$, and so, when $B = r\frac{\sigma_y}{\sigma_x}$,

this function of B is a minimum. We have thus been led to choose $A = 0$, and $B = r\frac{\sigma_y}{\sigma_x}$ in the equation $y = A + Bx$, and the result is

$$y = r\frac{\sigma_y}{\sigma_x}x,$$

as before, equation (5).

The actual value of $\frac{1}{N}\Sigma\delta^2 f$ in this case is also of interest, for it is obviously a measure of the closeness with which the dots cluster about the line of regression. By (9) and (10),

$$\frac{1}{N}\Sigma\delta^2 f = \sigma_y^2 + \sigma_x^2 r\frac{\sigma_y}{\sigma_x}\left(r\frac{\sigma_y}{\sigma_x} - 2r\frac{\sigma_y}{\sigma_x}\right) = \sigma_y^2(1 - r^2). \quad (11)$$

Since the left-hand side of (11) is by nature positive (or zero), so is the right-hand side. Therefore

$$1 - r^2 \geq 0, \text{ and } -1 \leq r \leq 1,$$

a fact we had stated to be true before, but had not proved.

For *Case* (*b*), $\delta = x - (A + By)$. In Problem 3 the student is asked to show that here $\frac{1}{N}\sum_{x,y}\delta^2 f(x, y)$ will be a minimum if $A = 0$, and $B = r\frac{\sigma_x}{\sigma_y}$, and that in this case

$$\frac{1}{N}\Sigma\delta^2 f = \sigma_x^2(1 - r^2). \quad (12)$$

Case (*c*). This case can be treated easily by least squares, but not by moments. This is why it has been postponed till now. Again, let us suppose the equation of the line to be in the form, $y = A + Bx$. The distance δ from this line to a point (x', y') off the line is, by analytics,

$$\delta = \frac{y' - A - Bx'}{\sqrt{B^2 + 1}}, \quad (13)$$

and we may drop the primes, if, instead of calling the point (x', y'), we call it (x, y), but we must be careful not to confuse

it with a point (x,y) on the line itself. Our problem is now so to choose A and B that

$$\frac{1}{N}\sum_{x,y}\left(\frac{y - A - Bx}{\sqrt{B^2+1}}\right)^2 f(x, y) \tag{14}$$

shall be a minimum. This function may be written (cf. equation (9)):

$$\frac{1}{B^2+1}(\sigma_y^2 + A^2 + B^2\sigma_x^2 - 2Br\sigma_y\sigma_x). \tag{15}$$

To make this a minimum, we first put $A^2 = 0$. Then the function is

$$\frac{\sigma_y^2 + B^2\sigma_x^2 - 2Br\sigma_y\sigma_x}{B^2+1}. \tag{16}$$

The value of B necessary to minimize this is not a simple expression, except in an important special case, *viz.*, when $\sigma_x = \sigma_y = 1$. Let us continue the discussion with this special case only in mind. The function (16) becomes

$$1 - \frac{2Br}{1+B^2}, \tag{17}$$

and the student of the calculus again has an easy task to show that this is a minimum when $B = \pm 1$. Such students are asked to prove this in Problem 4, and others are led, in Problem 5, to an approximate proof in a numerical case. The plus sign for B is to be used when $r > 0$, the minus sign when $r < 0$. If $r = 0$, there is no minimum value, all lines (for which $A^2 = 0$) fitting equally well. So finally our equation becomes

$$\textit{Case (c):} \quad \left.\begin{array}{l} y = x \quad \text{if } r > 0 \text{ and } \sigma_y = \sigma_x = 1 \\ y = -x \text{ if } r < 0 \text{ and } \sigma_y = \sigma_x = 1 \end{array}\right\} \tag{18}$$

It is now interesting to write also the regression lines for the case where $\sigma_x = \sigma_y = 1$. They are:

$$\left.\begin{array}{l} \textit{Case (a):}\ y = rx \\ \textit{Case (b):}\ x = ry \end{array}\right\} \tag{19}$$

COROLLARY 1. *The values of* $\frac{1}{N}\Sigma\delta^2 f$ *in the three cases are:* (*a*) $1 - r^2$, (*b*) $1 - r^2$, (*c*) $1 - |r|$.

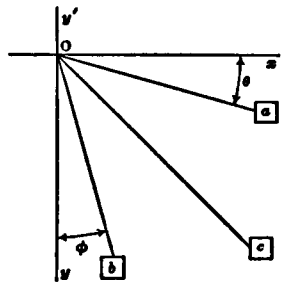
Proof of (*c*). By (15), when $\sigma_x = \sigma_y = 1$, and $A = 0$ and $B = \pm 1$,

$$\frac{1}{N}\Sigma\delta^2 f = \frac{1 + 1 - 2r}{2} = 1 - r \text{ if } r > 0,$$

$$= \frac{1 + 1 + 2r}{2} = 1 + r \text{ if } r < 0.$$

COROLLARY 2. *By Corollary* 1, $|r|$ *measures the closeness with which the dots cluster about the geometrically best-fitting line;* r^2 *the closeness with which they cluster about the regression lines (distances in the last case being measured parallel to the* y- *and* x-*axes, respectively).*

COROLLARY 3. *If* $\sigma_x = \sigma_y = 1$, *line* (*c*) *bisects the angle between lines* (*a*) *and* (*b*)



Proof (when $r > 0$). By (18) and (19) all these lines go through the origin. Now, by (19), for the line (*a*), $y/x = r$. So $r = \tan\theta$, in the figure. For the line (*b*), $x/y = r$. So also $r = \tan\phi$, in the figure. Therefore $\phi = \theta$. The angle between line (*a*) and line (*c*) is $(45° - \theta)$, and the angle between lines (*b*) and (*c*) is $(45° - \phi)$. Since $45° - \theta = 45° - \phi$, the corollary is established.

Because $\tan\phi = \tan\theta = r$, it follows that, as r increases from 0 to 1, lines (*a*) and (*b*) start from coincidence with the x- and y-axes, respectively, and rotate with equal angular velocities in the direction of (*c*). When they reach (*c*), the three lines coincide, and $r = 1$.

When r is negative, similar remarks hold for the quadrant xoy'. In the figure, we have supposed y positive when drawn downwards.

COROLLARY 4. *Let the standard deviations be chosen as units. Then the coefficient of correlation measures the degree to which it is true that a change in one variable determines an equal change in the other.*

This is probably the best simple description of the character of the coefficient of correlation which can be given in words, without the aid of mathematical symbols.[1] Before we can prove it, we must state it in more precise language: The larger $|r|$ is, the more closely do the dots lie to the line (c); and for points on (c) it is exactly true that a change in one variable determines an equal change in the other.

Proof. The first part of this statement is contained in Corollary 2. To prove the second part, let (X', Y') and (X'', Y'') be any two points on (c). Then the slope of the line joining them is unity, and so

$$\frac{Y'' - Y'}{X'' - X'} = 1.$$

Hence, $X'' - X' = Y'' - Y'$, which is the same as saying that the change in X in going from one point to the other equals the change in Y.

COROLLARY 5. *The coefficient of correlation measures the degree to which it is true that a relative change in one variable determines an equal relative change in the other. By a relative change is meant the ratio of the absolute change to the standard deviation.*

This of course is a restatement of Corollary 4.

PROBLEMS CHAPTER IX

1. Find the regression-lines in the following cases, and plot them across the tables (pp. 144–147):

In Chapter VIII: (a) Problem 3, (b) Problem 4, (c) Problem 5, (d) Problem 6, (e) Problem 8, (f) Problem 9, (g) Problem 10.

[1] It does not, as sometimes thought, presuppose that the distribution is "normal," or that the "regression is linear."

2. Actually compute $\frac{1}{N}\Sigma\delta^2 f$ for the regression of y on x in the following cases, and show that it equals $\sigma_y^2(1 - r^2)$, as stated in equation (11):

In Chapter VIII: (*a*) Problem 3, (*b*) Problem 4.

3. Prove equation (12).

4. (For calculus students.) Prove that (17) has a minimum when $B^2 = 1$.

5. (For students who do not use the calculus.) Actually compute the function (17) for various values of B near to 1, taking $r = 0.3$, and show from the graph that (17) apparently has a minimum when $B = 1$.

6. Show that, in the (u, v) system, the line of geometric best fit has the equation

$$\frac{v - \bar{v}}{\sigma_v} = \pm\frac{u - \bar{u}}{\sigma_u}.$$

7. Plot the line of Problem 6, as well as the two regression lines in the following problems of Chapter VIII: 5, 8, 9.

8. State and prove Corollary 3 for the case where $r < 0$. Is it true when $r = 0$?

9. (*a*) Show that the maximum value possible for $\frac{1}{N}\Sigma\delta^2 f$ is unity in all three cases, (*a*), (*b*), (*c*), treated on page 161.

(*b*) Adopting, therefore, $\left(1 - \frac{1}{N}\Sigma\delta^2 f\right)$ as a measure of goodness of fit of the line to the points, show that it equals r^2 in Cases (*a*) and (*b*), and $|r|$ in Case (*c*).

(*c*) For Case (*c*) show that when $r = .5$ the fit is twice as good as when $r = .25$.

CHAPTER X

NORMAL SURFACE. CORRELATION OF NON-MEASURABLE CHARACTERS

1. The Normal Surface. Just as there is a normal curve which is useful as an approximation to the forms of many frequency distributions when one independent variable only is involved, so, when we have two independent variables, there is a normal surface.[1] Its equation is,

$$z = \frac{N}{2\pi\sqrt{1-r^2}\sigma_x\sigma_y} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)}, \qquad (1)$$

where, as before, $x = X - \bar{X}$, $y = Y - \bar{Y}$. If also $\sigma_x = \sigma_y = 1$, this equation becomes

$$z = \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(x^2 - 2rxy + y^2)}. \qquad (2)$$

If, in (2), we let x represent a constant $x = x_0$, we are confining our attention to a section of this surface which lies in a plane parallel to the yz plane and at a distance x_0 from the origin. This section is a curve whose equation has the form:

$$z = \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(x_0^2 - 2rx_0y + y^2)}.$$

Complete the square in the parenthesis:

$$z = \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(r^2x_0^2 - 2rx_0y + y^2 - r^2x_0^2 + x_0^2)}$$

$$= \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}(rx_0 - y)^2} \times$$

$$e^{-\frac{1}{2(1-r^2)}(x_0^2 - r^2x_0^2)} = C\, e^{-\frac{y^2}{2(1-r^2)}}. \qquad (3)$$

[1] See Figure 3, § 1, Chapter VIII. The student who has not studied solid analytic geometry may omit the remainder of this section.

where C is a constant, and $y' = rx_0 - y$. This is the equation of a normal curve. Similarly, a section parallel to the xz plane is a normal curve.

Now it often happens that a correlation table is such that it represents a two-way frequency distribution which can be approximately described by a properly chosen normal surface. In such a case, the letters which we have called σ_x and σ_y should be chosen equal to the standard deviations of the table, r should be placed equal to the coefficient of correlation of the table, the origin should be placed at its mean, and N should be chosen equal to the total frequency. These are the reasons why these letters were employed in equation (1), but any other constants would have given us a normal surface. In such a table each array is normal. This fact is proved for the columns by equation (3). Also, each marginal total is normal. This could be easily established from (1) or (2) by the use of the integral calculus. If we fix our attention on those cells of the table in which the frequency has a fixed value, it will be found that they all lie on an ellipse. This can be proved from (2) by letting z equal a constant. Then it follows that

$$x^2 - 2rxy + y^2 = k, \tag{4}$$

where k stands for some constant. This is the equation of an ellipse. It pays to study its form a little. Some of its properties are:

(*a*) *The center is at the origin.*

(*b*) *Its axes make angles of* 45° *with the* x- *and* y-*axes.*

(*c*) *Its semi-axes* a *and* b *are*

$$a = \sqrt{\frac{k}{1-r}}, \quad b = \sqrt{\frac{k}{1+r}}, \quad \text{if } r \geq 0.$$

(*d*) *Hence if* r *is positive,* $r = 1 - \dfrac{2}{\dfrac{a^2}{b^2} + 1}$, *and the nearer* b *is to* a *the nearer also* r *is to* 0. *When* $r = 0$, *the ellipse is a*

circle; that is, all sections of the surface perpendicular to the z-axis are circles. We have then a surface of revolution.

The proofs of these properties are not difficult for one who has studied the equation of the ellipse containing the xy term, for (4) is such an equation. The property (d) may be used as a means of estimating r in a table, provided the scales are so chosen that the geometrical lengths of σ_x and of σ_y are nearly equal. This is illustrated in the diagram of Example 1. First we pick a frequency that occurs often. The frequency chosen in this diagram was 10. Then we estimate the approximate position of this frequency (10) on those arrays where it does not actually occur. Mark all these estimated and actual positions of this frequency by dots, and draw, free-hand, an ellipse which will, as nearly as possible, go through these points and at the same time have axes making 45° with the x- and y-axes. Measure the ratio a/b for this ellipse and substitute in (d). Our actual estimate in this case was $a/b = 34.5/8.5$; hence $r = 0.885$. The true value of r is 0.91.

Example 1. Find r approximately by the use of (d). Data from *Yule's* text.

Age of Wife

Age of Husband														
2	2													
16	173	46	4	1										
4	185	402	84	10	2	1								
1	41	265	411	84	12	2	1							
	9	69	251	369	80	12	2	1						
	3	17	71	219	309	66	12	2	1					
	1	6	20	66	178	252	59	10	2	1				
		2	8	19	57	146	195	44	10	2				
		1	3	8	18	46	110	141	35	6	1			
			1	3	8	16	39	81	101	23	4	1		
			1	1	3	6	11	26	53	58	13	2	1	
				1	1	2	5	8	18	31	31	6	1	
					1	1	2	3	5	10	14	12	2	
							1	1	1	2	4	5	3	1
										1	1	1	1	

EXERCISES § 1

1. Construct a numerical table which is nearly normal by finding the ordinates for various values of x and y of equation (2), for $N = 100$. This will correspond to Table I(a) for the one-dimensional case, except that we are now taking $N = 100$ instead of 1. Take $r = 0.8$. The approximate table is given below. The computation should yield one place more of accuracy if one computes (2) by the use of four-place logarithm tables.

y \ x	−2	−1	0	1	2
−2	2.9	2.2	0.1		
−1	2.2	15.2	6.6	0.2	
0	0.1	6.6	26.5	6.6	0.1
1		0.2	6.6	15.2	2.2
2			0.1	2.2	2.9

(a) Show that $f(x)$ and $f(y)$ are nearly normal, by reference to Table I(a), and that, for each, $\sigma = 1$.

(b) Why are these marginal totals not more nearly normal than they are?

(c) By reference again to Table I(a), show that the distribution $f(0, y)$ is nearly normal, and that its standard deviation is about $\sqrt{1 - r^2}$. (This would be true of all the arrays in a complete table.)

(d) Compute r directly from the table.

(e) Draw the ellipse which goes through the frequencies equal to 2.2.

(f) Will the lines $y = \pm x$ be lines of symmetry in all cases, no matter what the value of r?

2. Show from the equations that, when $r = 0$, the normal surface is such that all the columns are alike and all the rows are alike.

3. Show that, when $r = 1$, the surface is a normal curve in a

plane parallel to the z-axis. (Use equation (2), turning the axes through an angle of 45°.)

4. Find approximately, as in Example 1, the value of r in Problem 8 of Chapter VIII (p. 146).

2. Non-Measurable Characters. There are several possible methods of proceeding when we wish to find the correlation between two series which are ordered but not measured. Of course it is necessary to make some assumptions, and, whatever assumptions we may make, they will involve, explicitly or implicitly, a definition of what is to be understood by correlation in such a case. Certainly the old product moment definition is not immediately available, because by the definitions of moments we must have measurements to deal with. Of the various methods that have been suggested we shall mention only two in this chapter, reserving further discussion of the subject for Part II, Chapter V.

Method I. (Median point method.) Normalize the frequency distributions given by the marginal totals, as in Chapter VI, § 3, so as to find the median points of each class interval. Call these points x/σ_x, y/σ_y, since, automatically, now the means are the origins and the standard deviations are the units; and then proceed as with measured series.[1] The formula is

$$r = \frac{1}{N}\sum_{x,y} \frac{x}{\sigma_x}\frac{y}{\sigma_y} f(x, y), \tag{5}$$

and this is to be computed outright, as it stands.

Method II. (Mean point method.) Proceed as by Method I, except that the mean points of the intervals, instead of the median points, are to be found. This method is more commonly used than the other one, and in this text it will be convenient to use it exclusively when N is small (less than 51), because in that case Table VI enables one to write down immediately the values of x/σ_x and y/σ_y. When N is large

[1] It should be noticed that the x of our tables has the unit σ and is therefore the same as the $\frac{x}{\sigma_x}$, or $\frac{y}{\sigma_y}$, in the notation of this chapter.

and the number of cell divisions is also large, the author prefers Method I, because in that case it is the simpler method, and the theoretical disadvantage is at most slight. But if the material is grouped into a small number of cells, the grouping errors will be so large that neither method is trustworthy. This is often the case in practice, and the reader is advised to study at least the discussion in Part II, Chapter V, before applying the theory extensively. Each method will now be illustrated.

Example 2. Correlate ability in the two studies, given the following groupings (Method I). The small numbers in circles are the products $\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)$. These are to be multiplied by the respective frequencies and added over the whole table.

2nd Study \ 1st Study	A	B	C	D	E	$f(y)$	Cum $\frac{f(y)}{N}$ to Medians	$\frac{y}{\sigma_y}$
A	8 (+2.86)	1 (+1.37)	1 (+.822)			10	.05	−1.65
B	4 (+1.30)	5 (+.65)	4 (+.01)	6 (−.69)	1 (−1.46)	20	.20	−.84
C		6 (+.10)	14 (−.09)	10 (−.31)		30	.45	−.13
D		5 (−.52)	13 (−.01)	8 (+.86)	4 (+1.19)	30	.75	+.68
E		2 (−1.37)	3 (−.58)	1 (+1.26)	3 (+2.89)	10	.95	+1.65
$f(x)$	12	20	35	25	8	100		
Cum $\frac{f(x)}{N}$ to Medians	.06	.22	.495	.795	.96			
$\frac{x}{\sigma_x}$	−1.55	−.77	−.012	+.824	+1.75			

$$\sum_{x,y} \frac{x}{\sigma_x}\frac{x}{\sigma_y} f(x,y) = +36.86,$$

$$r = 0.37.$$

Example 3. Find, as well as possible, the correlation between achievement in the first half of this course and in the second half from the rank lists of twelve students. Fractions indicate ties.

Student	A	B	C	D	E	F	G	H	I	J	K	L
1*st Half*	1.5	1.5	3	4	5	6.5	6.5	8	9	10	11	12
2*nd Half*	2	3	1	7	5	9.5	8	5	9.5	11.5	5	11.5

In obtaining the x's from Table VI we do not actually use fractional ranks (cf. Chapter VI, § 3); they are inserted here because they afford a handy means of indicating which scores were ties.

Student	A	B	C	D	E	F	G	H	I	J	K	L
x_1/σ_1	−1.50	−1.50	− .82	−.55	−.32	0	0	.32	.55	.82	1.16	1.84
x_2/σ_2	−1.16	− .82	−1.84	.11	−.32	.68	.32	−.32	.68	1.50	−.32	1.50
$x_1x_2/\sigma_1\sigma_2$	1.74	1.23	1.51	−.06	.10	0	0	−.10	.37	1.23	−.37	2.76

Each $f = 1$. So our formula is:

$$r = \frac{1}{N}\Sigma\frac{x_1}{\sigma_1}\frac{x_2}{\sigma_2} = \frac{1}{12}(+\ 8.41) = 0.70.$$

These methods assume that good scales by which the attributes in question might have been measured would effect normal distributions in the marginal totals. They do not assume that the correlation surfaces would be normal, for there exist correlation surfaces which are not normal, for which nevertheless the marginal totals are normal. We saw when dealing with ordered series that the foregoing assumption about the scales was a common and a reasonable one, but again it is evident that it can only be assumed to be approximately true. In drawing conclusions from the r as thus computed, one must remember that it is not a precise measure.

3. Partly Measured Characters. If one of two series is measurable and the other is not, we may make a combination of one of the methods of this chapter with the method of Chapter VIII. Suppose the X series is ordered, merely, and the Y series measured. First, normalize the X series so as to find either the median (Method I) or the mean (Method

II) points of the class intervals. Call these x/σ_x as before. For the Y series choose an arbitrary origin and unit, as in Chapter VIII, and denote the mid-points of the intervals by v. Then our formula for r is

$$r = \frac{1}{N}\sum_{x,y}\frac{x}{\sigma_x}\frac{y}{\sigma_y}f(x,y) = \frac{1}{N}\sum_{x,v}\frac{x}{\sigma_x}\frac{v-\bar{v}}{\sigma_v}f(x,v). \qquad (6)$$

The last expression is to be computed outright, as illustrated in the examples.

Example 4. Using Method I, find the correlation between the number of divorces granted (per 1000 population) and the laxity

v	Laws / Divorces	A	B	C	D	E	F	$f(y)$	$uf(y)$	v^2f	$\frac{y}{\sigma_y}=\frac{v-\bar{v}}{\sigma_v}$
−3	0.0 −	3 (3.66)						3	−9	27	−1.819
−2	0.5 −	3 (2.29)	6 (1.40)	7 (.77)	8 (.13)	3 (−.62)	6 (−1.55)	33	−66	132	−1.137
−1	1.0 −		8 (.56)	8 (.31)	9 (.06)	4 (−.25)	3 (−.62)	32	−32	32	−.455
0	1.5 −		1 (−.28)	3 (−.15)	14 (−.03)	8 (.12)	5 (.31)	31	0	0	.227
1	2.0 −			3 (−.61)	5 (−.10)	11 (.49)	4 (1.24)	23	23	23	.909
2	2.5 −		2 (−1.96)			7 (.86)	3 (2.17)	12	24	48	1.592
3	3.0 −		1 (−2.80)				1 (3.11)	2	6	18	2.274
4	3.5 −						2 (4.04)	2	8	32	2.956
	$f(x)$	6	18	21	36	33	24	138 (N)	−46	312	
Cum $f(x)$ to Medians		3	15	34.5	63.0	97.5	126				
Cum $\frac{f(x)}{N}$		.022	.109	.250	.456	.706	.914				
	$\frac{x}{\sigma_x}$	−2.014	−1.231	−.674	−.110	.542	1.366				

$$\bar{v} = \frac{-46}{138} = -\frac{1}{3},$$

$$\sigma_v^2 = \frac{312}{138} - \bar{v}^2$$

$$= 2.150,\ \sigma_v = 1.466.$$

of the divorce laws. The data are taken from the *World Almanac* of 1927, and relate to all the states in the United States, except Nevada and South Carolina (and District of Columbia), in which the laws are quite exceptional. The laws were strictest in group *A*, most lenient in group *F*. Obviously, one could not measure the strictness of these laws precisely. At most, one could make a number of classifications as stated. Therefore this, the *X* series, is ordered merely. Obviously, also, the number of divorces granted is a precise number and affords the basis for a completely measured series *Y*.

$$\sum_{x,y} \frac{x}{\sigma_x} \frac{v - \bar{v}}{\sigma_v} f = 52.25,$$

$$r = \frac{52.25}{N} = 0.38.$$

Had Method II been employed, the result would have been $r = 0.40$.

Example 5. Find the correlation between weight and health for the following (fictitious) data. Health was graded as *A*, *B*, *C*, or *D*, *i.e.*, good, fair, poor, sick. Weight is in pounds. $N = 16$.

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Health	*A*	*B*	*D*	*A*	*B*	*B*	*A*	*C*	*B*	*A*	*C*	*B*	*A*	*A*	*C*	*B*
Weight	127	156	120	145	131	142	144	205	136	142	142	130	117	145	137	140

The six *A*'s in health are ties, and by Table VI their mean x_R is

$$-\tfrac{1}{6}(1.968 + 1.326 + 1.013 + .778 + .580 + .403) = -1.011.$$

For the *B*'s, $x_R = +.164 = \frac{x}{\sigma_x}$; for the *C*'s, + 1.039; for *D*, + 1.97.

So far we have obtained the second row of the table on page 173. Let v = weight − 100 to get the third row.

4. Correlation between Ranks. Suppose a small group N of individuals are assigned the numbers 1 to N in two different ways. The correlation between the two sets of numbers or

COMPUTATION FOR EXAMPLE 5

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	*Sums*
x_R	−1.01	.16	1.97	−1.01	.16	.16	−1.01	1.04	.16	−1.01	1.04	1.6	−1.01	−1.01	1.04	.16	*About* 0
v	27	56	20	45	31	42	44	105	36	42	42	20	17	45	37	40	659
v^2	729	3136	400	2025	961	1764	1936	11025	1296	1764	1764	900	289	2025	1369	1600	31219
$v - \bar{v}$	−14.19	14.81	−21.19	3.81	−10.19	.81	2.81	63.81	−5.19	.81	.81	−11.19	−24.19	3.81	−4.19	−1.19	
$\frac{(v - \bar{v})}{\sigma_v}$	−.74	.78	−1.11	.20	−.54	.04	.15	3.34	−.27	.04	.04	−.59	−1.27	.20	−.22	−.06	
$\frac{x}{\sigma_x} \cdot \frac{v - \bar{v}}{\sigma_v}$	.75	.12	−2.18	−.20	−.09	.01	−.12	3.47	−.04	−.04	.04	−.09	1.28	−.20	−.23	−.01	2.47

$$\bar{v} = \frac{659}{16} = 41.19,\ \sigma_v^2 = \frac{32983}{16} - \bar{v}^2 = 365.0,\ \sigma_v = 19.1;\ r = \frac{2.47}{16} = +0.15.$$

ranks thus obtained can be found by the usual product moment formula,

$$r = \frac{1}{N}\sum_{X,Y}\left(\frac{X - \bar{X}}{\sigma_x}\right)\left(\frac{Y - \bar{Y}}{\sigma_y}\right),$$

but in this case Pearson has shown [1] that it can be put into a simpler form:

$$r = 1 - \frac{6\Sigma(X - Y)^2}{N(N^2 - 1)}.$$

Example 6. Five judges, T, U, B, L, H, ranked according to merit the same twelve answers to a given problem, with these results (*Kelley*):

Answers	*T*	*U*	*B*	*L*	*H*
A	5	7	10	2	5
B	6	4	6	3	9
C	3	1	4	1	2
D	2	2	11	8	3
E	12	3	1	4	10
F	1	8	2	5	1
G	11	10	8	12	4
H	9	5	7	6	11
I	4	9	12	7	6
J	7	11	5	9	8
K	10	12	9	10	12
L	8	6	3	11	7

Find the correlation between the ranks of T and the ranks of U.

T	5	6	3	2	12	1	11	9	4	7	10	8
U	7	4	1	2	3	8	10	5	9	11	12	6
$T - U$	−2	2	2	0	9	−7	1	4	−5	−4	−2	2
$(T - U)^2$	4	4	4	0	81	49	1	16	25	16	4	4

$$\Sigma(T - U)^2 = 208, \quad r = 1 - \frac{6 \cdot 208}{12 \cdot 143} = 0.2727.$$

[1] See Problem 5.

The student should note carefully the distinction between the procedures in Example 5 and Example 6, and the reasons therefor. In this last example we treat the ranks as if they were measurements. We are not interested in the achievements of the students who gave the answers; we are concerned only with the degree of uniformity with which the two judges assessed their achievements. That is, we really want here a coefficient which will express the degree of uniformity in the two sets of *scores*. In Example 5, we were trying to do something more difficult. We were thinking of the students' real success, and supposed that it was distributed normally, and that the rank numbers gave only their relative order. We were not satisfied then with a coefficient which would measure the degree of uniformity between the rank numbers. We wanted one which would measure the degree of uniformity between the hypothetically normally distributed numbers which were supposed to measure more closely the actual achievements. We should not use the method of Example 6 in Example 5, and it would be equally incorrect to use the method of Example 5 in Example 6.

PROBLEMS CHAPTER X

1. Use the second method of § 2 to find the correlations indicated (*Lovitt and Holtzclaw*).

(*a*) Intelligence and ability in mathematics.
(*b*) Intelligence and ability in English.
(*c*) Ability in mathematics and ability in English.

I	*M*	*E*	*I*	*M*	*E*
63	85	75	49	65	78
45	72	75	75	82	72
59	85	85	59	40	75
68	82	90	80	61	78
49	40	76	46	85	78
52	85	82	49	40	40
70	80	86	79	85	76
59	83	81	55	40	40
53	40	75	66	75	85
48	55	75	68	95	91
66	77	87	65	83	77
79	93	87	53	73	72
71	76	63	51	35	35
69	75	95	70	88	85
45	65	70	69	75	80

I = *Thorndike* Intelligence Score
M = Score in Freshman Mathematics
E = Score in Freshman English

2. Use the first method of § 2 to find the correlations below.[1]

(*a*) Chapter VIII, Problem 3, supposing the categories were ordered merely, instead of measured in feet and pounds.

(*b*) Chapter VIII, Problem 6, to be treated like Problem 3 in (*a*).

(*c*) Chapter VIII. Problem 7, to be treated in the same way.

(*d*) *Hair color and eye color of British schoolboys.* (*Biometrika*, vol. 3, p. 460.)

[1] One should note carefully what was said just before Example 2 relative to the untrustworthiness of these methods when the grouping errors may be large. Problems of this sort are inserted here as numerical exercises merely. Several of them will be solved later by more advanced methods.

Hair / Eye	Fair	Brown	Dark
Light	210	107	47
Medium	117	158	113
Dark	23	63	125

(*e*) *Health of two brothers* (*N* = 1918). (*Biometrika*, vol. 3, p. 166.)

1st Bro. / 2nd Bro.	Very Strong	Strong	Normally Healthy	Rather Delicate	Very Delicate
Very Strong	24	31	11.5	4	
Strong	31	342	163.75	65.75	3
Normally Healthy	11.5	163.75	588.5	137.25	6
Rather Delicate	4	65.75	137.25	95	11
Very Delicate		3	6	11	2

(The fractional frequencies occur because there were, in the original data, border-line cases.)

(*f*) *Curliness of hair of two sisters* (*N* = 1908). (*Biometrika*, vol. 3, p. 169.)

1st Sister / 2nd Sister	Smooth	Wavy	Curly
Smooth	937.5	190.5	98
Wavy	190.5	213.5	52
Curly	98	52	76

(*g*) *Handwriting and drawing, Swiss pupils* ($N = 1405$). (*Biometrika*, vol. 7, p. 225.)

H. W. / *D.*	FORT	MOYEN	FAIBLE
FORT	82	182	30
MOYEN	180	556	209
FAIBLE	25	76	65

3. Use the method of § 3 in the following cases. (*Biometrika*, vol. 3.)

(*a*) *Age and hair color. Males, Lower Elsass* ($N = 1912$).

Age / *Hair*	BLOND	BRAUN	SCHWARZ
0–15	384	74	
15–30	122	204	15
30–45	119	261	44
45–60	120	292	75
60–75	58	105	39

(*b*) *Age and hair color, British schoolgirls* ($N = 1305$).

Age / *Hair*	RED	FAIR	BROWN	DARK	JET BLACK
4– 7	2	16	6	3	
7–10	10	86	79	20	
10–13	16	165	160.5	73.5	4
13–16	19	136.5	189.5	91	3
16–19	6.5	71.5	90.5	55	1.5

4. Find the following rank correlations, using the data of Example 6:

$$r_{TB},\ r_{TL},\ r_{TH},\ r_{BU},\ r_{BL},\ r_{BH},\ r_{LH}.$$

5. Prove the formula of § 4, making use of certain formulae which will be proved in Part II:

$$\sum_1^N X = \frac{N(N+1)}{2},\quad \sum_1^N X^2 = \frac{N(N+1)(2N+1)}{6}.$$

HINT: Let $t = X - Y$, and express the usual formula, Chapter VIII (10 *b*), in terms of t and N.

6. Find the correlation between cancer mortality (relative to general mortality) and age. *Mortality Statistics, U. S. Census Bureau.*

Relative Mortality	.032	.083	.136	.153	.129	.074
Age	29.5	39.5	49.5	59.5	69.5	79.5

Cancer is known to be an old-age disease. Why is the correlation so small?

7. From equation (1) show that all the columns of a normal frequency distribution have the same standard deviation, and that it equals $\sigma_y\sqrt{1-r^2}$. Cf. also Exercise 1 *c*, § 1.

PART II: THE MATHEMATICAL PART OF ELEMENTARY STATISTICS

CHAPTER I

PROBABILITY

1. Preliminary Definition. Some notion of what is meant by probability may be obtained from a rather common definition: If an event can happen in m ways and either happen or fail in $(m + n)$ ways, its probability is $\frac{m}{(m + n)}$.

Thus, if we inquire what the probability is that a cubical die will fall so that a given face (say the ace) is uppermost, we think of one "way" in which this can happen, and of six ways in which it can either happen or fail; hence, the probability is $\frac{1}{6}$. But this definition is not accurate as it stands, for if the die were not perfectly cubical and homogeneous it is clear that the true probability would not be $\frac{1}{6}$, although by this definition it would seem to be. This definition is also vague. The reader might inquire what the probability is that he will die within the year. Just what is he to understand to be meant by the number of "ways" in which he may die, or the number of "ways" in which he may either die or live? To obviate these and other objections it is desirable to phrase a better definition of probability, and for that purpose some preliminary definitions of other terms are necessary.

2. Events. DEFINITION 1. THE TRIAL AND THE EVENT. When we speak of probability we shall always have two sets of circumstances or conditions in mind. One of these sets comprises the conditions which are supposed to have occurred by the hypothesis of the problem. For short, this set is to be called the "trial." The other set may or may not occur when

the trial does. The probability that it will occur is the thing that interests us. This set will be called the "event."

Examples. What is the probability of obtaining *exactly one head* (the event) in *one toss of a coin* (the trial)? What is the probability of obtaining *exactly one head* (the same event) in *one toss of two coins* (a different trial)? What is the probability that *a man aged thirty* will *die within a year*? In this last case, the trial is usually thought of as a random choice of one man from a large group, all aged thirty. The event is then the choice of one of those members of that group who will actually die within the year.

DEFINITION 2. MUTUALLY EXCLUSIVE. *Two or more events are mutually exclusive, with respect to a given trial, if only one can occur if the trial occurs but once.*

Examples. If the trial is the toss of a coin, and two events are the results, head and tail, these two events cannot both occur as the result of one toss, and therefore are mutually exclusive. If the trial is a single toss of two coins, these same [1] two events may both occur as the result of one toss, and are not mutually exclusive.

DEFINITION 3. EQUI-PROBABLE EVENTS. *Two events are equi-probable with respect to a given trial if they satisfy the following conditions. In N trials suppose the first event has happened x times, and the second y times. Then, as N becomes infinite, the ratio x/y shall approach unity as a limit.*

Example. If the trial is the toss of a single coin, the events, head and tail, are equi-probable, provided in the long run they will happen equally often, or, more precisely, provided the limit of the ratio, number of heads to number of tails, is one. Of course, no one really knows in the case of any special coin whether or not these conditions are truly satisfied, but if, for practical purposes, we are willing to assume that they are satisfied, then we may now say, more briefly, that we are willing to assume that these events are equi-probable. The theory that we shall build upon this hypothesis

[1] To make the events clearly identical in the two cases they should be described both times as exactly one head and exactly one tail.

will give in practice good or bad results, depending on whether or not this assumption is a valid one.[1]

EXERCISES § 1

1. Are E and E' mutually exclusive events in the following cases?

(*a*) *E:* to throw an ace with 1 die; *E':* to throw a deuce.

(*b*) *E:* to draw a black ball from a bag containing 3 black and 2 white balls; *E':* to draw a white ball.

(*c*) *E:* to throw 2 aces at 1 throw of 2 dice; *E':* to throw an ace and a deuce.

(*d*) *E:* to throw at least 1 ace on 1 throw of 2 dice; *E':* to throw at least 1 deuce.

(*e*) *E:* to draw at most 1 black ball on drawing 2 balls from a bag containing 3 black balls and 1 white ball; *E':* to draw at most 1 white ball.

2. Are E and E' in Exercise 1 equally likely?

DEFINITION 4. PROBABILITY. *With respect to a given trial let there be* $(m + n)$ *equi-probable and mutually exclusive events. The probability that one of the m events will happen as the result of this trial is* $p = \dfrac{m}{(m + n)}$.

Example. There are six equi-probable and mutually exclusive events which may happen as a result of a single toss of a cubical die: $m + n = 6$. One of these events is that the ace will be uppermost. So the probability of an ace is $\frac{1}{6}$. Two of these events are: the ace uppermost and the deuce uppermost. So the probability of either an ace or a deuce is $\frac{2}{6}$.

[1] There is always a discrepancy of this sort between theory and practice. The propositions of Euclidean geometry can be applied to nature only as approximations because there are no exactly straight lines in nature. The propositions of mechanics regarding the behavior of rigid bodies are only approximately true of physical bodies because such bodies are only approximately rigid. By various tests we can discover, again approximately, how nearly rigid any given body is, and also, by various tests, we can discover what the ratio x/y for two given events seems to be approaching as a limit.

COROLLARY. *If the trial is repeated N times, and u is the number of times one of the m events occurs, and v is the number of times one of the other n events occurs, then*

$$\lim_{N=\infty} \frac{u}{u+v} = \frac{m}{m+n}; \text{ i.e., } \lim_{N=\infty} \frac{u}{N} = p.$$

Proof. Let e_1 be the number of times in N trials that the first event happens; let e_2 be the number of times in N trials that the second event happens, etc.; similarly, let e_m be the number of times in N trials that the mth event happens; and finally let e_{n+m} be the number of times in N trials that the $(n+m)$th event happens.

By Definition 3

$$\lim \frac{e_1}{e_1} = \lim \frac{e_2}{e_1} = \cdots = \lim \frac{e_m}{e_1} = \cdots = \lim \frac{e_{n+m}}{e_1} = 1,$$

and therefore

$$\lim \frac{e_1 + \cdots + e_m}{e_1} = m, \lim \frac{e_1 + \cdots + e_{n+m}}{e_1} = n + m.$$

But

$$e_1 + \cdots + e_m = u, e_1 + \cdots + e_{n+m} = u + v,$$

and so

$$\lim \frac{\frac{u}{e_1}}{\frac{u+v}{e_1}} = \lim \frac{u}{u+v} = \frac{m}{n+m}.$$

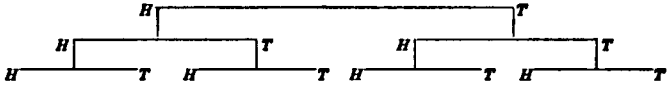
This corollary is the so-called limit definition of probability. It shows that we have really used this definition, although we have seemed to use limits only in defining equi-probable events. Elsewhere we have used instead the language of proportion:[1] $\frac{m}{(m+n)}$ is the proportion of favorable events in

[1] We shall not make use, in this text, of so-called continuous probability. Where continuous functions are to be used as probability functions, they will be thought of merely as convenient approximations to the discrete probabilities for which they are substituted.

the total group of $(m + n)$ events. But this group of events cannot be any group; it must be a group of equi-probable and mutually exclusive events, and when this is said, the limit idea is introduced.

Example 1. Ten balls, 3 red, 5 white, and 2 blue, are in a box. The proportion of reds is three-tenths. The probability of getting a red is also three-tenths, provided we carefully define our conditions. We must state the nature of the trial, that one ball only is to be drawn, and we must posit that all balls are equally likely to be drawn. If we do not thus define our conditions, the probability of a red may not be three-tenths. If, *e.g.*, the trial were so arranged that a red ball could not be obtained, the probability would be zero.

Example [1] 2. If a single toss of three coins is made, it is fairly obvious that all of the following events: head – head – head, head – head – tail, etc., are equi-probable and mutually exclusive. For convenience, the several events are arranged on a diagram like two family trees. Each line of descent is an event, the first being *HHH*.



Hence, the probability of 3 heads is $\frac{1}{8}$, of exactly 2 heads $\frac{3}{8}$, of exactly 1 head $\frac{3}{8}$, and of no heads $\frac{1}{8}$.

EXERCISES § 2

1. In Example 2 find the probability of (*a*) at least 2 heads; (*b*) at least 1 head; (*c*) at least 1 tail; (*d*) at most 2 heads; (*e*) as many heads as tails.

2. A single toss of 2 dice is made. Count up as in Example 2 the various equi-probable events, and find the probability of (*a*) ex-

[1] It is easy for a class of 20 to verify this. Each student should make 4 throws with 3 coins. The total number (80) of throws, when tabulated, will usually give results close to the theoretical values: 10, 30, 30, 10. This sort of demonstration is convincing to those students who feel that all these probabilities are equal.

actly 2 aces; (*b*) exactly 1 ace; (*c*) no aces; (*d*) exactly 1 ace and 1 deuce.

3. Elementary Theorems. DEFINITION 5. INDEPENDENT EVENTS. Events are mutually independent with respect to a given trial (1) if all of them can occur when the trial does, and (2) if the occurrence of any one cannot affect the probability of the occurrence of the others.

Example. If the trial is a single throw of a coin with each of one's hands, and if the events are, to obtain a head with the right hand, and to obtain a head with the left hand, these events are independent.

EXERCISE. Are mutually exclusive events independent?

Theorem I. (*The Product Theorem.*) *If, with respect to a given trial, two or more events are mutually independent, the probability that all of them will happen, as the result of a single trial, is the product of their several probabilities.*

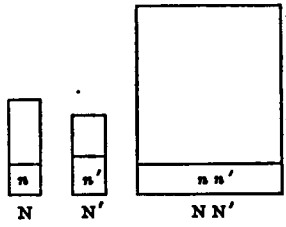
Theorem II. (*The Addition Theorem.*) *If, with respect to a given trial, two or more events are mutually exclusive, the probability that exactly one of them will happen, as the result of a given trial, is the sum of their separate probabilities.*

Proofs. These two theorems will be proved in the special case where the conditions may be accurately described by drawing balls from boxes. This is not always the situation, although it will be difficult for the student [1] to imagine it to be otherwise. Another way of describing this case is to say that we are supposing that the various given probabilities are rational fractions.

(*Theorem I.*) Let there be two boxes of balls. The first shall contain N balls, of which n are white. Let $p = \frac{n}{N}$. The second shall contain N' balls, of which n' are white. Let

[1] It would be necessary for him to know something of continuous probabilities — not defined in this text.

$p' = \frac{n'}{N'}$. A pair of balls is drawn, one from each box. The total number of different possible pairs is NN', of which nn' are pairs of which both balls are white. If all balls are equally likely, p is the probability that a ball chosen from the first box is white, p' the probability that a ball chosen from the second box is white, and $\frac{nn'}{NN'}$ is the probability that a pair is all white. These statements follow immediately from the definition of probability. Theorem I now follows from the fact that $pp' = \frac{nn'}{NN'}$.



(*Theorem II.*) Let a box contain N balls, of which n are red, n' are white, and n'' blue. Of course $n + n' + n'' = N$. Let $p = \frac{n}{N}$, $p' = \frac{n'}{N}$, $p'' = \frac{n''}{N}$. If all balls are equi-probable, these p's are the probabilities that, if one ball be drawn, it will be red, white, and blue, respectively. Likewise, from the definition of probability, the probability of obtaining either a red or a white is $\frac{(n + n')}{N}$, for this is equal to

$$\frac{\text{number of favorable events}}{\text{total number of events}}.$$

Therefore, since $p + p' = \frac{(n + n')}{N}$, we have shown that the probability of either a red or a white is the sum of the separate probabilities of red and white. The drawing of a red ball and the drawing of a white ball, if one ball only is drawn, are mutually exclusive events, and so the theorem is verified.

COROLLARY. *The probability that an event will either happen or fail on a given trial is unity. We give the name certainty to a probability which equals* 1.

Proof. By the definition of probability, the event in question is one of a group of exactly m equi-probable, mutually exclusive events, and there are $(m + n)$ such events which may happen as a result of the trial. The probability of success is $\frac{m}{(m+n)}$, and the probability of failure is $\frac{n}{(m+n)}$, and these two events are mutually exclusive; so the sum of their probabilities is the probability of either success or failure, and this sum is 1.

Example 3. In Example 2 find the probabilities of the following results:

(*a*) *At least* 2 *heads. Ans.*, $\frac{1}{8} + \frac{3}{8} = \frac{1}{2}$, for the events, exactly 2 heads, and 3 heads, are mutually exclusive, and it is required to find the probability that either the one or the other will occur.

(*b*) *At most* 2 *heads. Ans.*, $\frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{7}{8}$, for it is required to find the probability that one of the following events will occur: exactly 2 heads, exactly 1 head, or no heads.

(*c*) *Three heads. Ans.*, $\frac{1}{8}$. Let us show that this answer would also be obtained if one threw the coins one at a time. The probability that the first would be a head is $\frac{1}{2}$, that the second would be a head is $\frac{1}{2}$, and likewise $\frac{1}{2}$ is the probability that the third would be a head. These three events are independent, and by Theorem I the probability that all will happen is the product $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

Example 4. In two throws of three coins each, find the probabilities of the following results:

(*a*) 6 heads;

(*b*) 3 heads on the first throw, and exactly 2 on the second;

(*c*) exactly 2 heads on the first throw, and 3 on the second;

(*d*) exactly 2 heads on one of the throws, and 3 on the other.

Ans., (*a*) $\frac{1}{64}$; (*b*) $\frac{1}{8} \cdot \frac{3}{8} = \frac{3}{64}$; (*c*) $\frac{3}{8} \cdot \frac{1}{8} = \frac{3}{64}$; (*d*) $\frac{3}{64} + \frac{3}{64} = \frac{3}{32}$.

EXERCISES § 3

1. Are E and E' dependent in the following cases?

(*a*) *E:* to throw 3 heads with 1 coin; *E′:* then to throw 1 tail.

(*b*) *E:* to throw an ace with 1 die; *E′:* to throw an ace with another.

(*c*) *E:* to draw a white ball from a bag containing 10 white and 9 red balls; *E':* to draw a white ball from those that remain after the first draw.

(*d*) *E* as in (*c*); *E':* to draw a white ball after the first ball drawn has been replaced.

(*e*) *E: A*, now age 25, lives to age 60; *E': B*, now age 20, lives to age 60.

(*f*) *E: A*, now age 25, dies at 60; *E': B*, now age 20, lives 5 years after *A*'s death.

2. Find the several probabilities that both *E* and *E'* will happen in Exercise 1 *a*, *b*, *d*. *Ans.*, $\frac{1}{19}$, $\frac{1}{18}$, $\frac{1}{361}$.

3. Find the probability that both *E* and *E'* will happen in Exercise 1 *c*. (*Hint:* Restate this problem so as to use only independent events.) *Ans.*, .263.

4. In tossing a coin three times, what is the probability:

(*a*) of 2 heads and then 1 tail?

(*b*) of 1 head and then 2 tails?

(*c*) of 1 head, 1 tail, and then 1 head?

(*d*) of exactly 2 heads and 1 tail (in any order)?

5. Show that the answers to Exercise 4 are the same if 3 coins are thrown simultaneously, the coins having been marked 1, 2, 3.

4. Permutations and Combinations. If, in the last example, we had tried to find the probability of getting exactly 5 heads, we should have been obliged to consider, not only the two methods of obtaining 5 heads indicated in (*b*) and (*c*), but also all the other possible methods. Counting up the various possibilities in a problem is frequently a complicated matter, but there are some general theorems and formulae which are often helpful. We first distinguish between two different kinds of things to be counted. The first is called a permutation. A permutation is an order, or arrangement. An illustration is the problem of finding the number of permutations in which 6 volumes may be arranged on a shelf. The second is called a combination. This is simply a group, in which the order is immaterial. I might select from the 6 volumes groups of 5 each. There would be

6 such groups or combinations, for an easy way to select t 5 is simply to eliminate one and take the rest, and there a 6 books which can be eliminated in turn, leaving 6 differe groups of 5 each. Note that these groups are not tota different, the one from the other. Some of the books th appear in one group will also appear in the other, but th are different in part. No two are exactly alike. The syml for the number of combinations of n things taken r at a ti is ${}_nC_r$. We have just seen then that

$$ {}_6C_5 = 6. $$

After selecting each of these 6 groups of 5 books each, might wish to arrange them on a shelf in every possible or or permutation, and to count up the total number of th permutations for all 6 groups. This would be called number of permutations of 6 things chosen 5 at a time, would be denoted by ${}_6P_5$, ${}_nP_r$ being the symbol for the n ber of permutations of n things taken r at a time. We r prove certain theorems.

Theorem III. ${}_nP_r = {}_nC_r \cdot {}_rP_r$.

Proof. This follows immediately from the definitions, it is obvious that one way of obtaining the total numbe permutations (${}_nP_r$) is, first, to select all possible groups (and then to arrange each group in all possible orders (${}_r$

Theorem IV. ${}_nP_r = n(n-1)(n-2)\cdots(n-r+$

COROLLARY 1. ${}_rP_r = r(r-1)(r-2)\cdots 2\cdot 1$.

COROLLARY 2. $${}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+}{r(r-1)(r-2)\cdots 1}$$

Proof of the Theorem. Suppose the objects are n le and that any permutation of letters is a word. The pro is to find out how many words of r letters each can be from these n letters. No letter is to be used twice i

the number of one-letter words

can be made is n. To find the number of two-letter words, we may take each of these one-letter words and adjoin to it any one of the given letters, except that one which we have already used, for the same letter must not be repeated in the same word. With *each* of the one-letter words we can therefore form $(n - 1)$ two-letter words. That means $n(n - 1)$ two-letter words in all. To find the number of three-letter words, we may now take *each* of these two-letter words and adjoin any letter except one of those two already used; that is, we may adjoin any one of $(n - 2)$ letters, thus making $n(n - 1)(n - 2)$ three-letter words. This process can evidently be extended to r-letter words, and we shall obtain for the number of such words: $n(n - 1) \cdots$, to r factors, proving the theorem.

Corollary 1 is obtained by putting $n = r$ in the theorem. Corollary 2 is obtained by solving Theorem III for ${}_nC_r$ thus:

$$ {}_nC_r = \frac{{}_nP_r}{{}_rP_r}, $$

and then substituting the formulae of the theorem and of Corollary 1.

DEFINITION. FACTORIALS. *If n is a positive integer, the product of the n factors, $n(n - 1)(n - 2) \cdots (1)$, is called factorial n and is written, $\lfloor n$, or $n!$* This definition does not apply to the case where $n = 0$, and so we may define $0!$ in any way that suits our convenience. It is convenient to let $0! = 1$.

COROLLARY[1] 3.

$$ {}_rP_r = r!; \quad {}_nP_r = \frac{n!}{(n - r)!}; \quad {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n - r)!}. $$

COROLLARY 4. ${}_nC_r = {}_nC_{n-r}$.

This follows immediately from the last equation of Corollary 3.

[1] To be proved in Problem 4.

COROLLARY 5. *The binomial theorem may be written:*

$$(a+b)^n = a^n + {}_nC_1 a^{n-1}b + {}_nC_2 a^{n-2}b^2 + \cdots + {}_nC_r a^{n-r}b^r + \cdots + b^n,$$

or, since $1 = {}_nC_0$,

$$(a+b)^n = \sum_{r=0}^{n} {}_nC_r\, a^{n-r}b^r.$$

Example. Six books are on a shelf. (*a*) Find the number of different ways in which groups of 4 may be selected.

$$Ans.,\ {}_6C_4 = \frac{{}_6P_4}{4!} = \frac{6\cdot 5\cdot 4\cdot 3}{1\cdot 2\cdot 3\cdot 4} = 15.$$

(*b*) In how many different ways may 4 books be taken from the 6 and arranged on another shelf? *Ans.*, ${}_6P_4 = 360$.

(*c*) Show that the number of ways in which groups of 4 of these books may be selected equals the number of ways in which groups of 2 may be selected. This follows from the fact that whenever a group of 4 is taken, a group of 2 is left. It follows also from Corollary 4: ${}_6C_4 = {}_6C_2$.

(*d*) Answer the question in (*b*) if Volume 1 must always be taken and placed in the first position on the other shelf. Like many questions, this cannot be answered simply by applying a formula; but, if Volume 1 and the first place are always together, we may as well say that we have only 5 other books to choose from and only 3 other places to put them in, and so the answer is given by ${}_5P_3 = 5\cdot 4\cdot 3 = 60$.

Example 5. How many committees of 5 Republicans and 3 Democrats can be formed from a senate of 63 Republicans and 23 Democrats?

Select the Republican members first: there will be ${}_{63}C_5 = 7{,}028{,}847$ possible different groups. Then choose the Democratic members: there are ${}_{23}C_3 = 1771$ possible different groups. Now put these Republicans and Democrats together in as many ways as possible and count up the total number of composite groups. With each one of the 7,028,847 groups of Republicans, each one of the 1771 groups of Democrats may be combined, and so the total number of composite groups or committees is the product of these two numbers:

$${}_{63}C_5 \cdot {}_{23}C_3 = 12{,}448{,}088{,}037.$$

Example 6. Find the number of combinations of 1000 things taken 998 at a time.

Ans., $\frac{1000!}{998!\,2!}$, but this is difficult to compute outright.

However, by Corollary 4, we know that ${}_{1000}C_{998} = {}_{1000}C_2$, and this is

$$\frac{1000 \cdot 999}{1 \cdot 2} = 499{,}500.$$

The last two examples suggest the desirability of finding a shorter method of computing high factorials than by multiplying together all the factors. If $n \leq 500$, $\log n!$ is given in Table V. If $n > 500$, a very nearly correct value may be obtained from Stirling's formula:

$$\text{Approximately, } n! = e^{-n}n^n\sqrt{2\pi n}; \tag{1}$$

and so, approximately,

$$\log_{10} n! = -n(0.43429\ 44819) + (n + \tfrac{1}{2})\log_{10} n + 0.399090. \tag{1a}$$

The proof of this formula is beyond the scope of this book, but we shall see it verified in particular cases. The reason the value of $\log_{10} e$ is given to so many places of decimals is that it is multiplied by n and n may be a large number, and a smaller number of places might have resulted in an inaccurate product. Suppose we had written 0.4342945 and that $n = 10{,}000$; then, apparently, $n \log e = 4342.945$, which, if 4-place logarithms are to be used, would have to be interpreted as 4342.9450, but the true value is 4342.9448. Similarly, in the second term, $(n + \frac{1}{2})\log n$, of formula (1*a*), if n is very large, $\log n$ should be found to a large number of decimal places.

Example 7. Compute approximately ${}_{1000}P_{800}$.

By (1*a*), $\log 1000! = -434.2945 + (1000.5)(3) + 0.3991 = 2567.6046$.

By Table V, $\log 200! = 386.4343$. So $\log {}_{1000}P_{800} = 2181.1703$, and ${}_{1000}P_{800} = 1.480 \times 10^{2181}$.

Example 8. Thirteen cards are drawn from a pack of playing cards. What is the probability that exactly 7 hearts are included? To get the number of favorable cases, we first select 7 hearts. The number of ways in which this can be done is ${}_{13}C_7$. Then we select from the remaining 39 cards a group of 6. This may be done in ${}_{39}C_6$ ways. The total number of favorable cases is then ${}_{13}C_7 \cdot {}_{39}C_6$. The total number of possible cases is ${}_{52}C_{13}$ and so the probability is

$$p = \frac{\lfloor\underline{13}}{\lfloor\underline{7}\ \lfloor\underline{6}} \frac{\lfloor\underline{39}}{\lfloor\underline{6}\ \lfloor\underline{33}} \frac{\lfloor\underline{13}\ \lfloor\underline{39}}{\lfloor\underline{52}} = \frac{(\lfloor\underline{13})^2\ (\lfloor\underline{39})^2}{(\lfloor\underline{6})^2\ \lfloor\underline{7}\ \lfloor\underline{33}\ \lfloor\underline{52}}.$$

By the tables:

$$\begin{aligned}
2 \log \lfloor\underline{13} &= 19.5886 \\
2 \log \lfloor\underline{39} &= 92.6192 \\
-2 \log \lfloor\underline{6} &= -\ 5.7146 \\
-\log \lfloor\underline{7} &= -\ 3.7024 \\
-\log \lfloor\underline{33} &= -\ 36.9387 \\
-\log \lfloor\underline{52} &= -\ 67.9066 \\
\hline
\log p &= 7.9455 - 10 \\
p &= 0.00882.
\end{aligned}$$

It is better, when using a machine which subtracts easily, to subtract the logarithms of the numbers in the denominator, rather than to add their cologarithms, in finding $\log p$.

EXERCISES § 4

1. Three balls are selected from a bag containing 5 black and 3 white balls. In how many ways may 3 black balls be chosen?

2. Four balls are selected from the same bag. In how many ways may 2 black balls and 2 white balls be chosen?

3. Given 5 different colored flags to choose from, how many different signals can be made with 5 different colored flags in line? with 4? with 3? with any number under 6? *Ans.*, 120, 120, 60, 325.

4. How many straight lines are determined by 10 points, no 3 of which are in the same straight line? *Ans.*, 45.

5. How many planes are determined by 10 points, no 4 of which are in the same plane? *Ans.*, 120.

6. How many handshakes could be exchanged among 7 persons, each greeting the other once? *Ans.*, 21.

7. How many groups of 13 cards can be selected from a pack of 52?

8. How many committees of 5 Republicans and 3 Democrats can be formed from a Senate of 23 Republicans and 63 Democrats?

5. The Point Binomial. Theorem V. *Let p be the probability of an event e in one trial, q the probability of failure. The successive terms of the binomial expansion,*

$$(p+q)^n = p^n + {}_nC_1 p^{n-1}q + \cdots + {}_nC_t p^{n-t}q^t + \cdots + q^n, \quad (2)$$

give the respective probabilities that, in n trials, this event will occur exactly $n, n-1, \cdots, n-t, \cdots, 0$ times.

The expansion (2) is called the point binomial. It was studied by J. Bernoulli,[1] and is often referred to as the series of Bernoulli.

Proof. Let us put $s = n - t$ so that the general term is

$${}_nC_t\, p^s q^t.$$

The problem is, essentially, to prove that the value of this term is the probability that, in n trials, the event e will happen exactly s times; and therefore fail to happen exactly t times. Consider the following sequences (E) of events which can happen as a result of n trials, and notice that these sequences are themselves events which are mutually exclusive with respect to any given set of n trials:

E_1 is the event: e happens on each of the first s trials and fails on the other t;

E_2 is the event: e happens on each of another group of s trials, and fails on the other t;

E_3 is the event: e happens on each of a third group of s trials, and fails on the other t; etc.

[1] His research was written in Latin and was published in 1713, after his death, under the title *Ars Conjectandi.*

How many E events of this sort can there be? Obviously as many as there are different groups of s trials to be used. This number is ${}_nC_s = {}_nC_t$. Let P_1 denote the probability of E_1, P_2 the probability of E_2, etc. By Theorem I,

$$P_1 = p^s q^t. \quad \text{Also } P_1 = P_2 = P_3, \text{ etc.} \tag{3}$$

Now what we are seeking ultimately is the probability that in a set of n trials *one* of the mutually exclusive events E_1, E_2, etc., will occur. By Theorem II, this is given by the sum of their several probabilities, $P_1 + P_2 + \cdots$. By (3) these several probabilities are all equal and their number has been found to be ${}_nC_t$. Therefore

$$P_1 + P_2 + \cdots = {}_nC_t\, p^s q^t.$$

COROLLARY. (*a*) *The probability that e will happen at least r times in n trials is the sum of all those terms of the series in which the exponent of p,* $s \geq r$.

(*b*) *The probability that e will happen at most r times is the sum of all those terms in which the exponent of p,* $s \leq r$.

(*c*) *The probability that e will happen at least once is* $1 - q^n$.

For by (*a*) it is the sum of all those terms of the series except the last, q^n. But the sum of all the terms of this series must be 1 because it equals $(p + q)^n$, and $p + q = 1$. So $1 - q^n$ does represent the sum of all the terms of the series except the last.

Example 9. If a coin is tossed 3 times, what are the probabilities of exactly (*a*) 3 heads, (*b*) 2 heads, (*c*) 1 head, (*d*) no heads? The results must turn out the same as in Example 2 (page 187), for it does not make any difference whether one coin is tossed thrice or three coins tossed once; we shall now arrive at those results by way of a series of Bernoulli. For this case the series is:

$$\begin{aligned}(\tfrac{1}{2} + \tfrac{1}{2})^3 &= (\tfrac{1}{2})^3 + 3(\tfrac{1}{2})^2(\tfrac{1}{2}) + 3(\tfrac{1}{2})(\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 \\ &= \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8}.\end{aligned}$$

These four fractions are the answers to (*a*), (*b*), (*c*), and (*d*), respectively.

6.[1] **The Finite Hypergeometric Series. Theorem VI.** *If there are m balls in a bag of which pm are white and qm are black, and if $n \leq m$ are drawn, without replacements, then the successive terms of the series,*

$$\frac{1}{{}_mC_n}\Big[{}_{pm}C_n\,{}_{qm}C_0 + {}_{pm}C_{n-1}\,{}_{qm}C_1 + \cdots + {}_{pm}C_{n-t}\,{}_{qm}C_t + \cdots + {}_{qm}C_n\Big],$$

give the respective probabilities that there will be, in the samples drawn, exactly $n, n-1, \cdots, n-t, \cdots, 0$ white balls.

Proof. As before, set $n - t = s$, and we need to show merely that $\frac{{}_{pm}C_s\,{}_{qm}C_t}{{}_mC_n}$ is the probability of exactly s white balls. The reasoning is exactly like that of Example 5. The number of possible different groups of s balls that can be obtained from the pm white balls is ${}_{pm}C_s$. The number of groups of t black balls that can be obtained from qm blacks is ${}_{qm}C_t$. The number of composite groups is the product of these two numbers, and this is the numerator of our fraction. The denominator is the total number of groups of n things that can be obtained.

COROLLARY. *The sum,*

$$\frac{1}{{}_mC_n}\sum_{t=0}^{n} {}_{pm}C_s\,{}_{qm}C_t = 1.$$

EXERCISES §§ 5–6

1. A coin is tossed 6 times. Find the probability of at least 2 heads, of at most 2 heads, of exactly 4 heads, of either 4 heads or 4 tails, of at least 4 heads or 4 tails. *Ans.*, $\frac{57}{64}$, $\frac{22}{64}$, $\frac{15}{64}$, $\frac{32}{64}$, $\frac{22}{64}$.

2. A bag contains 6 black balls and 4 white balls. After each drawing, the ball drawn is replaced. Find the probability of getting in 7 drawings exactly 4 black and 3 white balls; of 6 black balls and 1 white ball. *Ans.*, .29, .13.

[1] This section is of considerable interest and very important in certain applications, but it is less important than §§ 1–5 and may be omitted if desired.

3. In Exercises 1 and 2, what are the most likely results? *Ans.*, 3 heads and 3 tails; 4 black and 3 white.

4. Three balls are drawn, all at once, from a bag containing 4 black and 4 white balls. What is the probability that all are black? What is the probability that all are of one color?

5. In Exercise 4, what is the probability that: (*a*) at least 1 is black? (*b*) 2 are black and 1 is white? (*c*) at least 2 are of one color? (*d*) at most 2 are of one color?

6. A coin is tossed 6 times. If it comes down heads at least 3 times, the player is to receive \$10. What is the value of his expectation? (If one is to receive a certain sum of money in case a certain event takes place, the value of his expectation is defined as the product of that sum times the probability of the event.) *Ans.*, \$6.56.

7. A coin is tossed 6 times. If the player is to receive \$10 for every succession of 3 heads or 3 tails, what is the value of his expectation? (A succession of 4 heads is to be interpreted as two successions of 3 heads, etc.) *Ans.*, \$10.

8. What is the probability of exactly 4 aces in 5 throws with 1 die? *Ans.*, .00322.

9. What is the probability of exactly 4 aces in 5 throws with 2 dice? *Ans.*, .054.

10. What is the probability that the balls will be alternately of different colors in Exercise 2? *Ans.*, .0138.

11. Thirteen cards are drawn from a pack of 52. What is the probability of exactly 2 aces? *Ans.*, .2135.

12. In Exercise 11, what is the probability of at least 2 aces?

PROBLEMS CHAPTER I

1. How many semaphore signals can be made with 2 similar flags, if each may be used in any one of 8 positions, and the use of 2 flags in the same position does not count as a signal? *Ans.*, 28.

2. Prove that ${}_nC_r + {}_nC_{r-1} = {}_{n+1}C_r$. Illustrate.

3. Four constitute a quorum of an executive board of 25 of a certain chamber of commerce, provided the president and secre-

tary are 2 of the 4. How many different quorums of 4 can be formed? *Ans.*, 253.

4. Prove Theorem IV, Corollary 3.

5. Use Corollary 5 of Theorem IV to show that the total number of different combinations of n things, taken any number at a time, from 1 to n inclusive, is $2^n - 1$.

6. How many automobile registration numbers may be made by the use of any 5 of the 10 digits, if repetitions are not allowed? *Ans.*, 27,216.

7. How many may be made with any number of digits less than 6, if repetitions are not allowed? *Ans.*, 32,490.

8. How many, in the preceding problem, if repetitions are allowed?

9. There are 50 possible questions which may be asked in a certain subject, of which a student knows the answers to exactly 30. A two-hour examination involves 10 questions. In how many ways may a paper be made out which will give this student a grade of 100%? of 90%? of 60%? of 10%? (Each answer is to be regarded as either wholly correct or wholly incorrect.)

10. Show that the number of distinguishable orders in which 7 keys can be arranged on a ring is 6!.

11. Find the greatest value of r such that $n - r + 1 \geqq r$, n being fixed. Why is this the value of r that makes ${}_nC_r$ a maximum?

12. How many dominoes in a set: (*a*) from double 0 to double 6? (*b*) from double 0 to double 9? (*c*) from double 0 to double 12? *Ans.*, 28, 55, 91.

13. (*a*) A coin has been tossed twice with the result, 2 heads. Are heads and tails equally likely on the next throw? (*b*) If heads have come up 1000 times in succession, are heads and tails equally likely on the next throw?

14. Are E and E' equally likely if 2 dice are thrown, where E is the event, the sum of the numbers shown is 2; and E' is the event, the sum is 3?

15. If there are 99,999 registration numbers, what is the chance of meeting a car on which the same digit occurs more than once? *Ans.*, 0.675.

16. In Problem 9 what is the probability of exactly 60%? What is the probability of either 50%, 60%, or 70%?

17. *A*, *B*, *C*, and *D*, in the order named, throw a die. The one who gets an ace first gets \$1. Find the value of the expectation of each, assuming the game to continue indefinitely. (Cf. Exercise 6, § 6.)

18. If the probability of hitting a target is $\frac{1}{3}$, find the probability of no hits in 3 shots; of at least 1 hit in 3 shots.

19. If the probability of 2 successive shots both being hits is .9, what is the probability of a hit in 1 shot?

20. If an event has a probability as great as .99, artillerymen say that it is "morally certain." In Problem 18, how many shots are necessary to make at least one hit morally certain?

21. To choose partners for a game of tennis, four racquets are thrown, the two smooths and the two roughs to indicate partners. What is the probability of a choice on the first throw? on one of the first two throws?

22. If one player says he will "go with the odd," and only the racquets of the other three players are thrown, what are the corresponding probabilities?

23. If one player says he will go with the odd on the second throw in case there should be no choice on the first throw, what is the probability of a choice on one of the first two throws?

24. If the probability that *A* will solve a problem is $\frac{1}{3}$ and the probability that *B* will solve it is $\frac{1}{4}$, what is the probability that it will be solved if both try it?

25. *A* and *B* play a game with a die. If it comes up ace the first time, *A* gives *B* \$1. If it comes up 2 the second time, *A* gives *B* \$2, etc., the amount increasing by \$1 each time. The game is to stop whenever the die does not come up favorably to *B*, or in any event after 6 throws. What is the value of *B*'s expectation? (Cf. Exercise 6, § 6.)

26. *A* and *B* toss a die at most 6 times. If it comes up 1 the first time, *A* gives *B* \$1, and the game ends. If it comes up 2 the second time, *B* gives *A* \$2, and the game ends, etc., the amount

increasing by \$1 each time. Find the values of the expectations of each.

27. In one throw with a pair of dice what sum is most likely? What are the odds against it?

28. If the probability of hitting a target is $\frac{1}{3}$, find the probability of at least 10 hits in 13 shots.

29. If 20 hits are expected out of every 300 shots, find the probability of exactly 5 hits in 10 shots.

30. In Problem 29, how many shots are necessary to make it morally certain that the target will be hit at least twice?

31. Alter Problem 17 so that 2 dice are thrown, and the player who first throws as much as 10 on a single throw gets \$10, and then the game is over. What is the value of the expectation of each?

32. How many drawings with replacements from a bag containing 20 black and 5 white balls are necessary to make: (*a*) the probability of getting at least 1 black .99? (*b*) the probability of getting at least 2 blacks .99?

33. If the probability of each justice on the supreme bench rendering a correct decision is .9, and there are 9 justices, all of whom must vote, what is the probability of getting a correct decision by at least a majority vote?

34. (*a*) Draw a graph illustrating the various probabilities of all possible results when a coin is tossed 9 times.

(*b*) Plot on the same diagram the curve:

$$y = \phi(x),\ x = t - \tfrac{9}{2},\ t = \text{number of tails.}$$

35. (*Fry*) "The letters of the word *tailor* are written on cards. The cards having been thoroughly shuffled, four are drawn in order (without replacements). What is the probability that the result is *oral*? *Ans.*, $\frac{1}{210}$."

36. (*Fry*) Same for the words *pepper* and *peep*. *Ans.*, $\frac{1}{70}$.

37. (*Fry*) "A batch of 1000 (electric) lamps is five per cent bad. If five are tested, what is the chance that no defectives will appear? What is the chance that the test batch will be (at least) forty per cent defective?"

38. From an urn containing 8 white and 6 black balls, 7 are drawn and placed in a second urn. From the second urn 4 are drawn. What is the probability that 2 are white and 2 are black?

39. (*Bertrand's* "Box Paradox.") Three boxes have in them 2 coins each. In one box both coins are gold, in one both are silver, in the other they are mixed. Outside, the boxes are of identical appearance. A man chooses a box and takes out a coin which proves to be gold. What is the chance that the other coin in his box is also gold? *Ans.*, $\frac{2}{3}$.

CHAPTER II

APPROXIMATIONS TO THE POINT BINOMIAL

1. Properties. Before considering the main problem of this chapter it will be advantageous to note some of the properties of the point binomial. We shall give below its mean and some of its higher moments. By the moments of a point binomial we mean the uncorrected moments of the histogram which represents it. In this case we want the uncorrected moments, for these are found on the assumption that the entire frequency or probability represented graphically by each rectangle is concentrated at the middle of the rectangle, and this is the actual fact in the point binomial. The general term of the point binomial represents the relative frequency or probability of exactly s successes, not the relative frequency of a multitude of numbers varying from $s - \frac{1}{2}$ to $s + \frac{1}{2}$, as was the case with ordinary frequency distributions. Sometimes authors use the term "loaded ordinates" to apply to frequency rectangles when, as here, the entire frequencies are supposed loaded on the ordinates at the mid-points.

Theorem I. *Let the unit of measurement be the same as the unit of t.*

(*a*) *The mean of $(p + q)^n$ is, relative to the origin of t, $\bar{t} = nq$. Relative to the origin of s, the mean is $\bar{s} = np$. More briefly, the mean number of failures is nq, and the mean number of successes is np, in a point binomial distribution.*

(*b*) *The standard deviation is $\sigma = \sqrt{pqn}$.*

(*c*) $\alpha_3 = \dfrac{p - q}{\sigma}$, *the positive direction being the direction of increasing t;* $\alpha_4 = 3 + \dfrac{1}{\sigma^2} - \dfrac{6}{n}$.

(*d*) *The mode, i.e., the greatest term, is within an interval extending to a distance of* 1 *on either side of the mean, i.e.,*

$$|\textit{mean} - \textit{mode}| \leqq 1.$$

Proof of (*a*). Relative to the origin of t,

$$\begin{aligned}\bar{t} &= \sum_{t=0}^{n} t \cdot {}_nC_t p^s q^t = nq \sum_{t=1}^{n} t \frac{{}_nC_t}{n} p^s q^{t-1} \\ &= nq \sum_{t=1}^{n} \left(\frac{t}{n} \cdot \frac{n(n-1)\cdots(n-t+1)}{t(t-1)\cdots(1)}\right) p^s q^{t-1} \\ &= nq \sum_{t=1}^{n} {}_{n-1}C_{t-1} p^s q^{t-1} \\ &= nq(p+q)^{n-1} = nq.\end{aligned}$$

Proof of (*b*). Relative to the origin of t, the second moment,

$$\begin{aligned}\nu_2 &= \sum_{t=0}^{n} t^2 {}_nC_t p^s q^t = \sum_{t=0}^{n} [t(t-1) + t] {}_nC_t p^s q^t \\ &= \sum_{t=2}^{n} t(t-1) {}_nC_t p^s q^t + \sum_{t=0}^{n} t {}_nC_t p^s q^t \\ &= n(n-1)q^2 \sum_{t=2}^{n} \left[\frac{t(t-1)}{n(n-1)} \frac{n(n-1)(n-2)\cdots(n-t+1)}{t(t-1)(t-2)\cdots(1)}\right] \\ &\qquad \left[p^s q^{t-2}\right] + \bar{t}, \text{ by } (a), \\ &= n(n-1)q^2 \sum_{t=2}^{n} {}_{n-2}C_{t-2} p^s q^{t-2} + nq \\ &= n(n-1)q^2(p+q)^{n-2} + nq = n(n-1)q^2 + nq \\ &\qquad = n^2q^2 - nq^2 + nq.\end{aligned}$$

Since $\sigma^2 = \mu_2 = \nu_2 - \bar{t}^2$, we now have

$$\sigma^2 = n^2q^2 - nq^2 + nq - n^2q^2 = nq(1-q) = nqp.$$

Proof of (*c*). (See Problem 21.) It is first necessary to obtain ν_3, μ_3, and μ_4. These quantities will be found to have the following values:

$$\left.\begin{aligned}\nu_3 &= nq(n^2q^2 - 3nq^2 + 2q^2 + 3nq - 3q + 1), \\ \mu_3 &= npq(p-q), \\ \mu_4 &= npq[1 + 3(n-2)pq].\end{aligned}\right\} \quad (1)$$

Proof of (*d*). (Method of *Laplace*.) We wish to find the greatest term, or the two greatest terms, in the expansion of $(p + q)^n$. Let T be one of them.

Write $T = {}_nC_t p^t q^t$. Now our problem is to find the required value of t. The term preceding T is $T \frac{p}{q} \frac{t}{n - t + 1}$, and the term succeeding T is $T \frac{q}{p} \frac{n - t}{t + 1}$, as will be immediately obvious if the coefficients of these terms, such as ${}_nC_t$, be written out in full. By the hypothesis that T is one of the greatest terms,

$$\left.\begin{aligned} T &\geqq T \frac{p}{q} \frac{t}{n - t + 1}, \\ T &\geqq T \frac{q}{p} \frac{n - t}{t + 1}. \end{aligned}\right\} \tag{2}$$

Now solve each of these inequalities for t. Reducing the first, we have

$$q(n - t + 1) \geqq pt, \; q(n + 1) \geqq (p + q)t,$$

and so

$$qn + q \geqq t. \tag{3}$$

From the second of the inequalities of (2) it follows that

$$p(t + 1) \geqq q(n - t), \; (p + q)t \geqq qn - p,$$
$$t \geqq qn - p. \tag{4}$$

Putting (3) and (4) together, we learn that

$$qn - p \leqq t \leqq qn + q, \tag{5}$$

that is, that t differs from qn by at most p or q, whichever is greater. But both p and q are less than one, save in the trivial case when p or $q = 1$; and therefore, in every case, t differs from qn by at most 1.

The case where there are two equal terms, both greater than any other, occurs when $q(n + 1)$ is an integer. For then the next lower integer is $qn + q - 1 = qn - p$. In this case there are two values of t which satisfy (5). One

is $qn - p$ and the other is $qn + q$. A simple special case occurs when $q = \frac{1}{2}$ and n is odd; *e.g.*,

$$(\tfrac{1}{2} + \tfrac{1}{2})^3 = \tfrac{1}{8}(1 + 3 + 3 + 1).$$

Another simple case occurs when $q = \frac{1}{3}$ and $n + 1 = 3$:

$$(\tfrac{2}{3} + \tfrac{1}{3})^2 = \tfrac{4}{9} + \tfrac{4}{9} + \tfrac{1}{9}.$$

COROLLARY. *The mode lies within an interval only half as long as that indicated by the outside limits given in* (*d*). *It extends to a distance* p *on one side of the mean and to a distance* q *on the other side, and the total length is* $p + q = 1$.

This follows immediately from (5).

Example 1. Find the mean, mode, α_3, and α_4 of the point binomial $(\frac{1}{4} + \frac{3}{4})^7$ by means of the formulae. Verify the formulae by computing the moments outright.

Here $p = \frac{1}{4}$, $q = \frac{3}{4}$, $n = 7$, and, by the formulae of (*a*), (*b*), (*c*), (*d*),

$$\bar{t} = \frac{3 \cdot 7}{4} = 5.25; \quad \sigma = \sqrt{\tfrac{1}{4} \cdot \tfrac{3}{4} \cdot 7} = 1.1456;$$

$$\alpha_3 = \frac{-1}{2\sigma} = -.436; \quad \alpha_4 = 3 - \frac{2}{21} = 2.905.$$

There are two greatest terms, at $t = \dfrac{3 \cdot 8}{4} = 6$ and at $t = 5$.

Verification.[1] $(\frac{1}{4} + \frac{3}{4})^7 = \frac{1}{4^7}(1 + 21 + 189 + 945 + 2835 + 5103 + 5103 + 2187)$.

To find the moments it is as well to omit the coefficient, $\frac{1}{4^7}$.

t	f	u
0	1	− 5
1	21	− 4
2	189	− 3
3	945	− 2
4	2835	− 1
5	5103	0
6	5103	1
7	2187	2

$\Sigma f = 16\,384$, $\Sigma fu = 4096$, $\Sigma fu^2 = 22\,528$,
$\Sigma fu^3 = +\,5632$, $\Sigma fu^4 = 79\,360$.

[1] An easy method of constructing drill problems in finding moments is to choose as frequencies the successive terms of a point binomial. The correct answers are given by (*a*), (*b*), and (*c*).

Hence,

$$\bar{u} = \tfrac{1}{4},\ \mu_2 = \tfrac{1}{16},\ \mu_3 = -\tfrac{1}{32},\ \mu_4 = 5.0039;\ \text{and } \bar{t} = 5.25,$$
$$\sigma_u = 1.1456 = \sigma_t,\ \alpha_3 = -.436,\ \alpha_4 = 2.905.$$

EXERCISES § 1

1. Repeat Example 1 for the following point binomials:
(*a*) $(\frac{1}{2} + \frac{1}{2})^4$; (*b*) $(\frac{1}{3} + \frac{2}{3})^5$; (*c*) $(\frac{1}{5} + \frac{4}{5})^4$; (*d*) $(\frac{2}{5} + \frac{3}{5})^5$; (*e*) $(\frac{1}{10} + \frac{9}{10})^4$.

2. If T is as in the proof of (*d*), express as a multiple of T:
(*a*) the second term preceding T;
(*b*) the second term following T;
(*c*) the sum of the five terms nearest T, including T itself.

2. Normal Curve. In many of our problems in probability it was found necessary to add together the values of a set of consecutive terms of a point binomial. When the number of terms is large, this entails so much labor that a simple approximation to this sum is very desirable. The simplest of these is afforded by the normal curve, and in certain types of cases to be described it is sufficiently accurate.

We already know that, if $\bar{t}$ is the mean, the equation of the normal curve may be written:

$$y = \frac{N}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\bar{t})^2}{2\sigma^2}}. \tag{6}$$

Now let us try, by choosing $\bar{t}$ and σ properly, to make this normal curve fit the histogram of the point binomial (see figure for Example 2). Then, since Table I gives the area under any portion of this curve, it will also give the area of the corresponding portion of the histogram, *i.e.*, the sum of the corresponding number of terms of the point binomial.

Using the method of curve fitting outlined in Chapter V, Part I, we shall set $N = 1$, $\bar{t} = nq$, $\sigma^2 = npq$, and then we

shall expect that y will give approximately the point binomial for the various values of t. Before making use of this relation let us test it by a numerical example, and, in order that we may make use of Table I(a), let us substitute in (6)

$$x = \frac{t - \bar{t}}{\sigma}.$$

Then

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{\phi(x)}{\sigma}. \qquad (6a)$$

Example 2. Compute (6a) in the case of the point binomial $(\frac{3}{5} + \frac{2}{5})^6$. We first note that the true value of the terms is:

$$\tfrac{1}{5^6}(64 + 576 + 2160 + 4320 + 4860 + 2916 + 729)$$
$$= .0041 + .0369 + .138 + .277 + .311 + .186 + .047,$$

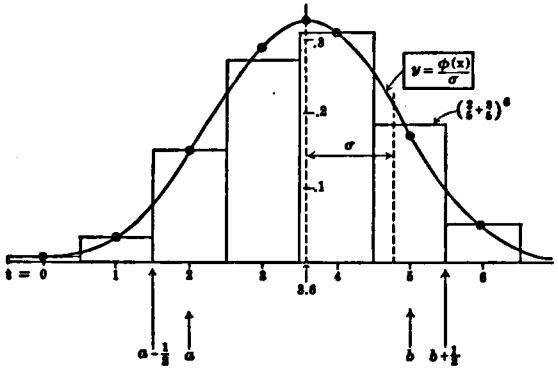
also that $\bar{t} = nq = 3.6$, and $\sigma = \sqrt{npq} = 1.2$.

COMPUTATION OF (6a)

t	$t-\bar{t}$	x	*Table I(a)* $\phi(x)$	*Formula (6a)* $y = \frac{\phi}{\sigma}$	*True y*
0	− 3.6	− 3.00	.0044	.004	.004
1	− 2.6	− 2.17	.0379	.033	.037
2	− 1.6	− 1.33	.1647	.138	.138
3	− 0.6	− 0.50	.3521	.293	.277
4	0.4	0.33	.3778	.315	.311
5	1.4	1.67	.2012	.168	.186
6	2.4	2.00	.0540	.045	.047
$\bar{t} = 3.6$	0.0	0.00	.3989	.331	

To help in plotting the graphs, the value of y is also computed at $t = \bar{t}$. The comparison between the approximate and true values, as indicated in the last two columns and also in the figure, shows how good the fit is. Whether it is good enough or not depends on when and how this approximation is to be used. In general, the fit is good for point

binomials in which n is large and p is nearly equal to q, and the fit is better near the middle than near the ends of a distribution. In making graphs of this sort, the curve should



Point Binomial and Normal Curve

be plotted and drawn before the histogram. Otherwise, it is hard to avoid drawing the curve so that it will fit the histogram as well as possible, rather than allowing it to pass as smoothly as possible through its own plotted points. The curve should be drawn so as to be symmetrical with respect to the ordinate at the mean, and should cross its tangent at a distance from the mean equal to σ. Except when $\bar{t}$ is exactly one of the t's, every point of the curve found by the computation furnishes two points for the diagram, one on one side of the mean ordinate and the other in the symmetrical position on the other side. Both points should be marked before the curve is drawn, but only the computed point should be left permanently marked on the diagram. These suggestions were made also in Chapter V of Part I.

3. First Approximation. We are now ready to express in a theorem the application to probability.

Theorem [1] **II.** *The sum of those terms of the point binomial* $(p + q)^n$ *in which* t *ranges from* a *to* b, *inclusive* $(a \leqq t \leqq b)$, *is approximately:*

$$\int_{x_1}^{x_2} \phi(x)dx,$$

where $$x_1 = \frac{a - \frac{1}{2} - qn}{\sigma}, \quad x_2 = \frac{b + \frac{1}{2} - qn}{\sigma}.$$

This approximation is good if a lies on one side of the mean and b on the other, at approximately equal distances. When these distances are exactly, or very nearly, equal, $qn - a = b - qn$, and therefore $x_1 = -x_2$, and then the formula can be written in a more convenient form, and the theorem can be expressed in language more easily applicable to problems, thus:

COROLLARY 1. *The sum of those terms in which* t *differs from* qn (*or what is the same thing,* s *differs from* pn) *by* k *or less is:*

$$2\int_0^x \phi(x)dx,$$

where $$x = \frac{k + \frac{1}{2}}{\sigma}.$$

COROLLARY 2. *The sum of those terms in which* t *differs from* qn (*or* s *from* pn) *by* k *or more is:*

$$2\int_x^\infty \phi(x)dx = 2\left(1 - \int_{-\infty}^x \phi(x)dx\right),$$

where $$x = \frac{k - \frac{1}{2}}{\sigma}.$$

In both these corollaries, as indicated above, $qn - a = b - qn = k$, and so k and qn must be such that $qn - k$ is an actually occurring exponent a, and that $qn + k$ is an actually occurring exponent b; or at least this must be very nearly true. In practical cases it is usually very nearly true, but almost never exactly true.

[1] Sometimes incorrectly attributed to Bernoulli.

Proof. If we admit that the curve fits the histogram, the proof is graphically obvious, for the sum of the terms in which t ranges from a to b is the same as the sum of the areas of the corresponding rectangles, and to find this it is sufficient to obtain the area under that part of the curve which these rectangles occupy. But we must be careful to include whole rectangles, not half rectangles merely, at the ends. Referring to the figure on page 211, one will note that $t = a$ is the coördinate of the middle of the first rectangle of the four which are nearest the mean point, and so $t = a - \frac{1}{2}$ is the coördinate of the left end. Likewise, b is the coördinate of the middle of the last of these rectangles, and $b + \frac{1}{2}$ is the coördinate of its right end. In this case, then, the area of the middle four rectangles is approximately the area under the curve from $a - \frac{1}{2}$ to $b + \frac{1}{2}$, and this is in general the area required; that is, the area under y from $t_1 = a - \frac{1}{2}$ to $t_2 = b + \frac{1}{2}$, in the t-unit:

$$\int_{t_1}^{t_2} y\,dt, \quad y = \frac{\phi(x)}{\sigma}, \quad x = \frac{t - qn}{\sigma}. \tag{7}$$

This is the same as finding the area under $\phi(x)$ from $x_1 = \dfrac{a - \frac{1}{2} - nq}{\sigma}$ to $x_2 = \dfrac{b + \frac{1}{2} - nq}{\sigma}$, in the x-unit:

$$\int_{x_1}^{x_2} \phi(x)\,dx. \tag{8}$$

For the area from t_1 to t_2 under y in the t-unit is the same as the area from x_1 to x_2 under $\dfrac{\phi(x)}{\sigma}$ in the t-unit, since x_1, x_2 refer to the same points as t_1 and t_2 and since $y = \dfrac{\phi(x)}{\sigma}$. But any area under $\dfrac{\phi(x)}{\sigma}$ in the t-unit is the same as the corresponding area under $\phi(x)$ in the x-unit, since the unit of x is σ times the unit of t. That is, in going from $\dfrac{\phi(x)}{\sigma}$ to $\phi(x)$ we multiply

each ordinate by σ, but if in doing this we also multiply our horizontal unit by σ we have left the area unchanged.[1]

COROLLARY 3. *Whenever a and b lie on opposite sides of the mean, it is more convenient to write the formula of the theorem thus:*

$$\int_{x_1}^{x_2} \phi(x)dx = \int_{-\infty}^{-x_1} \phi(x)dx + \int_{-\infty}^{x_2} \phi(x)dx - 1,$$

when making use of Table I.

Proof.

$$\int_{x_1}^{x_2} = \int_{-\infty}^{x_2} - \int_{-\infty}^{x_1}. \tag{9}$$

By the symmetry of the curve,

$$\int_{-\infty}^{x_1} = 1 - \int_{-\infty}^{-x_1}, \tag{10}$$

as may be seen immediately if the letters are placed on the graph. Now substitute (10) in (9) and get the result desired.

Example 3. Find the sum of those terms in which $t = 2, 3, 4,$ or 5 in $(\frac{2}{5} + \frac{3}{5})^6$. This is the sum of the four middle rectangles in the graph of Example 2. By the formula of Corollary 3:

$$a = 2,\ b = 5,\ \bar{t} = 3.6 = qn,\ \sigma = 1.2,$$

$$x_1 = \frac{2 - .5 - 3.6}{1.2} = -1.75,\ x_2 = \frac{5 + .5 - 3.6}{1.2} = 1.58.$$

By Table I, $\int_{-\infty}^{1.75} + \int_{-\infty}^{1.58} - 1 = .9599 + .9429 - 1 = 0.9028.$

The true value, found by adding the terms, is 0.9125. A closer approximation was not to be expected in this case, for, although $x_1 = -x_2$, approximately, n is quite small.

Example 4. Use Corollary 1 in Example 3. The question might have been put: find the sum of those terms in which t differs from

[1] The student of the calculus will observe that (8) is obtainable by direct substitution in (7). The question of units is cared for in the relation, $dt = \sigma\, dx$.

qn by ($k = 1.5$) or less, since $3.6 - 2 = 1.6$ and $5 - 3.6 = 1.4$. By Corollary 1, then, we have

$$x = \frac{1.5 + .5}{1.2} = 1.667, \quad 2\int_0^{1.667} \phi(x)dx = 2(.4522) = 0.9044.$$

The result happens to be a little closer to the true value than the answer to Example 3, but this must be regarded as an accident. In general, neither result can be relied on to more than two places. Usually the method of Example 3 will prove the better.

Example 5. The philosopher Buffon one day threw a coin 4040 times and noted 2048 heads. Should he have been surprised, either (a) that he came so close to the ideal number 2020, or (b) that he did not come closer to it?

The question (a) may be put thus: What is the probability that, in 4040 throws with a perfect coin, one would obtain a number of heads that would differ from 2020 by 28 or less? The answer is given by Corollary 1: $k = 28, nq = np = 2020, \sigma = \sqrt{1010} = 31.78$, $x = \frac{28.50}{31.78} = 0.898.$

$$2\int_0^x \phi(x)dx = 0.631.$$

He should not have been surprised at the closeness of the result, for, if his experiment were to be repeated indefinitely, this closeness would be expected in 63% of the trials.

The question (b) may be put thus: What is the probability that one would obtain a number which would differ from 2020 by 28 or more? The answer is given by Corollary 2: $k = 28, \sigma = 31.78$, $x = \frac{27.50}{31.78} = 0.865.$

$$2\int_x^\infty \phi(x)dx = 0.387.$$

He should not have been much surprised at this because a deviation as great as he obtained would be expected 39% of the time.

EXERCISES § 3

Use the normal curve in finding approximations to the following sums. The answers are given to four decimal places in order that the student may check his work; but, in general, they represent

the true values of the sums desired only to about two significant figures.

1. The terms of $(\frac{2}{3} + \frac{1}{3})^{450}$ in which $131 \leqq t \leqq 179$. *Ans.*, .9728.

2. The terms of $(\frac{1}{3} + \frac{2}{3})^{90}$ in which $50 \leqq t \leqq 70$. *Ans.*, .9812.

3. The terms of $(\frac{1}{2} + \frac{1}{2})^{30}$ in which $5 \leqq t \leqq 13$. *Ans.*, .9343.

4. The terms of $(\frac{2}{3} + \frac{1}{3})^{450}$ in which the exponent of $\frac{1}{3}$ is greater than 140 and less than 155. *Ans.*, .5025.

5. The terms of $(\frac{1}{3} + \frac{2}{3})^{90}$ in which the exponent of $\frac{1}{3}$ is greater than 45 and less than 75.

6. The terms of $(\frac{1}{2} + \frac{1}{2})^{47}$ in which t differs from $23\frac{1}{2}$ by less than 7.5. *Ans.*, .9588.

7. The terms of $(\frac{1}{2} + \frac{1}{2})^{400}$ in which t differs from 200 by more than 25.

8. The terms of $(\frac{1}{2} + \frac{1}{2})^{400}$ in which t differs from 200 by at least 25. *Ans.*, .0142.

9. The probability that in throwing 500 coins one will obtain at least 260 heads. *Ans.*, .1982.

10. That the number of heads will differ from 250 by less than 20.

4. Closer Approximations. There are various methods of obtaining closer approximations to the sum of a number of consecutive terms of a point binomial. A closer approximation is really needed when p and q are quite different, and when the terms are not arranged symmetrically with respect to the mean. Next in point of simplicity to the normal curve is a curve which is expressible in a series of terms which involve the "polynomials of Hermite." These are designated by the letters $H(x)$ and defined in part as follows: $H_0(x) = 1$, $H_1(x) = -x$, $H_2(x) = x^2 - 1$, $H_3 = -x^3 + 3x$, $H_4 = x^4 - 6x^2 + 3$, $\cdots$. There are an infinite number of them. If each is multiplied by $\phi(x)$, new functions are obtained like the following:

$$\phi^{(0)}(x) = H_0\phi = \phi(x).$$
$$\phi^{(1)}(x) = H_1\phi = -x\phi(x).$$

$$\phi^{(2)}(x) = H_2\phi = (x^2 - 1)\phi(x).$$
$$\phi^{(3)}(x) = H_3\phi = (-x^3 + 3x)\phi(x).$$
$$\phi^{(4)}(x) = H_4\phi = (x^4 - 6x^2 + 3)\phi(x).$$

These functions are very useful. If we multiply each by a constant and add them, we shall get another new function $F(x)$:

$$F(x) = c_0\phi^{(0)} + c_1\phi^{(1)} + c_2\phi^{(2)} + \cdots. \tag{11}$$

This series is sometimes called a Gram-Charlier series in honor of two mathematicians who have worked with it.[1] By a proper choice of constants, c_0, c_1, etc., this series can be made to fit very closely almost any ordinary frequency curve, provided an infinite number of terms is used. If only three or four terms are used, it is often a very good approximation to a frequency distribution. Later on we shall see how to determine the c's in order to make it fit various histograms, including the point binomial, but we postpone for the present this theoretical discussion and use only the result. It yields the following theorem, which is analogous to Theorem II.

Theorem III. *The sum of those terms of the point binomial* $(p + q)^n$ *in which* t *ranges from* a *to* b, *inclusive* $(a \leq t \leq b)$, *is approximately* (*if* n *is fairly large*):

$$\int_{x_1}^{x_2}\phi(x)dx + \left[\frac{q - p}{6\sigma}\phi^{(2)}(x) + \frac{1}{24}\left(\frac{1}{\sigma^2} - \frac{6}{n}\right)\phi^{(3)}(x)\right]_{x_1}^{x_2},$$

where, as in Theorem I,

$$x_1 = \frac{a - \frac{1}{2} - qn}{\sigma}, \quad x_2 = \frac{b + \frac{1}{2} - qn}{\sigma};$$

[1] Pearson uses essentially the same series and calls it a series of tetrachoric functions. The student of the calculus can easily show that $\phi^{(1)}(x) = \frac{d\phi}{dx}$, $\phi^{(2)}(x) = \frac{d^2\phi}{dx^2}$, etc. The definitions of the H's are obtained from these derivatives.

and, as in Corollary 3, we may write, if we choose,

$$\int_{x_1}^{x_2} = \int_{-\infty}^{-x_1} + \int_{-\infty}^{x_2} - 1;$$

and where the square bracket has the same meaning as in the calculus, thus:

$$\Big[f(x)\Big]_{x_1}^{x_2} \quad \text{means} \quad f(x_2) - f(x_1).$$

The functions $\phi^{(2)}(x)$ and $\phi^{(3)}(x)$ are given in Tables II and III. Let us note that, by their definitions,

$$\phi^{(2)}(-x) = \phi^{(2)}(x), \text{ but that } \phi^{(3)}(-x) = -\phi^{(3)}(x).$$

COROLLARY. *It is often sufficient to use but two terms of the formula:*

$$\int_{x_1}^{x_2} \phi(x)dx + \left[\frac{q-p}{6\sigma}\phi^{(2)}(x)\right]_{x_1}^{x_2}.$$

When but one term is used, we have Theorem II. We have called this a first approximation. When two terms are used, we shall call it a second approximation, and, when three terms are used, a third approximation. Some authors compute more terms of the series (11) and obtain higher approximations, but this is of doubtful value unless a large number of terms is employed, and then more labor is encountered and more tables necessary.[1] There is no simple formula which will give a sufficiently good approximation in all cases, and therefore the general problem of finding such approximations is rather complicated. It is discussed more fully in Part III, but we may say here that it is usually safe to use Theorem III if x_1 and x_2 are not very far from the mean and if p is not very different from q. To be more specific, we may use Theorem III if

$$R > 0.5, \quad x < 5, \quad n > 25, \tag{12}$$

where R is the smaller of the quantities $\dfrac{(a-1)p}{(n-a+2)q}$,

[1] If more terms are used, it is desirable to add them in groups, rather than singly. (Cf. *Fry*, page 255.) Cf. also Appendix, § 3.

$\frac{(n-b-1)q}{(b+2)p}$, and x is the larger of the quantities x_1, x_2. These conditions are sufficient rather than necessary, and we may in fact use Theorem III when one of them is not quite satisfied, if the others are easily satisfied. When n is as small as 25, it is rather better to compute enough terms outright. This is not very difficult if a machine is used and the work is well planned. However, we may use the formulae for even smaller values of n in the more symmetrical cases.

Example 6. For the point binomial $(\frac{1}{2}+\frac{1}{2})^{450}$, find the sum of those terms in which the exponent t of $\frac{1}{2}$ lies in the interval $131 \leqq t \leqq 179$.

The author has computed the true value in this case. It is 0.9735. We shall now approximate it by the use of Theorem III. Incidentally we shall find how good the approximation would have been had we used two terms or one term instead of three terms of our series:

$$\bar{t} = 150,\ \sigma = 10,\ x_1 = \frac{131 - .5 - 150}{10} = -1.95,$$

$$x_2 = \frac{179 + .5 - 150}{10} = 2.95,$$

$$\frac{q-p}{6\sigma} = -\frac{1}{180},\ \frac{1}{24}\left(\frac{1}{\sigma^2} - \frac{6}{n}\right) = \frac{-1}{7200}.$$

1st approximation: $\int_{x_1}^{x_2} \phi(x)dx = 0.9728.$ (*a*)

2nd approximation: $.9728 - \frac{1}{110}[\phi^{(3)}(x_2) - \phi^{(3)}(x_1)] = 0.9735.$ (*b*)

3rd approximation: $.9735 - \frac{1}{7200}[\phi^{(5)}(x_2) - \phi^{(5)}(x_1)] = 0.9735.$ (*c*)

In the inequalities of (12), R is the smaller of the quantities $\frac{130\cdot 2}{321\cdot 1} = .81$, $\frac{270\cdot 1}{181\cdot 2} = .75$, both of which are greater than 0.5.

EXERCISE § 4. Repeat the methods of Example 6 in the following cases: § 3, Exercises 2, 3, 4, 5.

5.[1] **Theorem IV.** *The mean of the hypergeometric series of* § 6, *Chapter I, page* 199, *is, relative to the origin of t,* $\bar{t} = nq$.

[1] This section, like § 6 of Chapter I, may be omitted if desired.

The proof is very similar to the proof of Theorem I(a):

$$\bar{t} = \frac{1}{{}_mC_n}\sum_{t=0}^{n} t\,{}_{pm}C_s\,{}_{qm}C_t = \frac{\lfloor n \lfloor m-n}{\lfloor m}\sum_{t=1}^{n} t\,{}_{pm}C_s\frac{\lfloor qm}{\lfloor qm-t \lfloor t}$$

$$= \frac{nqm}{m}\frac{\lfloor n-1 \lfloor m-n}{\lfloor m-1}\sum_{t=1}^{n} {}_{pm}C_s\frac{\lfloor qm-1}{\lfloor qm-t \lfloor t-1}$$

$$= nq\left(\frac{1}{{}_{m-1}C_{n-1}}\sum_{t-1=0}^{t-1=n-1} {}_{pm}C_s\,{}_{qm-1}C_{t-1}\right).$$

Now, by Theorem VI, § 6, Chapter I, the part in parentheses is the hypergeometric series corresponding to a sample of $(n-1)$ balls drawn from a bag containing $(m-1)$ balls of which pm are white and $(qm-1)$ are black. (Note that $pm + qm - 1 = m - 1$, $s + t - 1 = n - 1$.) Hence, by the corollary to Theorem VI, this part is 1. So $\bar{t} = nq$.

PROBLEMS CHAPTER II

1. Prove Theorem I (c).

Find the sums of the terms of the point binomials as indicated below. In each case the true values are given, and in the first case the answers which the student should obtain by the use of the approximate formulae are also given. A good idea of the accuracy of these formulae is afforded by these examples.

2. $(\frac{2}{3} + \frac{1}{3})^{25}$. *Cf.* p. 375.

t	*True* S_{t+1}	*Approximation*
4	.04620	.0480
6	.22215	.2224
8	.53758	.5374
10	1 − .17799	1 − .1794
12	1 − .04151	1 − .0428
14	1 − .00560	1 − .00591
16	1 − .000415	1 − .000452
$7 \leqq t \leqq 15$	.77620	.7758
$6 \leqq t \leqq 16$	.8881	.8860

3. $(\frac{2}{3} + \frac{1}{3})^{450}$.

t	True S_{t+1}
75	1.4342×10^{-15}
115	.000214
125	.0066
135	.0726
145	.3282
165	$1 -$ .0614
175	$1 -$ .00582
185	$1 -$ .000237
195	$1 - 4.1 \times 10^{-6}$
$126 \leqq t \leqq 165$	.9321
$136 \leqq t \leqq 175$	.9216

4. $(\frac{8}{9} + \frac{1}{9})^{25}$.

t	True S_{t+1}
1	.2141
2	.4632
3	$1 -$ .2999
5	$1 -$ .05202
7	$1 -$ .00440
9	$1 - 1.94 \times 10^{-4}$
$1 \leqq t \leqq 8$	.94639
$2 \leqq t \leqq 10$	.7859

5. $(.99 + .01)^{500}$.

t	True S_{t+1}
3	.2635
4	.4395
5	$1 -$ .3841
6	$1 -$ .2372
7	$1 -$ .1324
8	$1 -$ .06724
12	$1 -$ .001902
15	$1 - 6.143 \times 10^{-5}$
$3 \leqq t \leqq 8$	.8095

6. At age 65 the probability of death in one year is about .04, and so the cost of insuring a man of that age for one year would be \$40 per \$1000, if loading and interest is neglected. If a company is carrying 500 cases of this sort and charges \$42, what is the probability that the premiums collected will be insufficient to pay the death claims?

7. At one time it was supposed that male and female children were equally likely. Out of 10,000 births it was found that 5098 were male children (Laplace). Was there either a surprisingly large or a surprisingly small deviation from the expected number?

8. Answer the questions of Problem 7 accepting a theory of sex determination which would make the proportion of male to female children 36 to 34.

9. (*a*) A coin is tossed 81 times. What is the probability that the number of heads will differ from 40.5 by 5.5 or less? (*b*) Answer (*a*) after rephrasing the question by multiplying each of those numbers by 100.

10. A census report showed that in general 59.58% of New York City children went to school, but that only 56.8% of the negro children went to school. The number of negro children was 20,000. Was the difference due to chance?

11. In a certain university the proportion of freshmen failing in mathematics was 10%. (*a*) Twenty-five per cent failed in a section of 16. How abnormal was the discrepancy? (*b*) Same as (*a*) for a section of 160.

12. In general about 21.6% of men alive at age 25 are dead before they reach 50. If it were found that, of 890 athletes living at age 25, 205 were dead before reaching 50, would one be justified in concluding that there is a really different mortality rate applicable to athletes?

13. If a baseball player has a batting average of 22%, what is the probability of at least 22 hits out of 100 times at bat?

14. A speculator can guess correctly the daily changes in the stock market 55% of the time, and he wagers \$1000 every day for 300 days. (*a*) What is the probability that he will clear at least \$50,000? (*b*) What is the probability that he will lose \$10,000 or more in the first 100 days?

15. (*Fry*) "There is a case on record where a die was thrown 315,672 times with the result that either 5 or 6 appeared 106,602 times." Is it reasonable to suppose that the die was a true one?

16. (*Coolidge*) "In 1850 the Swiss astronomer Wolff threw two dice 100,000 times. The two showed the same face 16,647 times." Comment on this result.

17. Two dice are thrown 500 times. What is the standard deviation of the probability distribution for the number of times the sum 7 will appear?

18. In general, 10% of persons afflicted with a certain disease die. Of a number of persons thus afflicted and subjected to various methods of treatment the following data were gathered. Do any of these methods appear worth further study? That is, in each case find the probability that a percentage as different from 10 as that observed would occur by chance.

Treatment	*A*	*B*	*C*	*D*	*E*
Deaths	35	15	7	3	900
Number Diseased	400	200	100	50	10,000
Per Cent Deaths	8.75	7.5	7.0	6.0	9.0

19. The probable error of a gun battery is 40 yards in longitude. One hundred shots are fired at an enemy position which is 80 yards deep and extends indefinitely to the right and left of the line of fire. What is the probability that more than 60% of the shots are hits?

20. A dean's report showed the following figures:

Subject	*Honor Grades*		*Failures*		*Number Examined*
	Number	%	*Number*	%	
German	187	36	33	6.3	521
Mathematics	162	35	38	8.2	466
Music	11	50	0	0	22
All Subjects		38		5.4	

Find the probability: (*a*) that, in selecting 521 students at random (from a supposedly infinite number), one would obtain as few honor grades as were obtained in German, (*b*) as many failures; (*c*) that, in selecting 466 at random, one would obtain as few honor grades as were obtained in mathematics, (*d*) as many failures; (*e*) that, in selecting 22, one would obtain no failures, (*f*) eleven or more honor grades.

21. Prove (*c*) of Theorem I, § 1.

CHAPTER III

FREQUENCY CURVES

1. The Gram-Charlier Series. It was stated in Chapter II that the series

$$\sum_{i=0}^{\infty} c_i\phi^{(i)} \equiv c_0\phi^{(0)} + c_1\phi^{(1)} + c_2\phi^{(2)} + \cdots \qquad (1)$$

could be made to fit almost any ordinary frequency distribution if the c's were chosen properly. The formulae which determine the proper choice of these c's are commonly given in the more advanced books. We shall be content here to discuss the simple and most useful case where there are but five terms,

$$F(x) \equiv c_0\phi^{(0)} + c_1\phi^{(1)} + c_2\phi^{(2)} + c_3\phi^{(3)} + c_4\phi^{(4)}. \qquad (1a)$$

Let us then think of $F(x)$ as representable by a curve made by adding together the five curves indicated by the five terms on the right of this equation. Let us suppose we have a given frequency distribution, $f(t)$, and that we wish to choose c_0, c_1, c_2, c_3, c_4 so that this curve will fit the distribution as well as possible. The method of fitting will be the method of moments used often before. That is, we shall have five equations which will determine the five constants. The first equation will say that the $0th$ moment of f is equal to the $0th$ moment of F; the second that the first moment of f is equal to the first moment of F; etc.

It will simplify matters a little if we suppose that the total given frequency is 1 instead of N, if we take as our horizontal unit the given standard deviation σ_t, and if we use the given mean as our origin. To accomplish this, we let $f(t)$ designate

as before the frequency in the interval whose mid-point is a given coördinate t, or $f(u)$ the frequency in the interval whose mid-point is an arbitrary coördinate u, and put $x = \frac{t - \bar{t}}{\sigma_t} = \frac{u - \bar{u}}{\sigma_u}$; then $f(x)$ will designate the frequency in the interval whose mid-point is x. Moreover, the origin and mean of x will be zero, and its unit will be the given standard deviation. The equations to determine the c's will now become:

$$\left.\begin{aligned} \textit{Area of } F(x) &= 1, \textit{ i.e., } \int_{-\infty}^{\infty} F(x)dx = 1, \\ \textit{Mean of } F(x) &= 0, \textit{ i.e., } \int_{-\infty}^{\infty} xF(x)dx = 0, \\ \mu_2 \textit{ of } F(x) &= 1, \textit{ i.e., } \int_{-\infty}^{\infty} x^2F(x)dx = 1, \\ \mu_3 \textit{ of } F(x) &= \textit{given } \alpha_3, \textit{ i.e., } \int_{-\infty}^{\infty} x^3F(x)dx = \alpha_3, \\ \mu_4 \textit{ of } F(x) &= \textit{given } \alpha_4, \textit{ i.e., } \int_{-\infty}^{\infty} x^4F(x)dx = \alpha_4. \end{aligned}\right\} \quad (2)$$

From these equations the following values of the c's may be determined:[1]

$$c_0 = 1,\ c_1 = 0,\ c_2 = 0,\ c_3 = -\frac{\alpha_3}{6},\ c_4 = \frac{\alpha_4 - 3}{24}, \quad (3)$$

so that our curve (1a) may be written in the form:

$$F(x) = \phi(x) - \frac{\alpha_3}{6}\phi^{(3)}(x) + \frac{\alpha_4 - 3}{24}\phi^{(4)}(x), \quad (1b)$$

where, as stated, $x = \frac{u - \bar{u}}{\sigma_u}$, and σ_u, α_3, and α_4 apply to the given distribution.

This expression for $F(x)$ may be expected to yield a curve which will fit the given histogram in the following sense: The area of the curve over any given interval will equal approximately the area of that part of the histogram

[1] See § 2.

whose base is that interval, provided the total area or frequency happens to be 1. If it is N, instead of 1, then $F(x)$ yields in each case the relative frequency. To get the absolute frequency we must multiply this by N. That is, if the interval goes from $x = a$ to $x = b$:

$$\textit{Frequency over } (a, b) = N\int_a^b F(x)dx. \tag{4}$$

If in addition we require y the ordinate of the graduating curve, we must transfer back to the given (t) unit. Analogous to equation (6) of Part I, Chapter V, page 71, we have then

$$y = \frac{N}{\sigma_t} F(x). \tag{5}$$

2. Properties of the ϕ's. By (1*b*)

$$\int_a^b F(x)dx = \int_a^b \phi(x)dx - \frac{\alpha_3}{6}\int_a^b \phi^{(3)}(x)dx + \frac{\alpha_4 - 3}{24}\int_a^b \phi^{(4)}(x)dx, \tag{4a}$$

but so far we have found no means of computing the last two of these integrals. It turns out that the ϕ's have certain simple properties which make this computation easy. The first property is:

$$(a)^1 \qquad \int_a^b \phi^{(3)}(x)dx = \phi^{(2)}(x)\Big]_a^b, \quad \int_a^b \phi^{(4)}(x)dx = \phi^{(3)}(x)\Big]_a^b,$$

and similar relations hold for all the ϕ's.

Proof. This cannot be completely proved without the calculus, but it is very easy to see that it is at least approximately true in special cases.

Example 1. Find, from Table IV, $\int_{.05}^{.55} \phi^{(4)}(x)dx$, approximately.

By Table IV the histogram is given by the following data:

[1] To students of the calculus: this property is a result of the fact that the ϕ's are successive derivatives of ϕ^0.

x	$\phi^{(4)}(x)$
.1	1.1671
.2	1.0799
.3	0.9413
.4	0.7607
.5	0.5501

The area is the class interval times $\Sigma\phi^{(4)}(x)$, *viz.*, $(.1)(4.4991) = 0.44991$.

Let us now notice that this result is approximately equal to

$$\Big[\phi^{3}(x)\Big]_{.05}^{.55} = \phi^{(3)}(.55) - \phi^{(3)}(.05) = .5088 - .0597 = 0.4491,$$

by Table III.

Example 2. Find, from Table III, $\int_{.975}^{1.575} \phi^{(3)}(x)$, approximately, taking the class interval equal to 0.05.

x	$\phi^{(3)}(x)$
1.00	.4839
1.05	.4580
1.10	.4290
1.15	.3973
1.20	.3635
1.25	.3282
1.30	.2918
1.35	.2550
1.40	.2180
1.45	.1815
1.50	.1457
1.55	.1111
Total	3.6630

Area $= (.05)(3.6630) = 0.18315$.

The formula gives

$$\phi^{(2)}(1.575) - \phi^{(2)}(0.975) = .1709 - (-.0122) = 0.1831.$$

The agreement is better than in Example 1, as was to have been expected, because of the smaller class interval.

From (a) and $(4a)$ we derive:

$$\int_a^b F(x)dx = \int_a^b \phi(x)dx - \frac{a_3}{6}\Big[\phi^{(2)}(b) - \phi^{(2)}(a)\Big] + \frac{a_4 - 3}{6}\Big[\phi^{(3)}(b) - \phi^{(3)}(a)\Big]. \quad (4b)$$

This enables us in any given case to write down immediately from the tables the value of the partial area of $F(x)$. (Cf. Example 3.) The ϕ's have other properties also that are valuable, (*b*) and (*c*):

$$(b)\quad \begin{cases} \int_{-\infty}^{\infty} \phi(x)dx = 1, \int_{-\infty}^{\infty} x^2\phi(x)dx = 1, \int_{-\infty}^{\infty} x^4\phi(x)dx = 3; \\ \int_{-\infty}^{\infty} x^n\phi(x)dx = 0, \text{ if } n \text{ is odd.} \end{cases}$$

These were noted in Part I. They say in words that the area and the second moment of ϕ are both unity, the *4th* moment is 3, and the odd moments all vanish.

Proof of (3). These sets of properties enable us to derive the first two equations of (3) from the first two equations of (2); for, writing out in full the first equation of (2), we have

$$1 = \int_{-\infty}^{\infty} F(x)dx = c_0\int_{-\infty}^{\infty}\phi(x)dx + c_1\int_{-\infty}^{\infty}\phi^{(1)}(x)dx + c_2\int_{-\infty}^{\infty}\phi^{(2)}(x)dx + c_3\int_{-\infty}^{\infty}\phi^{(3)}(x)dx + c_4\int_{-\infty}^{\infty}\phi^{(4)}(x)dx.$$

Insert the property (*a*) and this becomes, since at $\pm\infty$ all the ϕ's vanish,

$$1 = c_0 + 0.$$

Now write the second equation of (2) in full:

$$0 = \int_{-\infty}^{\infty} xF(x)dx = c_0\int_{-\infty}^{\infty} x\phi(x)dx + c_1\int_{-\infty}^{\infty} x\phi^{(1)}(x)dx + c_2\int_{-\infty}^{\infty} x\phi^{(2)}(x)dx + c_3\int_{-\infty}^{\infty} x\phi^{(3)}(x)dx + c_4\int_{-\infty}^{\infty} x\phi^{(4)}(x)dx.$$

Use (*a*) again and also the definitions of the ϕ's in § 4, Chapter II:

$$0 = c_0\cdot 0 + c_1\int_{-\infty}^{\infty} - x^2\phi(x)dx + c_1\int_{-\infty}^{\infty}(x^3 - x)\phi(x)dx + c_3\int_{-\infty}^{\infty}(-x^4 + 3x^2)\phi(x)dx + c_4\int_{-\infty}^{\infty}(x^5 - 6x^3 + 3x)\phi(x)dx.$$

Finally, use property (*b*):

$$0 = -c_1 + 0 + c_3(-3 + 3) + 0; \text{ so } c_1 = 0.$$

This is the second of equations (3). To obtain the rest of (3) we shall need:

$$(c) \quad \int_{-\infty}^{\infty} x^6\phi(x)dx = 3 \cdot 5, \quad \int_{-\infty}^{\infty} x^8\phi(x)dx = 3 \cdot 5 \cdot 7, \text{ etc.}$$

We cannot prove (c) without the calculus. The application to equation (3) is one of the problems (9) given at the close of this chapter.

3. Graduation.

Example 3. Find and plot the F curve [1] which best fits the following distribution. Also graduate the distribution by means of this curve.

CORNSTALKS

Height, ft.	f	*Arbitrary* u	fu	fu^2	fu^3	fu^4
3–4	3	− 4	− 12	48	− 192	768
4–5	7	− 3	− 21	63	− 189	567
5–6	22	− 2	− 44	88	− 176	352
6–7	60	− 1	− 60	60	− 60	60
7–8	85	0	0	0	0	0
8–9	32	1	32	32	32	32
9–10	8	2	16	32	64	128
Sums	$217 = N$		− 89	323	− 521	1907
$\frac{1}{N}\Sigma$			− .410	1.488	− 2.401	8.788

Hence, $\bar{u} = -.410$; corrected $\sigma_u = 1.113$, $\alpha_3 = -.4643$, $\alpha_4 = 3.706$. Therefore, $c_3 = 0.0774$, $c_4 = 0.0294$, and the equation of the F curve is

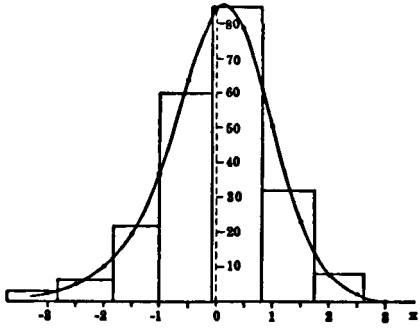
$$F(x) = \phi(x) + .0774\ \phi^{(3)}(x) + .0294\ \phi^{(4)}(x), \quad \text{by (1}b\text{).}$$

Hence, by (5), § 1, and since $N/\sigma = 194.97$, we have as our ordinate y at any point x,

$$y = 194.97\ \phi(x) + 15.09\ \phi^{(3)}(x) + 5.73\ \phi^{(4)}(x).$$

[1] We shall reserve the letter F to designate the curve (1a), just as the letter ϕ has been reserved for the normal curve. This use of F is peculiar to this text; the use of ϕ for the normal curve is very common.

For the purpose of plotting the curve any desired values of x may be chosen. It is convenient to take them as they are taken here, at equal intervals of $\frac{1}{2}$ each from -3 to 3. Table IV will be found convenient and adequate for this purpose. The values of $\phi^{(4)}(x)$ are not given except at the infrequent points of this table because in this text they are not needed at other points.[1]



Cornstalks and F Curve

THE ORDINATES, y

x	$194.97\,\phi(x)$	$15.09\,\phi^{(3)}(x)$	$5.73\,\phi^{(4)}(x)$	y
− 3.0	.858	1.204	.762	2.824
− 2.5	3.412	2.149	.458	6.019
− 2.0	10.528	1.630	− 1.547	10.611
− 1.5	25.249	− 2.199	− 4.035	19.015
− 1.0	47.183	− 7.302	− 2.773	37.108
− .5	68.649	− 7.305	3.152	64.496
0	77.774	0	6.858	84.632
.5	68.649	7.305	3.152	79.106
1.0	47.183	7.302	− 2.773	51.712
1.5	25.249	2.199	− 4.035	23.413
2.0	10.528	− 1.630	− 1.547	7.351
2.5	3.412	− 2.149	.458	1.721
3.0	.858	− 1.204	.762	0.416

We graduate the distribution by the use of (4b), placing a and b in that formula equal to the coördinates of the end points of our successive intervals. $N = 217$.

[1] However, they could be found approximately from Table III at intermediate points by the use of the following formula:

$$\phi^{(4)}(x) = 100[\phi^{(3)}(x + .01) - \phi^{(3)}(x)];$$

e.g., $\phi^{(4)}(2.60) = 100\,(- .1317 + .1328) = 0.11.$

GRADUATION

OBSERVED f	END u	END x	$\int_{-\infty}^{x}\phi\,dx$	(1) $\frac{}{\int_a^b \phi\,dx}$	$\phi^{(3)}(x)$	(2) $\frac{}{c_3[\phi^{(3)}(b)-\phi^{(3)}(a)]}$	$\phi^{(3)}(x)$	(3) $\frac{}{c_4[\phi^{(3)}(b)-\phi^{(3)}(a)]}$	(1)+(2)+(3) $\frac{f}{N}$	THEORETICAL f
	− 4.5	− 3.67	.0001		.0059		.0182			
3				.0026		.0039		.0027	.0092	2.00
	− 3.5	− 2.78	.0027		.0563		.1100			
7				.0274		.0090		− .0012	.0352	7.64
	− 2.5	− 1.88	.0301		.1727		.0685			
22				.1334		− .0141		− .0165	.1028	22.31
	− 1.5	− 0.98	.1635		− .0098		− .4933			
60				.3042		− .0298		.0117	.2861	62.08
	− 0.5	− .081	.4677		− .3950		− .0964			
85				.3262		.0233		.0188	.3683	79.92
	0.5	0.82	.7939		− .0934		.5440			
32				.1634		.0210		− .0158	.1686	36.59
	1.5	1.72	.9573		.1780		.0065			
8				.0382		− .0078		− .0041	.0263	5.71
	2.5	2.61	.9955		.0769		− .1317			

In each case, $x = \dfrac{u - \bar{u}}{\sigma_u}$, thus: $x_1 = \dfrac{-4.5 + .410}{1.113} = -3.67$. The rectangles of the graph are drawn by the use of the x's in column 3 of this computation.

EXERCISES §§ 1-3

1. Plot the equation (1b) in the special case: $\alpha_3 = 0.6$, $\alpha_4 = 4.2$, taking $x = 0, \pm 1, \pm 2, \pm 3$.

2. Plot the histogram of which the central ordinates are indicated by the points of Exercise 1. For the frequency distribution thus represented, find the area, σ, α_3, and α_4. By (2), α_3 should be equal to 0.6, approximately, and α_4 to 4.2. The area should equal 1 if the unit is σ, otherwise the area should equal σ.

3. Display the effect of the skewness factor by placing on the same diagram: (a) $\phi(x)$, (b) $\phi(x) - \dfrac{\alpha_3}{6}\phi^{(3)}(x)$, and ($c$) $\phi(x) + \dfrac{\alpha_3}{6}\phi^{(3)}(x)$, taking α_3 as in Exercise 1.

4. Display the effect of the kurtosis factor by placing on the same diagram: (a) $\phi(x)$, and (b) $\phi(x) + \dfrac{\alpha_4 - 3}{24}\phi^{(4)}(x)$, taking α_4 as in Exercise 1.

5. Using the corrected values of the moments of the distribution of wages in Example 1, Chapter IV, Part I, page 50: (a) compute the ordinates of the F curve which fits it and plot as in Example 3; (b) graduate the distribution.

6. Do the same for the data of Example 2, Chapter IV, Part II, page 250.

4. Other Frequency Curves and Their Uses. The Gram-Charlier series may be thought of as a system of frequency curves, infinite in number, of which the F curve is a particular case. Karl Pearson's system of curves is another infinite group, and here again there are one or two particular curves that are of special interest. Both these systems have as a very particular case the normal curve. Probably the two

most important of Pearson's more general curves are his so-called *Type I* curve and *Type III*.

Type I: $y = at^b(c - t)^d$, where $0 \leqq t \leqq c$, $0 < a, b, c, d$. This function occurs frequently in higher mathematics. Its partial area from 0 to any fixed point t is called the incomplete Beta function:

$$\int_0^t at^b(c - t)^d dt.$$

Type I is a limited curve in the sense that y actually reaches the value zero at the end points, $t = 0$ and $t = c$. It rises to a maximum at the intermediate point $t = \frac{bc}{(b + d)}$. Therefore it appears to be well suited to represent a uni-modal frequency distribution.

Type III: $y = ae^{-bt}(b + t)^c$, where $-b \leqq t$, $0 < a, b, c$. This is limited on one side only. The value of y is exactly zero at the end point $t = -b$, and approaches zero as a limit as t becomes infinite. It rises to a maximum at the intermediate point, $t = -b + \frac{c}{b}$. This function also occurs frequently in higher mathematics, and its partial area from $-b$ to any fixed point t is called the incomplete Gamma function:

$$\int_{-b}^t ae^{-bt}(b + t)^c dt.$$

Except when special tables are available, the process of graduating a distribution by means of these curves is rather more tedious than the process of graduating by the F curve, and as graduation is not very important in elementary statistics, we shall not explain how it is carried out for Pearson's types.

5. Uses of Frequency Curves. The reader may well inquire: What is the use of a frequency curve? Why try

to find a mathematical expression which will approximate the given frequencies of an observed distribution? The answers may be grouped under three heads:

(*a*) *Smoothing.* Instead of making use of the observed frequencies, which often contain irregularities, it is sometimes better to use the theoretical frequencies obtained by graduation. In other words, if we desire to smooth out the irregularities of our data by a mathematical process, we may resort to graduation by means of one of these curves. Many instances of the desirability of smoothing may be found in actuarial data. Consider the mortality table itself. This tells us the number of persons in a given group who will be alive at each age from one year up. The original data from which such a table is constructed are obtainable from the census reports. But, as they stand initially, these reports are not wholly reliable, for they contain too many irregularities.[1] If one should construct a mortality table from these reports without smoothing them out, it would be unreliable. Naturally the actuary prefers in this case the theoretical numbers to the observed numbers. The curve used to graduate the mortality table is not one of the frequency curves recently discussed, for the curve required is a continuously descending one, instead of being high in the middle and low at both ends, but this is not an important distinction, because it depends on an artificial choice of what is recorded. For example, tables of the reported deaths at each age exhibit the same sort of irregularity at the five-

[1] One particularly interesting irregularity is that, according to these reports, there are more persons alive at age 30 than at age 29, more at age 35 than at age 34, more at age 40 than at 39, and so on up to 85 or so. At every age which is a multiple of 5 the curve, which in general is a smoothly descending curve, has a bump in it. This is doubtless due to the fact that people do not tell the truth about their own ages or, more especially, about the ages of members of their households. "How old is so and so?" says the census taker, and so and so, being really 31 or 32, and known to be 30 or so, is said to be 30.

year intervals, and such tables are frequency distributions of the type we have studied.[1]

(*b*) *Classification.* If all the frequency distributions with which the statistician has to deal could be separated into general classes, so that one could say with some assurance that this distribution would be an F curve, that a Pearson Type I curve, etc., it would be very helpful in theoretical investigations. Indeed, most of the sampling theory which will be explained in the next chapter can be proved to be valid only if certain of the distributions involved are of comparatively simple types; in many instances it is necessary to assume that they are actually normal. So it is obviously necessary that the practical investigator should examine a wide variety of samples and find out what types may reasonably be assumed in a theoretical investigation. This seems to the author the most important of the several possible objectives of curve fitting, and it is quite fundamental to mathematical progress in statistics.

(*c*) *Testing A Priori Theories.* Suppose one has a theory as to the origin of a certain group of data, and that by means of this theory one could predict the type of frequency curve which the data should fit. The fit turns out to be very good. Therefore the theory is, to that extent, substantiated. This sort of argument is a bit perilous, for, it will be remembered, it is not considered good logic to derive the hypothesis from the conclusion, simply because it does yield the conclusion. However, the reasoning is valid, provided we are careful to regard our theories not as statements of the causes of phenomena but merely as terse summaries of the known facts. Theory and observation do not need to be thought of as the hypothesis and conclusion of a syllogism. We may think of theory merely as a formula which describes the observations.

[1] For numerous examples, see Elderton, *Frequency Curves and Correlation.*

Example 4. *The Law of Error.* It has long been known that it is possible to derive the normal law as the curve of the distribution which errors in a set of physical observations should follow, provided one makes certain assumptions about what are called elementary errors. The idea can be illustrated rather well when applied to the firing of a gun. Suppose there are a certain fixed number of errors that may be made when the gun is fired. For simplicity, suppose each is either plus or minus. If it is plus, the shot will go too far by a certain amount. Let plus and minus errors be equally likely. From these assumptions it follows that the chance combinations of plus and minus errors will produce, at the target, a dispersion of normal type. Moreover, one would get the same curve if these elementary errors were not all equal in effect. Similarly, if one performs a physical experiment such as measuring the length of a steel rod, or timing the passage of a star behind a micrometer wire, or striking at a golf ball, it is supposed that there are in one's physical, mental, and moral make-up certain little impulses, some of which work one way and some the other. The final act depends on the chance combination of these impulses. No one can predict any single act, but the total effect of a million acts will be the "normal law of error." This is the *a priori* theory. The question as to whether it is justified by the facts is obviously a problem in curve fitting, in this case normal curve fitting.[1]

PROBLEMS CHAPTER III

1. From Table IV obtain the approximate value of

$$\int_{.55}^{1.25} \phi^{(4)}(x)dx,$$

and compare the result with the true value as given by § 2 (*a*).

2. Do the same for $\int_{2.05}^{2.55} \phi^{(4)}(x)dx.$

[1] One can push this theory further. Suppose that, as one measures, the "true" value changes, due to progressive changes in the instruments, perhaps, in the case of the steel rod; due to increasing fatigue or loss of morale in the case of the golf ball. Then the resulting distribution will not be normal, but may depart from it widely. It will usually be uni-modal, but may be saddle-shaped.

3. Fit an F curve to the distribution of freshman weights of Part I, Chapter II, Problem 1*a*. (Mean weight = 142.25 lbs. For a unit of 11 lbs., corrected $\mu_2 = 2.56$, $\mu_3 = 2.42$, $\mu_4 = 24.10$.) (*a*) Find and plot the ordinates, F, on the same diagram as the histogram. (*b*) Graduate this distribution (p. 33).

4. Do the same for the following distribution of school grades (*Fishwild*).

Grade	f	*Grade*	f
50	1 000	80	13 000
55	1 000	85	13 000
60	2 000	90	25 000
65	2 000	95	23 000
70	5 000	100	9 000
75	6 000	Total	100 000

Mean grade = 86.35, corrected $\sigma_u = 2.077$, unit of $u = 5$ grades, $\alpha_3 = -1.16$, $\alpha_4 = 4.19$.

5. Do the same for the ages of women at marriage (*Burgess*).

Age	f	*Age*	f
13–17	99	38–42	14
18–22	732	43–47	7
23–27	461	48–52	1
28–32	120	53–57	0
33–37	37	58–62	1

Mean age = 22.78, corrected $\sigma = 4.69$ years, $\alpha_3 = 1.814$, $\alpha_4 = 9.22$.

6. Show that the first of the relations (*c*), § 2, is approximately true by computing the ordinates of the curve, $y = x^6\phi(x)$, at intervals of 0.5 from $x = 0$ to $x = 4$, and then finding the area of the histogram. Plot.

7. Do the same for the second of these relations.

8. (For students who use the calculus.) Find the mode of $F(x)$ when $\alpha_4 = 3$, and so prove approximately the rule for finding the mode given in Part I, Chapter 3, § 9.

9. Using the relations (*c*), § 2, where necessary, derive the last three equations of (3), § 1.

CHAPTER IV

SAMPLING

1. Nature of the Problem. The question to be investigated in this chapter is: How good is a sample? More precisely: How well does a sample selected from a given larger group, to be called the "universe" or "population," describe that larger group? More precisely still: How nearly does the mean of the sample agree with the mean of the universe? How well do the standard deviation and the higher moments of the sample agree with the standard deviation and higher moments of the universe? How near is the frequency curve, taken as a whole, of the sample to the frequency curve of the universe? Usually it is necessary to make assumptions with regard to the nature and constants of the universe in order to answer these questions. Often indeed it is necessary to require that the universe be normal. Also, it is always necessary to lay down a very important restriction with regard to the sample, and that is that the sample must be a truly random one, chosen from the universe in exactly as random a manner as in theory one chooses random cards from a pack. Since in practice this is a condition which is seldom satisfied, the application of the theory of sampling is manifestly limited. What is really obtained, then, is not an estimate of the error of one's sample but an estimate of what the error would be if the sampling were random. This ideal error must be thought of as smaller than the true error. In a practical case, we know that the error is not less than the one indicated by this theory; we do not know more than this; but to know this much is often of real value. I see before me, on shelf num-

ber 3, a group of 20 mathematical textbooks. Suppose I am interested in their average size. As regards size, they are not truly a random sample from the college library, because, being textbooks, certain book sizes are practically prohibited, and others are specially handy and therefore specially common. So, if I get the average size from this group, I shall not expect it to be as near the average size of all the library books as I should were the books chosen in a truly random fashion. Moreover, a sample which is nearly random with respect to one character may be not at all random with respect to another. If I am interested, not in the sizes of books, but in the number of times the letter x occurs in books, it would manifestly be a very biased answer I should get if I averaged the number of x's in the mathematical texts on this shelf.

We now consider a few elementary theorems of sampling theory. Our view will be limited to the sort of random samples that may be obtained by drawings *with replacements*. This means that, if for example a sample of 5 cards is drawn from a pack of 52, each card is drawn individually, its character noted, and then it is replaced in the pack before the next is drawn. Another way of stating this condition is to say that the pack contains an *infinite* number of cards and that the 5 cards are drawn all at once.

2. Mean of a Sample. Theorem I. *If a sample of size N be drawn from an infinite universe, and if $\bar{t}$ be the mean of the sample, then the mean of all possible such means equals the mean $\bar{t}$ [1] of the universe.*

The proof depends first of all on a clear understanding of what the theorem means. We have in mind something like a correlation table, except that it is infinitely extended in one direction. Across the top we place the probability (relative frequency) distribution of the universe: $p(a) \cdots p(t) \cdots p(b)$. Then each succeeding horizontal row represents

[1] The symbol $\bar{t}$ is to be read: *t curl*.

the relative frequency distribution of a sample of size N, drawn from this universe. As there are an infinite number of such samples that may be drawn, conceivably, we must think of an infinite number of such horizontal rows. This table is indicated below. A numerical illustration, in which the number of horizontal rows is fifty, is given in Example 1. Frequent reference to this numerical illustration, and especially to the figure accompanying it, will help the student to follow this demonstration.

THE UNIVERSE AND ITS SAMPLES (*Relative frequency distributions*)							MEANS
t		a		t		b	
Universe		$p(a)$		$p(t)$		$p(b)$	$\bar{t}$
SAMPLES	1^{st}	$\frac{f(a, 1)}{N}$		$\frac{f(t, 1)}{N}$		$\frac{f(b, 1)}{N}$	$\bar{t}_1$
	2^{nd}	$\frac{f(a, 2)}{N}$		$\frac{f(t, 2)}{N}$		$\frac{f(b, 2)}{N}$	$\bar{t}_2$
	etc.						
	i^{th}	$\frac{f(a, i)}{N}$		$\frac{f(t, i)}{N}$		$\frac{f(b, i)}{N}$	$\bar{t}_i$

In this table we are supposing that t ranges from a to b. The notation $f(t, i)$ means, as in a correlation table, the frequency at the point whose abscissa is t and whose ordinate is i. Also the marginal totals, $p(a)$, $\cdots$, $p(t)$, $\cdots$, $p(b)$,

although placed at the top to indicate the universe from which the successive samples beneath are drawn, really play the same rôle as the marginal totals, called $f(x)$, placed at the bottom of a correlation table; $p(a)$ is not, however, the actual sum of the infinite number of relative frequencies in the first column, being a relative and not a total frequency, but it is proportional to this sum; more precisely, as we shall now prove, it is the limit of the mean of s of these frequencies:

$$p(a) = \lim_{s=\infty} \frac{1}{s} \sum_{i=1}^{s} \frac{f(a, i)}{N}, \tag{1}$$

and, in general,

$$p(t) = \lim_{s=\infty} \frac{1}{s} \sum_{i=1}^{s} \frac{f(t, i)}{N}. \tag{2}$$

Note that in this expression, t remains fixed; we stay in one column. The result follows immediately from the limit definition of probability, for:

$$\frac{1}{sN} \sum_{i=1}^{s} f(t, i)$$

is the relative frequency with which t has occurred in drawing sN individuals from the universe:

$$\frac{1}{sN} \sum_{i=1}^{s} f(t, i) = \frac{\text{total number of } t\text{'s drawn}}{\text{total number of individuals drawn}}.$$

Hence, its limit is $p(t)$, the relative frequency with which t occurs in the universe.

This established, we can soon see the truth of the theorem itself. Let us now look at the last column of our diagram. This is a succession of t's, $\bar{t}_1, \bar{t}_2, \cdots, \bar{t}_i, \cdots$, an infinite number (not all different) of means of a corresponding, infinite number of samples; some of these $\bar{t}$'s are greater than the mean $\bar{t}$ of the universe, some less (cf. Example 1). Of course, any finite number of them could be arranged in a frequency distribution, or in a relative frequency or probability distribution, and we can even think of the probability distribution of the total infinite number if we imagine a limiting

distribution which will be approached as one includes a larger and larger number of samples. This limiting probability distribution of the $\bar{t}$'s we shall call the "curve of means" and shall designate it by $g(\bar{t})$. (Cf. figure with Example 1.) Now our theorem says that this curve of means has itself a mean value which is $\bar{t}$, the mean of the universe. We must think therefore of this infinite set of $\bar{t}$'s as clustering about the value $\bar{t}$ in the sense that

$$\lim_{s=\infty} \frac{1}{s} \sum_{i=1}^{s} \bar{t}_i = \bar{t}. \tag{3}$$

This is what it remains to prove. We get it from (2), thus:

Since
$$\bar{t} = \sum_t t p(t), \tag{4}$$

and
$$\bar{t}_i = \frac{1}{N} \sum_t t f(t, i),$$

$$\lim_{s=\infty} \frac{1}{s} \sum_{i=1}^{s} \bar{t}_i = \lim_{s=\infty} \frac{1}{s} \sum_{i=1}^{s} \frac{1}{N} \sum_t t f(t, i) = \sum_t t \cdot \lim_{s=\infty} \frac{1}{s} \sum_{i=1}^{s} f(t, i) \frac{1}{N}$$
$$= \sum_t t p(t) = \bar{t}.$$

Notation. In the proof of Theorem I, we have defined $g(\bar{t})$, the probability distribution of the curve of means, and we have actually found that its mean was $\bar{t}$. It has also, of course, a standard deviation and other moments, which we shall study. We shall use the following notation for the several moments:

	Universe	Curve of Means	The i^{th} Sample
Probability Distribution...	$p(t)$	$g(\bar{t})$	$\frac{f(t, i)}{N}$
Mean.....	$\bar{t}$	$\bar{t}$	$\bar{t}_i$
Standard Deviation.....	$\bar{\sigma}$	$\bar{\sigma}$	σ
α's.....	$\bar{\alpha}_3$, $\bar{\alpha}_4$	$\bar{\alpha}_3$, $\bar{\alpha}_4$	α_3, α_4

Theorem [1] II. $\bar{\sigma}^2 = \frac{1}{N}\tilde{\sigma}^2$, $\bar{\alpha}_3 = \frac{1}{\sqrt{N}}\tilde{\alpha}_3$, $\bar{\alpha}_4 - 3 = \frac{\tilde{\alpha}_4 - 3}{N}$.

A rigorous proof of this theorem requires the development of several other theorems in sampling by methods similar to those used in Theorem I, and this will not be attempted here. Instead, we shall use a numerical illustration to help us to appreciate further its meaning, and then proceed with the applications.

Example 1. Suppose the probability distribution of the universe to be of a very simple type:

t	0	1	2	3
$p(t)$	$\frac{3}{27}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{6}{27}$
i.e.	.0370	.2222	.4444	.2963

Let us draw from it samples numbering 27 each. This can be understood easily if one thinks of the universe initially as composed of an infinity of data distributed in the proportions indicated. An actual sampling [2] of this universe gave the 50 samples on page 246.

[1] The last two equations of this theorem indicate that the curve of means, $g(\bar{t})$, is more nearly normal than the universe. This is the fact, for similar relations can be proved for all the higher moments. As N becomes infinite, g approaches the normal curve as a limit.

We have already made use of the first of these equations in the theory of errors, Part I, Chapter V, page 86, where we stated that the probable error of the mean was $1/\sqrt{N}$ times the probable error of the given group of observations.

[2] By the use of Tippett's *Random Sampling Numbers*. This is an excellent source from which actually random samples from any desired universe may be obtained.

50 SAMPLES

	Universe and Its Samples (Relative Frequencies)*				*Means*
t	0	1	2	3	
*Universe**	1	6	12	8	$\bar{t} = 2$
25 SAMPLES	1	4	11	11	2.185
	0	5	14	8	2.111
	1	4	11	11	2.185
	1	4	11	11	2.185
	1	4	15	7	2.037
	0	9	12	6	1.889
	1	6	16	4	1.852
	0	4	14	9	2.185
	2	6	14	5	1.815
	2	10	10	5	1.667
	0	7	9	11	2.148
	2	5	10	10	2.037
	1	4	12	10	2.148
	1	4	14	8	2.074
	0	11	7	9	1.926
	2	5	13	7	1.926
	2	7	10	8	1.889
	1	5	16	5	1.926
	0	7	10	10	2.111
	2	6	11	8	1.926
	3	6	10	8	1.852
	3	4	15	5	1.815
	0	8	10	9	2.037
	2	10	10	5	1.667
	1	9	10	7	1.852

* For convenience in printing, the common denominator (27) of the relative frequencies is omitted throughout.

	Universe and Its Samples (Relative Frequencies)*				*Means*
t	0	1	2	3	
*Universe**	1	6	12	8	$\bar{t} = 2$
25 SAMPLES	1	4	15	7	2.037
	0	8	7	12	2.148
	2	6	12	7	1.889
	1	6	14	6	1.926
	1	5	13	8	2.037
	0	4	17	6	2.074
	0	2	16	9	2.259
	2	7	9	9	1.926
	2	4	14	7	1.963
	1	6	16	4	1.852
	1	5	18	3	1.852
	0	6	11	10	2.148
	1	7	13	6	1.889
	0	4	12	11	2.259
	1	6	12	8	2.000
	1	4	15	7	2.037
	0	7	13	7	2.000
	0	4	15	8	2.148
	0	6	14	7	2.037
	2	7	8	10	1.963
	1	3	10	13	2.296
	0	7	11	9	2.074
	1	9	13	4	1.741
	0	9	9	9	2.000
	0	7	10	10	2.111
Totals	47	297	612	394	

It is to be noticed first that the proportions in the totals of the several columns are near to the proportions in the universe. This was expected. In fact, this set of totals may be looked upon as a single sample containing $50 \times 27 = 1350$ individuals.

PROPORTIONS

t	0	1	2	3
Universe.....	.037	.222	.444	.296
Totals.....	.035	.220	.453	.292

Next we look at the column of means ($\bar{t}$'s) at the extreme right of the two tables on page 246. This is the frequency distribution whose form approximates the "curve of means." It becomes the curve of means if an infinite number of samples is taken and their relative distribution depicted. The approximation may be summarized as follows:

It has the following constants: mean = 2.00222, standard deviation = .1466, $\alpha_3 = -.1768$, $\alpha_4 = 2.5584$.

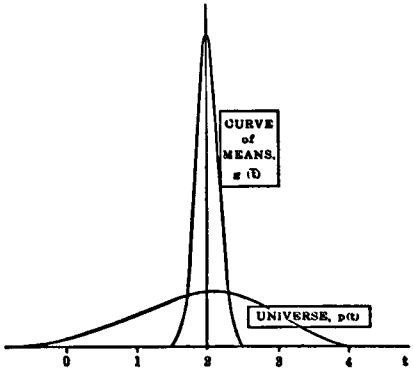
Since the corresponding constants for the universe are:

$\bar{t} = 2.00$, $\tilde{\alpha}_3 = -.4081$,
$\tilde{\sigma} = .817$, $\tilde{\alpha}_4 = 2.5051$,

we know from the theorem that the constants for the true curve of means to which the distribution of $\bar{t}$'s was an approximation are: $\bar{t} = 2.00$, $\bar{\tilde{\sigma}} = .157$, $\bar{\alpha}_3 = -.0785$, $\bar{\alpha}_4 = 2.9817$. In the figure we have graphs of this universe $p(t)$ and of the true curve of means $g(\bar{t})$. See also[1] Problem 5.

DISTRIBUTION OF $\bar{t}$'s

$\bar{t}$	f
1.667	2
1.741	1
1.815	2
1.852	5
1.889	4
1.926	6
1.963	2
2.000	3
2.037	7
2.074	3
2.111	3
2.148	5
2.185	4
2.259	2
2.296	1



F Curves Approximating $p(t)$ and $g(\bar{t})$

[1] The student may feel that the approximations to $\bar{t}$, $\bar{\sigma}$, and $\bar{\alpha}_4$ are satisfactory but that this is not true of $\bar{\alpha}_3$. It will be found later that all four of these approximations are as close as could be expected.

EXERCISES § 2

1. Ten freshmen are chosen at random from the group in Problem 5*a*, Part I, Chapter I, page 16. Can one tell in advance what the mean weight of these ten men will be? What is its most likely value?

2. Refer to the data of Problem 5*c*, Part I, Chapter I. In a certain district 100 persons died of tuberculosis in a certain year. Could one discover from the ages of these 100 what is the average age of death from tuberculosis in the general population? If not, could one discover anything about the general population? If one knew the average age of death in the general population, could one make any prediction with regard to this sample?

3. Consider the following frequency distribution:

Mid-t	5	10	15
f	200	500	300
Sample	///	//	///

Select from it a nearly [1] random sample of 10 individuals by means of the following device. Look at the first 10 numbers of any column of a tabulated function, *e.g.*, the first column of Table IV. Whenever the last digit is 0 or 1, place a mark in the cell for which $t = 5$; when it is 2, 3, 4, 5, or 6, place a mark in the cell for which $t = 10$; and when it is 7, 8, or 9, place a mark in the cell for which $t = 15$. (The marks actually recorded for the sample above were obtained from the first column of the tabulated function $\phi^{(4)}$ in Table IV.)

Find $\bar{t}$ and $\tilde{t}$. Repeat the process until you have 15 samples. Show by a diagram how your 15 $\bar{t}$'s cluster about $\tilde{t}$. What is their mean?

4. For Exercise 1, how closely would the means cluster about the general mean of the large group?

5. Answer the same question relative to Exercise 2, assuming that σ for the 100 is approximately equal to σ for the general population.

6. For Exercise 3, compute your $\bar{\sigma}$. What is its theoretical value?

[1] To obtain a *truly* random sample use *Tippett's* numbers instead.

3. Applications. We are now ready to apply some of this theory, and for this purpose we use Theorem III, which is really a corollary of Theorem II.

Theorem III. *The probability that a random sample from an infinite universe will have a mean, $\bar{t}$, which will be within any chosen amount δ of the mean $\bar{t}$ of the universe is*

$$P_\delta = 2\int_0^\delta \phi dx + \frac{\tilde{\alpha}_4 - 3}{12N}\phi^{(3)}(\delta),$$

approximately, where δ is expressed in the $\tilde{\sigma}$ unit.

The approximation is usually very close. Very commonly the second term is negligible. If the probability function for the universe is an F curve, the equation is true exactly, not merely approximately.

Proof. The probability in question is by definition the area,

$$P_\delta = \int_{z=-\delta}^{z=\delta} g(\bar{t})d\bar{t}, \quad x = \frac{\bar{t} - \bar{t}}{\tilde{\sigma}}. \tag{5}$$

Now $p(t)$, if an F curve, can be expressed as (page 226, 1*b*)

$$F(x) = \phi(x) - \frac{\tilde{\alpha}_3}{6}\phi^{(3)}(x) + \frac{\tilde{\alpha}_4 - 3}{24}\phi^{(4)}(x),$$

where $x = \dfrac{(t - \bar{t})}{\tilde{\sigma}}$. By Theorem II, the corresponding Gram-Charlier series for $g(\bar{t})$ is

$$\phi(x) - \frac{\tilde{\alpha}_3}{6\sqrt{N}}\phi^{(3)}(x) + \frac{\tilde{\alpha}_4 - 3}{24N}\phi^{(4)}(x) + \cdots, \tag{6}$$

where $x = \dfrac{(\bar{t} - \bar{t})}{\tilde{\sigma}}$. Moreover, the terms not printed in (6) are all zero because the higher c's of $g(\bar{t})$ vanish[1] if the higher c's of $p(t)$ do. If $p(t)$ is not an F curve, the terms involving the higher c's of the Gram-Charlier series for $p(t)$ should be considered, but usually they are small and the corresponding

[1] This was not proved but it was implied in the preceding footnote.

terms in (6) become much smaller, so that in almost all cases (6) is, as it stands, at least a very good approximation. Insert (6) in (5) and we have

$$P_\delta = \int_{-\delta}^{\delta} \phi(x)dx - \frac{\tilde{\alpha}_3}{6\sqrt{N}}[\phi^{(2)}(\delta) - \phi^{(2)}(-\delta)] + \frac{\tilde{\alpha}_4 - 3}{24N}[\phi^{(3)}(\delta) - \phi^{(3)}(-\delta)],$$

by (*a*), page 227. Since, by the definition of the ϕ's, $\phi^{(2)}(-\delta) = \phi^{(2)}(\delta)$, and $\phi^{(3)}(-\delta) = -\phi^{(3)}(\delta)$, the second term on the right of the above equation vanishes, and the last term can be compressed, so that P_δ is as stated in the theorem.

Example 2. The following series of deaths of a certain group of women is taken from Elderton's *Frequency Curves and Correlation:*

Ages	*Deaths*
30–34	1
35–39	5
40–44	8
45–49	12
50–54	28
55–59	82
60–64	128
65–69	253
70–74	342
75–79	525
80–84	438
85–89	265
90–94	53
95–99	18
100–104	4
Total	2162

Here, $\bar{t} = 75.98$ years,
corrected $\tilde{\sigma}_t = 1.89 \times 5 = 9.45$,
corrected $\tilde{\alpha}_3 = 0.704$,
corrected $(\tilde{\alpha}_4 - 3) = 0.996$.

Suppose that, from an infinite group similar to this one, an insurance company had a random 100 persons among its policy holders. What

is the likelihood that the mean age of death in the group of 100 would differ from $\bar{t}$ by as much as 1 year?

Here δ is 1 year, but it must be expressed in the $\bar{\sigma}$ unit, *i.e.*, $\delta = \frac{1}{\bar{\sigma}}$. By Theorem II, $\bar{\sigma} = \frac{9.45}{\sqrt{100}}$. So $\delta = 1.058$.

$$P_\delta = 2\int_0^{1.058} \phi(x)dx + \frac{.996}{(12)(100)}\phi^{(3)}(1.058) = 0.710.$$

This is the probability that the difference would be as little as 1 year. The probability that it would be as much as 1 year is $1 - .710 = 0.290$.

Example 3. The mean age of death of men who are alive at age 20 is, in the United States, 59.13. For the city of Chicago it is 58.98, and in 1910 the male population of age 20 was 24,000. Can the difference between the United States and Chicago be explained on the hypothesis of chance? Assume $\bar{\sigma} = 10$ years, and that the distribution of the universe is near enough to normal to permit the omission of the second term in Theorem V.

Here $\delta = \frac{59.13 - 58.98}{\bar{\sigma}} = \frac{.15}{\bar{\sigma}}$, $\bar{\sigma} = \frac{10}{\sqrt{24000}} = .0645$, $\delta = 2.32$, $P_\delta = 0.98$. Therefore there are only 2 chances in 100 that a deviation as great as this would be obtained in selecting 24,000 young men at random from the whole of the United States. The difference could possibly be explained on the hypothesis of chance, but not without difficulty.

Example 4. A fraternal organization wishes to be very sure that the average age of death in its group of men now aged 20 will not differ from the expected 59.13 years by more than 1 year. By "very sure" it means that P_δ must equal .999 or more. How large should the group be?

Assuming as before that $\bar{\sigma} = 10$; $\bar{\sigma} = \frac{10}{\sqrt{N}}$, $\delta = \frac{1}{\bar{\sigma}} = \frac{\sqrt{N}}{10}$, $.999 = 2\int_0^\delta \phi(x)dx$; $.4995 = \int_0^\delta \phi(x)dx$. Coming out of the tables at .9995, we find $\delta = 3.29 = \frac{\sqrt{N}}{10}$. So $N = 1082$.

In all the examples just considered, the constants of the universe have been given. When we do not know the constants of the universe, except in so far as they are indicated approximately by the constants of the sample, our results are less certain. This is a situation which in practice is very common, and is illustrated in Example 5.

Example 5. How surely is the mean height of cornstalks, as given by the sample of Example 3, § 1, Chapter III, page 230, within .1 ft. of the mean height of the whole field?

We have $N = 217$, $\bar{l} = 7.090$ ft., $\sigma = 1.113$, $\alpha_4 = 3.71$. If we assume that $\bar{\sigma} = 1.113$, and that $\bar{\alpha}_4 = 3.71$, we can complete our problem as in Example 1. But what more right have we to assume that $\bar{\sigma} = \sigma$, and that $\bar{\alpha}_4 = \alpha_4$ than that $\bar{l} = \bar{l}$, which is the very point at issue? A satisfactory answer to this question could be given only after an investigation [1] of the fluctuations of σ and of α_4, from sample to sample, similar to the investigation we have just made of the fluctuations of the mean from sample to sample; but we can give a partial answer now. Even if $\bar{\alpha}_4$ differs from α_4 considerably, the coefficient $(\bar{\alpha}_4 - 3)/12N$ will not differ much from $(\alpha_4 - 3)/12N$, and so the effect of this error on P_δ will be slight. Also, if $\bar{\sigma}$ differs considerably from σ, the expression for $\bar{\sigma}$, namely, $\bar{\sigma}/\sqrt{N}$, will not differ much from $\sigma/\sqrt{N}$, provided N is fairly large, and the effect therefore on δ and hence on P_δ will be small. It is important, though, that N be large in order that this be true. Also, the larger N is, the more nearly may we expect σ to equal $\bar{\sigma}$, and α_4 to equal $\bar{\alpha}_4$. On both accounts, therefore, one should not make the assumption of this example in cases where N is small.[2] We now have:

$$\bar{\sigma} = \frac{1.113}{\sqrt{217}} = .0756, \delta = \frac{.1}{\bar{\sigma}} = 1.323, \frac{\alpha_4 - 3}{(12)(217)} = .00027, P_\delta = 0.814.$$

[1] And not perfectly even then, for it would be found that the fluctuation of these constants depended again on the values of $\bar{\sigma}$, $\bar{\alpha}_4$, and other constants of the universe which could only be estimated from the sample.

[2] The case where N is small has been considered by a number of authors. See especially Fisher, R. A., *Statistical Methods for Research Workers*.

Making allowance for errors due to our assumptions, we can hardly be confident of more than the first figure in the result. So finally $P_\delta = 0.8$.

It will have been observed in these examples that, as predicted, the second term of the formula has played a very unimportant rôle, also that the size of δ is all important in the determination of P_δ. Statisticians often say that a deviation δ is significant if it is so large that $P_\delta = .999$ or more. They mean that, if a result has been obtained which differs from the expected result by so much that only once in 1000 times would it have happened by chance, then this result is strikingly different: some special cause should be looked for. The following theorem shows that this will almost always be the case if δ is as large as $3\frac{1}{2}$. Therefore it is commonly said that $\delta = 3\frac{1}{2}$ or more indicates a "significant" difference.

Theorem IV. *P_δ for $\bar{l}$ is almost always .999 or more, if δ is 3.5 or more, and $N \geq 25$.*

Proof. It will very seldom happen that $|\tilde{\alpha}_4 - 3| \geq 4$. So let us assume $|\tilde{\alpha}_4 - 3| < 4$. Then

$$P_\delta = 2\int_0^\delta \phi dx \pm \frac{\phi^{(3)}(\delta)}{75}.$$

If $\delta \geq 3.5$, $\phi^{(3)}(\delta)$ is negative and increases from $-.00038$ to zero as δ increases. Also

$$2\int_0^\delta \phi dx \geq 0.99953.$$

Hence, at the least,

$$P_\delta = .99953 - .00038 = 0.99915.$$

EXERCISES § 3

1. In Example 2 suppose the sample group contained 25 persons. (*a*) What would have been the answer? (*b*) What would have been the likelihood that the mean age of death in the group of 25 would have differed from $\bar{l}$ by as much as 2 years? (*c*) What would have been the probability (approximately) that it would have exceeded 77 years? *Ans.*, .595, .289, .295.

2. Answer the question of Example 3 for New York City, in which the mean age was 58.68, and the male population at age 20 numbered 49,000. *Ans.*, No.

3. Answer the question of Example 4 when a deviation of 2 years is allowed and a P_δ of .995. *Ans.*, 198.

4. Fill in the blanks: (*a*) P_δ for $\bar{t}$ is almost always .995 or more if δ is or more. (*b*) P_δ for $\bar{t}$ is almost always .99 or more if δ is or more. (*c*) P_δ for $\bar{t}$ is almost always or more if δ is 4 or more.

4. Moments of a Sample. In § 2 we answered as well as we could the first of the specific questions proposed at the outset of this chapter: How nearly does the mean of the sample agree with the mean of the universe? We found the curve which would tell us how the mean fluctuated, and for this curve we found the standard deviation, etc. If now we try to ask the analogous question about the other moments, we should, for each of these, find the curve which would describe its fluctuations, and for this curve we should find the standard deviation, etc. These problems have been partly worked out, but they are much more complicated than the first one, although results that are approximately true may be expressed in simple form. We summarize a few of them.

Theorem V. *If the universe is nearly normal and N is large:*

(*a*) *The mean of the curve of σ's is $\bar{\sigma}$ and its standard deviation is* $\sigma_\sigma = \dfrac{\bar{\sigma}}{\sqrt{2N}}$.

(*b*) *The mean of the curve of α_3's is $\bar{\alpha}_3$ and its standard deviation is* $\sigma_{\alpha_3} = \sqrt{\dfrac{6}{N}}$.

(*c*) *The mean of the curve of α_4's is $\bar{\alpha}_4$ and its standard deviation is* $\sigma_{\alpha_4} = \sqrt{\dfrac{24}{N}}$.

(*d*) *The mean of the curve of medians is* $\tilde{M}$ *and its standard deviation is* $\sigma_M = \sqrt{\frac{\pi}{2N}}\,\tilde{\sigma}$.

In place of the standard deviations of these curves, some authors prefer to use the product of the standard deviations and 0.6745. These products are then defined as the probable errors. In this connection the use of a probable error thus defined would seem to be without value, and ought to be abandoned, but in deference to current usage we state the most important of these probable errors in a corollary.

COROLLARY. *For a sample of size N:*

Probable error of the mean $= .6745\,\frac{\tilde{\sigma}}{\sqrt{N}}$.

Probable error of the standard deviation $= .6745\,\frac{\tilde{\sigma}}{\sqrt{2N}}$.

Probable error of the median $= .6745\sqrt{\frac{\pi}{2N}}\,\tilde{\sigma}$.

Now consider first the curve of σ's, whose mean and σ are given in (*a*). From this we can compute what we have called δ in a problem like the following:

Example 6. In Example 2, what would be the probability that the σ of the sample of 100 would be within 3 years of the $\tilde{\sigma}$ of the universe?

The difference between the σ of the sample and the $\tilde{\sigma}$ of the universe is supposedly 3 years. This is δ when expressed in σ_σ as the unit. By Theorem V (*a*):

$$\sigma_\sigma = \frac{9.45}{\sqrt{200}} = .668, \quad \text{and so} \quad \delta = \frac{3}{.668} = 4.49.$$

We now require P_δ but we cannot find it without knowing the equation of the curve of σ's, and we do not yet know this equation. Nor are we willing to assume that this curve, like the curve of means, will be nearly normal. The same sort of difficulty arises, in a more acute form, in the cases of the higher moment coefficients, and, as

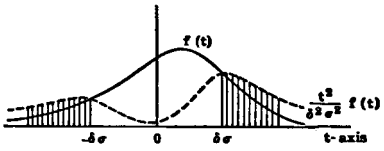
we shall see later, in the case of the coefficient of correlation. Therefore we insert here, parenthetically, some theorems which give approximate values for P_δ even when the precise form of the curve is unknown.

Theorem VI. (*Tchebycheff.*) *For any probability distribution* $f(t)$ *whatever,* $1 - P_\delta \leqq 1/\delta^2$ *where, as before,* $P_\delta = \int_{t=-\delta\sigma}^{\delta\sigma} f(t)dt$, and for convenience the origin is chosen at the mean.

Proof. Since $\int_{-\infty}^{\infty} f(t)dt = 1$,

$$1 - P_\delta = \int_{-\infty}^{-\delta\sigma} f(t)dt + \int_{\delta\sigma}^{\infty} f(t)dt \tag{7}$$

$$\leqq \int_{-\infty}^{-\delta\sigma} \frac{t^2}{\delta^2\sigma^2} f(t)dt + \int_{\delta\sigma}^{\infty} \frac{t^2}{\delta^2\sigma^2} f(t)dt, \tag{8}$$



because $\frac{t^2}{\delta^2\sigma^2} \geqq 1$ in the intervals over which the integrals are extended, and $f(t)$ is essentially positive. But (8) is less than

$$\int_{-\infty}^{\infty} \frac{t^2}{\delta^2\sigma^2} f(t)dt, \tag{9}$$

since, by examining the limits, it appears that a larger interval is involved in (9) than the two intervals of (8) taken together. But (9) equals

$$\frac{1}{\delta^2\sigma^2}\int_{-\infty}^{\infty} t^2 f(t)dt = \frac{\sigma^2}{\delta^2\sigma^2} = \frac{1}{\delta^2}.$$

Example 6*a*. So, in Example 6, one would know, no matter what the curve of σ's was, that

$$1 - P_\delta \leqq .0496, \quad \text{and that} \quad P_\delta \geqq 0.95.$$

Theorem VI is a very interesting one because it is true of all types of distributions. One simply cannot set down a set of positive numbers for which it is not true. (It is interesting to try to do this.) But the theorem has also the faults that go with its virtues. Because it says something which is true of every distribution, it does not give us adequate information about the particular distribution we have before us. Suppose it should happen that our particular distribution were normal. From the tables we should learn that $1 - P_\delta = 0.000{,}0072$. We do not think it is really normal, but we do think it lies somewhere between a normal distribution and the most peculiar types to which Tchebycheff's theorem would apply. There is a great difference between the values, .05 and .000,0072; hence the need of an intermediate theorem which will make some assumptions with regard to our distribution and yet not require it to be a normal one. This is the rôle of the following theorem and corollary.

Theorem VII. *If a distribution is uni-modal and if the mode is within σ of its mean* ($|skewness| \leqq 1$)*:*

$$1 - P_\delta < \frac{\alpha_{2r}}{\left(\frac{\delta + 2r\delta}{2r}\right)^{2r}}, \textit{ plus a small amount,}^1$$

where r is any positive integer.

COROLLARY 1. If $r = 1$, $1 - P_\delta < \frac{1}{2.25\delta^2}$.

Example 6b. So in Example 6, one would now conclude that $1 - P < 0.022$. This is an improvement, but still differs so much from .000,0072 that it seems almost certain that it would be better to increase our assumption still further than to put up with a number as large as 0.022. This is done in the following corollary.

[1] This amount is negligible and will be omitted in the applications given here. A formula and table for it were given by the author in the paper in which he announced this theorem, *Bulletin American Mathematical Society*, vol. 28, pp. 427–432. See also *Biometrika*, vol. 15, p. 253, for a correction and further tables.

COROLLARY 2. *If the distribution is either an F curve, or one of Pearson's fundamental types, with customary values of* α_3, α_4 ($|\alpha_3| < \sqrt{1.5}$, $2 \leqq \alpha_4 \leqq 4$), *then* $1 - P_\delta$ *is less than the maximum values given in the table below:*

δ	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Max. $(1-P_\delta)$	.20	.11	.05	.02	.01	.005	.002	.001

Example 6c. So in Example 6, we may now say that $1 - P_\delta < 0.002$. This is still much larger than the normal value but small enough to use in drawing practical conclusions. Moreover, even if we did know that a distribution was very nearly normal, we would not dare conclude that $1 - P_\delta$ was as small as .000,0072 when $\delta = 4.5$, for even a slight deviation from normality might affect this result very severely. We must not talk about these exceedingly small probabilities unless we are sure we have the conditions which lead to them exactly fulfilled.

Proof of Corollary 2. The proof for the case of Pearson's curves depends on a relation between their high and low moments. We have referred to these curves but have not discussed them. Therefore the proof will be given only for the F curve.

For the F curve the c's of higher index than 4 vanish. Had the higher ones been written, it would have been seen that

$$c_6 = \frac{\alpha_6 - 15\alpha_4 + 30}{\underline{|6}}, \quad c_8 = \frac{\alpha_8 - 28\alpha_6 + 210\alpha_4 - 315}{\underline{|8}}.$$

By hypothesis,

$$2 \leqq \alpha_4 \leqq 4, \text{ and } c_6 = c_8 = 0.$$

Therefore,

$$\alpha_6 + 30 = 15\alpha_4 < 60; \quad \alpha_6 < 30.$$

Also, $\alpha_8 + 210\alpha_4 = 28\alpha_6 + 315 < 840 + 315 = 1155;$

$$\alpha_8 < 1155 - 2(210) = 735.$$

Consider now the case where $\delta = 4$. By the theorem, $1 - P_\delta < \frac{\alpha_{2r}}{4^{2r}\left(\frac{2r+1}{2r}\right)^{2r}}$, and in this we are free to choose r as any positive integer. We shall choose that integer, by trial, which will yield the smallest number on the right of this inequality. Let us first try $r = 2$. We get $1 - P_\delta < \frac{4}{(4^4)(2.44)} = 0.064$. Try $r = 3$. $1 - P_\delta < \frac{30}{(4^6)(2.52)} = 0.0029$. This is smaller than the number in the table, and therefore that inequality is proved. A still smaller value would be obtained in this instance by using $r = 4$, but it is not necessary, since in our table we are not differentiating between the values obtainable for F curves and those obtainable for Pearson's curves.

The values at the other places in the table are provable by the same method. Just as in the case of the normal curve, we are supposing in this proof that the given curve is exactly an F curve or exactly a Pearson curve. If it deviates a little, we cannot be sure that this corollary will apply, and one might prefer to fall back on Corollary 1, or even on Tchebycheff's theorem. Our risk is not great, however, so long as we do not extend the table so as to obtain very small probabilities. This is one reason why we have not separated the two cases, the Pearson curves and the F curves, and obtained much smaller quantities in the latter case. For this sort of thing the forms of the tails of a curve are very important. Now, some of Pearson's curves are limited. They do not extend indefinitely to the right or left as do the normal and the F curves. Therefore for *this* purpose they are to be the more relied on as approximations to the distributions we shall have in practice.

Example 7 The following 100 data were actually drawn at random [1] from a universe in which the distribution of frequency was described by the successive terms of the point binomial $(\frac{1}{4} + \frac{3}{4})^7$ of Example 1, Chapter II, page 208. For each of the quantities, mean, median, σ, α_3, α_4, of this sample, find δ and the corresponding probability $1 - P_\delta$.

t	0	1	2	3	4	5	6	7
Proportions in Universe	$\frac{1}{4^7}$	$\frac{21}{4^7}$	$\frac{189}{4^7}$	$\frac{945}{4^7}$	$\frac{2835}{4^7}$	$\frac{5103}{4^7}$	$\frac{5103}{4^7}$	$\frac{2187}{4^7}$
Sample Frequencies	0	0	0	4	19	35	26	16

For the sample: $\bar{t} = 5.31$, $\sigma = 1.074$, $\alpha_3 = -.059$, $\alpha_4 = 2.30$.
For the universe: $\bar{t} = 5.25$, $\bar{\sigma} = 1.146$, $\bar{\alpha}_3 = -.436$, $\bar{\alpha}_4 = 2.90$.

For $\bar{t}$: $\sigma_{\bar{t}} = \dfrac{1.146}{\sqrt{100}} = .1146$, and so $\delta = \dfrac{.31 - .25}{.1146} = 0.54$.

By Theorem III, $1 - P_\delta = 0.41$.

For σ: $\sigma_\sigma = \dfrac{1.146}{\sqrt{200}} = .0810$, and so $\delta = \dfrac{1.146 - 1.074}{.0810} = 0.89$.

By Theorem VII, Corollary 1, $1 - P_\delta < 0.56$.

For α_3: $\sigma_{\alpha_3} = .245$, and so $\delta = \dfrac{.436 - .059}{.245} = 1.5$.

By Theorem VII, Corollary 2, $1 - P_\delta < 0.2$.

For α_4: $\sigma_{\alpha_4} = .490$, and so $\delta = \dfrac{2.90 - 2.30}{.490} = 1.22$.

By Theorem VII, Corollary 1, $1 - P_\delta < 0.30$.

For M: The median of the universe is $t = 5.32$, and the median of the sample is $t = 5.27$. Since $\sigma_M = \sqrt{\dfrac{\pi}{200}}\,(1.146) = .1436$, $\delta = \dfrac{5.32 - 5.27}{1.436} = .35$, and this is so small that Theorem VII is not applicable and not needed.

In no one of these cases is δ large enough to cause surprise; *i.e.*, in no case is $1 - P_\delta$ so small that a δ as large as the one found would appear quite unlikely. The largest values of δ occur in the cases of

[1] Obtained by using *Tippett's* numbers.

α_3 and α_4, but they must be regarded as inaccurate; for in these cases the values of the standard deviations as given by Theorem V are not close except when N is larger and the universe more nearly normal than in the present case.

EXERCISES § 4

1. Plot the figures in the table of Corollary 2, and estimate the value of max $(1 - P_\delta)$ at $\delta = 2.7$. (It is better to use ratio paper.) *Ans.*, .034.

2. (*a*) Using this graph, find max $(1-P_\delta)$ for the point binomial $(\frac{1}{2} + \frac{1}{2})^9$ when $\delta = 1\frac{1}{2}$, $2\frac{1}{2}$, and 4. (*b*) What are the true values of $(1 - P_\delta)$ in these cases? (*c*) What are the values given by the normal law? *Ans.* when $\delta = 2\frac{1}{2}$; .036, .0039, .0076.

3. Compute $(1 - P_\delta)$ exactly for a rectangular distribution (f = constant from $x = -a$ to $x = a$), given that $\sigma^2 = a^2/3$. Take $\delta = 1.5$, 2, and 3. Compare your results with those given by the table.

4. Do the same for the distribution of cornstalks (Example 5). *Ans.* when $\delta = 2$; .067, .11.

5. Coefficient of Correlation. So far we have thought of our sample as taken from a one-way frequency distribution, but this was not necessary. Our universe may be a frequency distribution of any number of dimensions. In particular, it may be a two-way distribution representable by a correlation table. An important question to be answered when one has a sample from a two-way table is: How reliable is the observed coefficient of correlation? That is, how close may it be expected to be to the coefficient pertaining to the universe? We treat this problem like those in § 3; we need to find the curve of r's, its mean and standard deviation. The partial results are as follows:

Theorem VIII. *If $\bar{r}$ applies to the universe and r to the sample, the mean of the curve of r's is $\bar{r}$, its standard deviation is* $\sigma_r = \dfrac{(1 - \bar{r}^2)}{\sqrt{N}}$, *and therefore its so-called probable error is*

$.6745\sigma_r = \dfrac{.6745(1-\bar{r}^2)}{\sqrt{N}}$, *approximately. The approximation* [1] *is good only if the universe is nearly normal and N is large.*

Rather thorough investigations have been made of the form of the curve of r's, but we shall not attempt to reproduce them here; instead we shall solve our probability problems by means of Corollary 2 of Theorem VII, assuming merely that we are dealing either with an F curve or with one of Pearson's fundamental types.

Example 8. A truly random sample of 400 was drawn from a universe (*Gavett*) in which the length and breadth of leaves were displayed, the unit being a millimeter and $\bar{r}$ being 0.61. Compute δ and $1 - P_\delta$ for the coefficient of correlation.

400 LEAVES. A RANDOM SAMPLE

Length \ Breadth	16 –	19 –	22 –	25 –	28 –	31 –	34 –
22 –	6	7	6				
27 –		5	19	28	8		
32 –		7	21	36	24	19	
37 –			7	33	24	21	
42 –			7	21	32	12	
47 –				11	18	9	8
52 –						9	2

[1] The closer approximation,

$$\sigma_r = \frac{1-\bar{r}^2}{\sqrt{N}}\sqrt{1 - \frac{\bar{r}^2}{1-\bar{r}^2}\left(\frac{{}_x\alpha_4 - 3 + {}_y\alpha_4 - 3}{4}\right)},$$

clearly fails in such a case as the following: ${}_x\alpha_4 = {}_y\alpha_4 = 4$, $\bar{r}^2 > \frac{1}{2}$, for it becomes imaginary. It is a common error to rely on the formula of Theorem VIII in cases where this usually closer approximation fails. In a large portion of practical cases the formula is quite unreliable, and can hardly be thought of as giving us more than a general idea of the order of size of σ_r; and it is then both misleading and a waste of time to compute it to any high degree of refinement. Two decimal places are usually enough.

For the universe: $\rho = .61$, and so $\sigma_r = 0.0314$.

For this sample: $r = .565$,

and so

$$\delta = \frac{.045}{.0314} = 1.433, \quad 1 - P_\delta < \frac{1}{(2.25)(1.433)^2} = 0.145.$$

Example 9. A sample is drawn from a normal universe and r is observed to be 0.5. May we be reasonably sure that ρ is truly greater than .4 if our sample numbers 100? if it numbers 900? Let us agree that a probability of .99 represents "reasonable" sureness.

The question has been put in a customary form, but strictly speaking it has no meaning in that form. That is, we should not ask, What is the probability that the universe is thus and so? since we do not have, theoretically, one sample and a number of universes, but instead one universe and a number of samples.[1] So we shall rephrase the question: If $\rho \leq .4$, is the probability as small as .01 that one would obtain a sample for which r would differ from ρ by as much as .5 does? If so, then we shall say we are reasonably sure that we were sampling in a universe[2] in which ρ was greater than 0.4.

So, let $\rho = .4$ and $N = 100$ and compute $1 - P_\delta$: $\sigma_r = .84/10 = .084$, $\delta = .1/.084 = 1.19$. By Corollary 1 of page 257, $1 - P_\delta < 0.32$. Evidently N is not large enough. We wanted $1 - P_\delta$ to be less than 0.01.

So, again let $\rho = .4$ and $N = 900$ and compute $1 - P_\delta$: $\sigma_r = .028$, $\delta = .1/.028 = 3.57$. By the table of Corollary 2, $1 - P_\delta < .01$, and so this value of N is satisfactory.

Exercise § 5: (*a*) Repeat the whole of Example 9, supposing r is observed to be .45 and $N = 1200$. (*b*) How great must N be in this case to make $1 - P_\delta < .005$?

[1] This brings up the question of inverse probability which is too difficult to discuss here. A meaning might be attached to the given question if one were willing to make certain assumptions.

[2] In this instance, as in like cases before, we are essentially assuming that what was *a priori* very unlikely did not actually happen. All statistical influence — which includes most of our knowledge — is based on this same assumption.

6. Chi Test. The student may well ask at this point: How about the first question of this chapter? How reliable is a sample? We have had theorems bearing on the questions, how reliable is the mean, how reliable is the standard deviation, and how reliable are certain other characteristics of a sample; but is there no way of estimating how reliable the sample is as a whole? This question was studied by K. Pearson in 1900 and the result was the now well-known chi test. The problem is not a very important one except in advanced statistics, and so we shall not give here a complete exposition of it. But, though not very important for us, it is interesting and the essential result can be stated quite simply. The question to be answered in the simplest case may be put thus:

Suppose a frequency table describing a universe (supposedly infinite) be divided into m cells. (We may have in mind here a tabulated frequency curve, or a tabulated correlation surface, or even a multiple correlation solid.[1]) Suppose a random sample be drawn from the universe. What is the probability, P, that, on the whole, the deviations between the expected frequencies and those observed will be as great as a given amount to be called [2] χ^2? The answer is given by a formula whose approximate values have been tabulated.[3] The quantity χ^2 is found as follows: Let p_i be the relative frequency that appears in the ith cell in the universe, N the size of the sample, f_i the absolute frequency that appears in the ith cell of the sample:

$$\chi^2 = \sum_{i=1}^{m} \frac{(f_i - Np_i)^2}{Np_i}. \tag{10}$$

[1] To be explained in a later chapter.

[2] The Greek letter *chi* is written χ. Its square is used merely because, as we shall see, the quantity for which χ^2 is to stand is essentially positive.

[3] *Tables for Statisticians and Biometricians* by K. Pearson. The conditions under which Pearson showed that his values were correct cannot be given in detail here (see a discussion by the author in *Transactions of the American Mathematical Society*, vol. 31 (1929), pp. 133–

Pearson's tables are too extensive to be repeated here, but a general idea of them may be obtained from the following short summary.

χ TEST. VALUES OF P

χ^2 \ m	3	4	5	6	7	8	9	10	11	12	13	14
2	.37	.57	.74	.85	.92	.96	.981	.991	.996	.998	.999	.9998
4	.14	.26	.41	.55	.68	.78	.86	.911	.947	.970	.983	.991
6	.050	.11	.20	.31	.42	.54	.65	.740	.815	.873	.916	.946
8	.018	.046	.09	.16	.24	.33	.43	.534	.629	.713	.785	.844
10	.007	.019	.04	.08	.12	.19	.27	.350	.440	.530	.616	.694
15	.0006	.0018	.0047	.0104	.0203	.0360	.0591	.0909	.1321	.1825	.2414	.3074
20	.0000	.0002	.0005	.0013	.0028	.0056	.0103	.0179	.0293	.0453	.0671	.0952
25	.0000	.0000	.0001	.0001	.0003	.0008	.0016	.0030	.0053	.0091	.0148	.0231
30	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004	.0009	.0016	.0028	.0047
40	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

Example 10. How improbable, taken as a whole, is the sample in Example 7?

Expected f, Np	*Observed f*	$\lvert Np - f \rvert$	$\frac{(Np-f)^2}{Np}$
.006	0	0	0
.128	0	.1	.08
1.154	0	1.2	1.25
5.77	4	1.8	.56
17.3	19	1.7	.17
31.1	35	3.9	.49
31.1	26	5.1	.83
13.4	16	2.6	.50
Totals	$100 = N$		$3.88 = \chi^2$

144), but in general it may be said that his tables cannot be guaranteed for cases where there exist cells such that $(f - Np)^2/Np$ is larger than 6, or where $f < 3$, or $m > 20$. In general, too, the very small values of the probabilities given in certain parts of his tables are unreliable. Usually, if we lump together small frequencies near the limits of a distribution, the required conditions will be satisfied, at least well enough to permit of an approximate solution.

Since m, the number of intervals or cells, is 8, we learn from the table [1] that P is about 0.8. Thus one would expect to get a sample which would be as far from perfect as this one 80% of the time, and so there is certainly no reason for suspecting that it was not a truly random sample, as it was claimed to be.

As another illustration we choose a famous dice problem.

Example 11. (*Weldon's Dice.*) Twelve dice were thrown 26,306 times, and those which fell so that either 5 or 6 points were uppermost were counted. The number of throws in which no such dice appeared was $f_0 = 185$; the number of throws in which one appeared was $f_1 = 1149$, etc., as indicated in the table below. The number of times either 10, 11, or 12 such dice appeared was 18. These last three cases are lumped together, because, as stated in a preceding footnote, the method is not good when some of the frequencies are small.

Dice	f	Np	$f - Np$
0	185	203	− 18
1	1149	1217	− 68
2	3265	3345	− 80
3	5475	5576	− 101
4	6114	6273	− 159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10, 11, 12	18	14	4

$\chi^2 = \Sigma \dfrac{(f - Np^2)}{Np} = 35.94; m = 11$. By our table, $P < 0.0009$.

By Pearson's Tables, $P = 0.0001$.

[1] In a preceding footnote it was stated that we should not apply the table to cases where $f < 3$. Hence, it would have been better here to have supposed the problem so stated that the first four intervals were lumped in one, with a total $Np = 7.06$ and $f = 4$. This would have given the results: $\chi^2 = 3.35$, $m = 5$, $P = 0.5$. The effect of this procedure is to substitute a new problem with a more reliable answer

This χ test has been illustrated only in the simplest case: where the universe is known (or assumed) in advance. It should not be applied, without some modification, to cases where the universe is partly inferred from the sample.[1]

7. Significance of a Difference. Sampling theory may be used to provide an answer to a common question with regard to the differences between two samples. Two samples may differ, of course, in many ways, but we shall consider two ways only, and the first is the case in which the means are different. The specific question to be studied here is this: Two samples are drawn at random from a given infinite universe, and the means are found to differ by a certain amount. What is the probability that a difference as great as that observed would happen by chance?

Before we can answer this question we must put and answer a more elementary one: Two individuals are selected, t' and t'', from infinite universes whose probability distributions are respectively $f'(t)$ and $f''(t)$. What is the probability that the numerical value of the difference $|t' - t''|$ will be as large as some previously prescribed amount? For this we may use

Theorem IX. *If f' and f'' have the same mean, the probability distribution, F, of the differences $(t' - t'')$ is a symmetrical one. Its mean is zero and its even moments are related to the moments of f' and f'' thus,*

$$\mu_2(F) = \mu_2(f') + \mu_2(f'').$$
$$\mu_4(F) = \mu_4(f') + 6\mu_2(f')\mu_2(f'') + \mu_4(f''), \text{ etc.}^2$$

---

for the one given, not to procure a better answer to the given problem. I shall not again insist on doing this in this text; and it is much less necessary to do so if one is seeking a value of P correct to only one place, as above.

[1] For a consideration of this case, compare R. A. Fisher, *Statistical Methods for Research Workers;* Irwin, *Journal of the Royal Statistical Society*, vol. 92 (1929), p. 264; Fry, *Probability and Its Engineering Uses*, Chap. 9.

[2] For the general formula see the author's paper in *Journal of American Statistical Association*, vol. 18, p. 976.

The proof of this theorem will not be given. It proceeds in much the same way as does the proof of Theorem II.

COROLLARY 1. (*a*) *If the two universes are identical,* $\alpha_4(F) - 3 = \frac{1}{2}[\alpha_4(f) - 3]$; (*b*) *if not,* $|\alpha_4(F) - 3|$ *is less than the larger of the two quantities* $|\alpha_4(f') - 3|$, $|\alpha_4(f'') - 3|$.

The student is asked to prove this corollary (Problem 30). The next theorem with its corollary will now provide a solution of our more fundamental problem.

Theorem X. *Let* $g(\bar{t})$ *be a probability distribution describing an infinite universe from which two random samples of different sorts are to be drawn, the first to be of size* N' *and the second of size* N''. *In general, there is a difference between the mean values of two such samples. The probability distribution,* F, *of these differences has the following characteristics: Mean* $F = 0$; F *is symmetrical with respect to its ordinate at* 0; $\sigma_F = \bar{\sigma}\sqrt{\frac{1}{N'} + \frac{1}{N''}}$; $\alpha_4(F) - 3$ *is usually negligible.*

Proof. We may use the following notation:

	Universe	1st Sample	2nd Sample	Mean of 1st Sample	Mean of 2nd Sample	Difference
Probability Function ...	$g(\bar{t})$	—	—	$g'(\bar{t})$	$g''(\bar{t})$	F
Mean Value	$\bar{t}$	$\bar{t}'$	$\bar{t}''$	$\bar{t}$	$\bar{t}$	0
Standard Deviation....	$\bar{\sigma}$	σ'	σ''	$\bar{\sigma}'$	$\bar{\sigma}''$	σ
μ_r	$\bar{\mu}_r$	μ_r'	μ_r''	$\bar{\mu}_r'$	$\bar{\mu}_r''$	μ_r

A little reflection will show that $g'(\bar{t})$ and $g''(\bar{t})$ are probability distributions which may play the same rôle as do f' and f'' in Theorem IX, and when they do, $\bar{t}'$ is a single individual drawn from $g'(\bar{t})$, $\bar{t}''$ a single individual drawn from $g''(\bar{t})$. These two "curves of means," as they were called in § 2, are not alike, if $N' \neq N''$, but they do have the same

mean, since, by Theorem I, the mean of each is $\bar{l}$. So, by Theorem IX,

$$\mu_2 = \bar{\mu}_2' + \bar{\mu}_2'', \quad \mu_4 = \bar{\mu}_4' + 6\bar{\mu}_2'\bar{\mu}_2'' + \bar{\mu}_4''.$$

But by Theorem II, $\bar{\mu}_2' = \tilde{\mu}_2/N'$, $\bar{\mu}_2'' = \tilde{\mu}_2/N''$. Hence,

$$\mu_2 = \tilde{\mu}\left(\frac{1}{N'} + \frac{1}{N''}\right), \quad i.e., \quad \sigma = \tilde{\sigma}\sqrt{\frac{1}{N'} + \frac{1}{N''}}.$$

Also, by Theorem II,

$$\bar{\alpha}_4' - 3 = \frac{1}{N'}(\tilde{\alpha}_4 - 3), \; \bar{\alpha}_4'' - 3 = \frac{1}{N''}(\tilde{\alpha}_4 - 3).$$

These quantities were usually negligibly small in numerical value. Therefore $\alpha_4 - 3$, which, by the corollary to Theorem IX, is smaller numerically than the larger of these, must also be negligible.

COROLLARY.[1] *The probability that the difference between the two means will exceed, numerically, an observed number δ times σ is $1 - P_\delta$, where, approximately,*

$$P_\delta = 2\int_0^\delta \phi(x)dx.$$

When $1 - P_\delta$ is quite small (δ large), the observed difference is commonly said to be significant. That is, it means something: it probably did not happen by chance. The idea is illustrated in the following example.

Example 12. The average grades of two college fraternities, numbering 60 and 40 men respectively, were found to be 79.98% and 74.20%. Is the difference significant?

What we would like to know here is whether a difference as large as that observed, 79.98 − 74.20 = 5.78, might have happened had the dean chosen 60 men by lot from the college body[2] to make up the first fraternity, and then 40 to make up the second. The mean

[1] Cf. Problem 31.

[2] Or, better, from an infinite group resembling these two fraternities.

of the universe here is not given and not needed. The standard deviation, $\tilde{\sigma}$, is needed. Let us suppose it given as 10%. We use the theorem to get σ and the corollary to get $1 - P_\delta$:

$$\sigma = 10\sqrt{\frac{1}{60} + \frac{1}{40}} = 2.04, \quad \delta = \frac{5.78}{2.04} = 2.83,$$

$1 - P_\delta = .0047$; a rather significant difference, therefore.

Note: When, as usually happens, $\tilde{\sigma}$ is not given, we shall assume that

$$N\tilde{\sigma}^2 = N'\nu_2' + N''\nu_2'' \text{ (p. 34, Problem 9).}$$

This is the same as taking for $\tilde{\sigma}$ the σ of the combination of the two given samples.

8. Difference between Proportions. The final question to be answered in this chapter is in some respects the most elementary, but it could not have been answered until after most of the preceding theory had been presented. We begin with a numerical example.

Example 13. (Taken from the *United States Census Report for* 1910, for white persons at age $x = 20$ in the registration states.)

	d_x	l_{x+1}	*Totals*, l_x
Males.....	1093	223 635	224 728
Females.....	1029	243 877	244 906
Totals.....	2122	467 512	469 634

From this it follows that the proportion of males who died between ages 20 and 21 was $\frac{1093}{224{,}728} = 0.00486$, and that the proportion of females was $\frac{1029}{244{,}906} = 0.00420$. Does this indicate a real difference in vitality in favor of the young woman or might such a difference have occurred by chance? We shall derive a formula by which this question may be answered, but let us first replace our numerical table by one using letters only.

Characteristic / Sample	α	Not α	Totals
A	$p'N'$	$q'N'$	N'
B	$p''N''$	$q''N''$	N''
Totals	pN	qN	N

Here $p'N'$ is the number of individuals in the sample A who possess the characteristic α, $q'N'$ the number who do not, N' the total number; so that $p' = p'N'/N'$ is the proportion who possess the characteristic and q' is the proportion who do not. It was not necessary, either in the numerical example or in this diagram, to have given all three columns; from any two the third could have been obtained. All that is required is the set of numbers: p', p'', N', N''. We use this notation in the following theorem.

Theorem XI. *If two random samples of sizes N' and N'' are drawn from an infinite universe in which the proportion which has the character α is p, the probability that the difference in the proportions obtained will be as great numerically as the observed difference, $p' - p''$, is $1 - P_\delta$, where, very nearly,*

$$P_\delta = 2\int_0^\delta \phi(x)dx, \quad \delta = \frac{|p' - p''|}{\sigma}, \quad \sigma = \sqrt{pq\left(\frac{1}{N'} + \frac{1}{N''}\right)}.$$

Proof. Let us first consider the distribution obtained if one draws samples of size N' from the universe. The distribution of α's is what we called earlier the distribution of "successes," and the form is that of a point binomial:

$$(p + q)^{N'} = p^{N'} + \cdots + {}_{N'}C_s p^s q^{N'-s} + \cdots + q^{N'}. \quad (11)$$

The general term of (11) gives the probability that in the sample there will be exactly s successes. Another way of stating this is to say that it gives the probability that the *proportion* of successes will be s/N'. Let $x = s/N'$. Then we may say that the general term of

$$\sum_{x=0}^{x=1} {}_{N'}C_{xN'}\, p^{xN'} q^{N'-xN'} \quad (12)$$

represents the probability that the proportion of successes will be x. Another way of expressing this is to say that if the histogram of (11) be constructed with the class interval $1/N'$ instead of 1, and the ordinates be made correspondingly higher, so as to preserve the areas unchanged, we shall have a representation of the distribution of proportions x in samples of size N'. Let us call this distribution $f(x)$. Its mean is at $x = p$, and its standard deviation is

$$\sigma' = \sqrt{\frac{pq}{N'}}.$$

Likewise, if $g(x)$ represents the distribution of samples of size N'', its mean is also p and its standard deviation is $\sigma'' = \sqrt{\frac{pq}{N''}}$. Our theorem now requires us to draw a single individual from f and a single individual from g and to find out whether the difference between them is significant. Clearly, f here plays the rôle of f' in Theorem IX, and g the rôle of f''. Let F again denote the curve of differences and σ, α_4 its constants. We have, then, $\sigma^2 = \sigma'^2 + \sigma''^2 = pq\left(\frac{1}{N'} + \frac{1}{N''}\right)$; $|\alpha_4 - 3|$ is less than the greater of the two quantities $|\alpha_4' - 3|$ and $|\alpha_4'' - 3|$, where these apply to f and g respectively. But f and g are point binomials in which the degrees N', N'' are usually large. Therefore both these quantities are usually very small, as we have seen (Chapter II, Theorem I (*c*), when n is large), and they may be omitted. Theorem XI now follows from Theorem IX.

We may now finish Example 13:

$$p = \frac{2122}{469634} = .00453, q = .99547, \sigma = \sqrt{pq\left(\frac{1}{224728} + \frac{1}{244906}\right)}$$
$$= .0001965, \delta = \frac{|.00486 - .00420|}{.0001965} = 3.36, 1 - P_\delta = 0.00098.$$

It is therefore almost impossible that such a difference should have been due to chance.

EXERCISES § 8

Repeat the example of this section with reference to each of the following sets of data:

1. Men between the ages 20 and 25, New York and Chicago (1910).

	d_s	Totals
Chicago.....	720	126 972
N. Y. C.....	1405	253 249

2. Same, Boston and Philadelphia. *Ans.*, .0023.

	d_s	Totals
Boston.....	154	32 763
Philadelphia...	474	76 027

3. Males and females between the ages 45 and 50, Philadelphia (1910).

	d_s	Totals
Males.....	840	44 354
Females.....	575	44 044

9. **Application to Physical Observations.** In Theorem II we saw that

$$\bar{\sigma}^2 = \frac{\tilde{\sigma}^2}{N}, \tag{13}$$

where $\tilde{\sigma}$ referred to an infinite universe of observations (x), and $\bar{\sigma}$ to the mean $(\bar{x})$ of N observations chosen at random from it. This equation is a special case of a more general formula in which it is supposed that each observation $(x_1, x_2, \cdots, x_N)$ is obtained from a different universe with individual standard deviations $(\sigma_1, \sigma_2, \cdots, \sigma_N)$, and in which $\bar{x}$ is not the simple mean but a weighted mean or any linear combination of the x's:

$$\bar{x} = c_1x_1 + c_2x_2 + \cdots + c_Nx_N.$$

Then $$\bar{\sigma}^2 = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_N^2\sigma_N^2. \tag{14}$$

Exercise. Show that (13) is a special case of (14) for a properly chosen set of c's.

If desired, one may multiply equation (14) through by $(.6745)^2$ and obtain

$$\bar{s}^2 = c_1^2 s_1^2 + \ldots + c_N^2 s_N^2, \tag{14a}$$

where s stands for probable error. These formulae help us to solve problems like the following one.

Example 14. Four sides of a field are measured with probable errors as indicated by the double signs: $80 \pm .01$ feet, $160 \pm .015$, $150 \pm .015$, $300 \pm .02$. Find the probable error of the perimeter.

The perimeter is

$$\bar{x} = 80 + 160 + 150 + 300 = 690 \text{ feet},$$

and its probable error is $\bar{s}$:

$$\bar{s}^2 = (.01)^2 + (.015)^2 + (.015)^2 + (.02)^2; \quad \bar{s} = 0.0308.$$

To obtain further applications we shall now incorporate our formulae in a theorem which has an even broader scope.

Theorem XII. *If $x_1, x_2, \cdots, x_N$ are mutually independent observations, and if $\bar{x}$ is such a function of them that when the definite errors, $\delta_1, \delta_2, \cdots, \delta_N$, are made in the x's the effect on $\bar{x}$ will be an error $\bar{\delta}$ which is a linear combination of the δ's, i.e.,*

$$\bar{\delta} = c_1\delta_1 + \cdots + c_N\delta_N;$$

then the σ of $\bar{x}$ is related to the σ's of the x's as follows:

$$\bar{\sigma}^2 = c_1^2\sigma_1^2 + \cdots + c_N^2\sigma_N^2.$$

Also the probable errors are similarly related:

$$\bar{s}^2 = c_1^2 s_1^2 + \cdots + c_N^2 s_N^2.$$

The student of the calculus has easy methods of finding a close approximation to the expression for the definite error $\bar{\delta}$ if he is given the functional relationship that exists between $\bar{x}$ and the several x's. For example, it is easy for him to show

that if, as on page 273, $\bar{x} = c_1x_1 + \cdots + c_Nx_N$, then $\bar{\delta} = c_1\delta_1 + \cdots + c_N\delta_N$, so that (14) is a special case of this theorem. There are two other cases of prime importance for which we give the results: *Let c and N be any constants.*

If $\bar{x} = cx^N$, *then, approximately,*

$$\frac{\bar{\delta}}{\bar{x}} = N\frac{\delta}{x}.$$

If $\bar{x} = cx_1 \cdot x_2 \cdots x_N$, *then, approximately,*

$$\frac{\bar{\delta}}{\bar{x}} = \frac{\delta_1}{x_1} + \cdots + \frac{\delta_N}{x_N}.$$

It will be obvious that both these formulae yield special cases of the theorem if they be multiplied through by $\bar{x}$. In the second case the constant quantities $\bar{x}/x_1, \cdots, \bar{x}/x_N$ take the places of $c_1, \cdots, c_N$. Hence, we have the following corollary.

COROLLARY. (*a*) *If* $\bar{x}$ *is c times a power of x,*

$$\frac{\bar{\sigma}}{\bar{x}} = |N|\frac{\sigma}{x}.$$

(*b*) *If* $\bar{x}$ *is c times the product of several different x's,*

$$\left(\frac{\bar{\sigma}}{\bar{x}}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \cdots + \left(\frac{\sigma_N}{x_N}\right)^2;$$

and the same is true if $\bar{x}$ *is obtained partly by multiplying the x's together, and partly by dividing by them, thus:*

if $$\bar{x} = \frac{x_1}{x_2}, \quad \left(\frac{\bar{\sigma}}{\bar{x}}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2.$$

We may prove this last equation if we first consider the case where $\bar{x} = 1/x$. This is obtained from (*a*) by setting $c = 1$, $N = -1$, and accordingly

$$\frac{\bar{\sigma}}{\bar{x}} = \frac{\sigma}{x}. \tag{15}$$

Now, going back to the case to be proved, we may write

$$\bar{x} = (x_1)\left(\frac{1}{x_2}\right), \quad \left(\frac{\bar{\sigma}}{\bar{x}}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2,$$

as desired.

Note that (*a*) is not a special case of (*b*). This is because in (*a*) the N x's are not independent, but are all alike instead. The difference is illustrated in Example 15. Note also that by these formulae one can obtain the relative precision of $\bar{x}$ if one knows merely the relative precisions of the individual x's. This is illustrated in Example 17.

Example 15. (*a*) A rectangle is supposed to be square, and its area is found by repeated measurements of one side (x_1), with the following result, $100 \pm .01$ feet. Find the probable error of the area.

By Corollary (*a*):

$$\frac{\bar{s}}{10000} = 2\left(\frac{.01}{100}\right), \quad \bar{s} = 2.$$

(*b*) Find the probable error of the same area if the rectangle is not supposed to be square, and measurements are made on the adjoining side with the same result, $100 \pm .01$ feet.

By Corollary (*b*):

$$\left(\frac{\bar{s}}{10000}\right)^2 = \left(\frac{.01}{100}\right)^2 + \left(\frac{.01}{100}\right)^2, \quad \bar{s} = \sqrt{2} \text{ feet.}$$

The reason why $\bar{s}$ should be smaller in this case than in (*a*) is that here a positive error in measuring x_1 may be compensated for by a negative error in measuring x_2, but in (*a*) a positive error in measuring x_1 is always propagated into the area.

Example 16. Find the precision with which the volume of a circular cone can be obtained from the following measurements of the radius of its base and its altitude:

$$r = 10, \ \sigma_r = .1, \ h = 10, \ \sigma_h = 0.05.$$

Since $V = \frac{\pi}{3} r^2 h$, we may first let x_1 play the rôle of r^2 and x_2 the rôle of h in Corollary (*b*), and we get

$$\left(\frac{\sigma_V}{V}\right)^2 = \left(\frac{\sigma_{r^2}}{r^2}\right)^2 + \left(\frac{\sigma_h}{h}\right)^2.$$

To obtain σ_{r^2}/r^2, we now employ Corollary (a), letting x play the rôle of r, and we have, for $N = 2$,

$$\frac{\sigma_{r^2}}{r^2} = 2\frac{\sigma_r}{r}.$$

Hence,

$$\left(\frac{\sigma_V}{V}\right)^2 = 4\left(\frac{\sigma_r}{r}\right)^2 + \left(\frac{\sigma_h}{h}\right)^2 = 4\left(\frac{.1}{10}\right)^2 + \left(\frac{.05}{10}\right)^2 = 0.000425.$$

Now $V = 1047.2$. So $\sigma_V = 21.59$.

Example 17. With what relative[1] precision (σ_x/x) should the height and diameter of a circular cone be measured in order to insure a relative precision of 1/1000 in the volume? It is to be assumed that the height and diameter are to be measured with the same relative precisions.

$$V = \frac{\pi}{12}(2r)^2 h, \quad \left(\frac{\sigma_V}{V}\right)^2 = 4\left(\frac{\sigma_{2r}}{2r}\right)^2 + \left(\frac{\sigma_h}{h}\right)^2 = 5\left(\frac{\sigma_h}{h}\right)^2,$$

$$\frac{1}{1000} = \sqrt{5}\left(\frac{\sigma_h}{h}\right), \quad \frac{\sigma_h}{h} = \frac{\sqrt{5}}{5000}.$$

EXERCISES § 9

Note: In Exercise 1 assume that p and v are to be measured with the same relative precisions; and make similar assumptions in all similar cases. Interpret the word precision as referring to standard deviation, and the double sign $\pm$ as referring to probable error.

1. The "gas constant" is $R = pv/T$, where T is the absolute temperature. If $T = 323$, how precisely must p and v be measured to make the relative standard deviation of R less than 1%?

Ans., .007.

2. How precisely must p, v, and T be measured to insure the same result?

[1] If instead of σ/x one uses the words relative precision in the sense of maximum possible error in x divided by x, the problem does not involve the theory of probability at all, only the definite errors δ. In this sense this sort of problem is common in current texts in calculus. One would have:

$$\frac{\delta_V}{V} = 2\frac{\delta_{2r}}{2r} + \frac{\delta_h}{h}, \quad \frac{1}{1000} = 3\frac{\delta_h}{h}, \quad \frac{\delta_h}{h} = \frac{1}{3000}.$$

3. The indicated horse power of an engine is given by the expression, $PLAN/33000$. How precisely should these four quantities be measured in order that the result may be reliable to 1%?

Ans., .005.

4. With what relative precision would it be necessary to measure the length and diameter of a right cylindrical column, which is approximately 12 inches long and 6 inches in diameter, in order to determine the volume to .10%? With what standard deviation in cubic inches would this correspond? *Ans.*, .00045, .339.

5. If it is desired to compute the area of a circle, approximately 10 sq. cm. in area, to 5 parts in 1000, what standard deviation is permissible in the measurement of the diameter? *Ans.*, .0089.

6. If the volume of a sphere is computed from a measurement of its diameter, how precisely should the latter be measured in order that the former may be reliable to .1%? *Ans.*, .00033.

7. To secure the same precision (as in Exercise 6) in the volume of a rectangular parallelepiped, how precisely should the three dimensions be measured? If the approximate lengths are 6, 5, and 2, to what standard deviation of the last dimension would this precision correspond? *Ans.*, .00058, .0012.

8. A freely falling body passes through the distance $h = gt^2/2$ feet in t seconds. It is desired to find h, given $g = 32.2 \pm .1$, $t = 10 \pm 0.5$. What is the probable error of h? *Ans.*, 161 feet.

9. In Exercise 8, to determine g from measurements of h and t, how precisely should h and t be measured, if the desired precision in g is .1%? *Ans.*, .00045.

10. In the figure of Exercise 7, suppose that the probable errors to which the dimensions are liable are, in the order indicated, .002, .001, and .0005 inch, respectively. Find the percentage probable error in the volume. *Ans.*, .046.

11. In order to obtain the lateral area of a right circular cone, the diameter of the base is measured and found to be $43.05 \pm .11$ cm., and the slant height measures $25.27 \pm .19$ cm. Find the area and its probable error.

12. The following measurements of the altitude of a hemispherical hill are used to determine its weight: 1121, 1123, 1123,

1122, 1120, 1122, 1123, 1125, 1122, 1123 feet. Supposing the weight is estimated at 100 ± 20 pounds per cubic foot, find the probable error (a) of the computed volume, (b) of the computed weight.

13. It is desired to measure the mass of the moon so that its relative probable error shall be less than .001 by the use of the formula

$$f = K\frac{Mm}{r^2},$$

where f is the acceleration, K is a constant whose probable error is negligible, M is the mass of the earth, m the mass of the moon, and r the distance between the two. What relative probable error is allowable in measuring the three quantities f, M, and r?

Ans., .00041.

14. A student's grade in mathematics depends on three factors: (a) the average of 36 daily assignments marked on a percentage scale by an assistant whose probable error, in marking each paper, is 15%; (b) the average of four formal tests marked by an instructor whose probable error is 10%; and (c) a final examination graded by a group of instructors; the probable error of the grade on examination is 5%. Find the probable error of the student's final grade if the weights given to a, b, and c are as 3, 4, and 5, and the averages are 80%, 70%, and 60%. By how much would this probable error have been reduced had the daily assignments been graded by the instructor? *Ans.*, 2.74, .04.

PROBLEMS CHAPTER IV

1. Repeat the whole of Example 1, page 245, using as the universe

t	0	1	2
$p(t)$	$\frac{7}{17}$	$\frac{1}{17}$	$\frac{9}{17}$

made by combining the first two categories of that universe, and using as samples those given samples, except that in each case the first two categories are to be combined, so that the first sample becomes 5, 11, 11.

2. Returning to Example 1, § 2, what is the probability that, in selecting a sample of 27×50 individuals from such a universe, the mean of the sample would differ from $\bar{l} = 2$ by as much as .00222? This was the observed difference. Is there reason to suppose, therefore, that the given 50 samples were not drawn, as stated, in a random fashion from the given universe?

3. In Example 1, again, what is the probability that the mean of a sample of 270 would differ from 2 by as much as that observed in the total of the first ten given samples?

4. In each of the following cases find the probability that the mean ($\bar{l}$) is in error by as much as 1% of itself. That is, assume an infinite universe having the parameters of the given sample and find the probability that the mean of a sample drawn from such a universe would differ from the mean (l) of the universe by as much as $\delta = \bar{l}/100$. In each case omit the term in α_4.

(*a*) Part I, Chapter IV, Example 1, page 50; (*b*) same, Example 2, page 51;

(*c*) Part I, Chapter II, Problem 4 (*a*), page 34; (*d*) same, Problem 4 (*b*), page 34;

(*e*) Part I, Chapter VIII, Example 2, $f(X)$, page 130; (*f*) same, $f(Y)$.

5. Fit an F curve to the distributions (*a*), (*b*), (*c*), obtainable from Example 1, page 245, as indicated below, and display the relations between them by graphs as follows:

Figure 1: (*a*) the universe; (*b*) the totals of the samples; (*c*) the given approximation to the curve of means.

Figure 2: (*a*) same as (*c*) of Figure (1); (*b*) the true curve of means.

6. For a group of college freshmen (*Gavett*) the mean height was 68.183 inches and $\sigma = 2.51$, $\alpha_4 = 3.16$. What is the probability that, in a sample of 25, the mean height will be between 67.183 and 69.183?

7. The dean of a large college was recently quoted as saying that the sons of graduates would be favored for admission because it had been found that on the average their grades were .2% higher than the general mean. Assume $\tilde{\sigma} = 12\%$, and that the dean had used a sample of "sons" as large as 1000. How significant a difference was the .2%?

8. A student selected what he claimed was a random sample of 100 books from the 900 of Problem 4, Chapter VIII, Part I, with the results shown in the table below. Judging by his average alone, do you believe the sample was truly a random one?

Length	10–	12–	14–	16–	18–	20–	22–	24–	26–	28–	30–	32–
Frequency	0	0	7	7	28	13	28	15	0	0	2	0

9. Compute σ for each of the first 20 samples of Example 1, page 245, thus finding a first approximation to the "curve of σ's." Compute its mean and its standard deviation and compare them with their theoretical values, $\tilde{\sigma}$ and σ_σ.

10. Repeat Example 7, page 260, for each of the following cases (data of Example 1, page 245): (*a*) the first sample; (*b*) the second sample; (*c*) the sum of the first five samples; (*d*) the sum of all the samples.

11. (*a*) For the year 1910, was the mean age of death of male infants in New Jersey significantly different from that of white male infants in the country as a whole? (It is sufficient here to regard the data for the whole country as the universe.) Data from the *United States Life Tables, Department of Commerce.* (*b*) Same for New York State.

Months	Infant Mortality		
	U.S.A.	*N.J.*	*N.Y.*
0–1	14 819	1 618	5 604
1–2	3 945	458	1 535
2–3	3 237	375	1 301
3–4	2 777	367	1 035
4–5	2 460	327	980
5–6	2 160	260	847
6–7	2 054	284	850
7–8	1 791	252	713
8–9	1 642	218	675
9–10	1 505	187	640
10–11	1 217	140	510
11–12	1 291	165	544

12. (*a*) Compare σ of the sample and σ of the universe. On the basis of this comparison is it very unlikely that the sample ($N = 40$) was a truly random one from this universe? (*b*) Make a similar study for α_3 and for α_4; and (*c*) for the sample as a whole (§ 6).

t	0	1	2	3	4
Proportions in Universe..	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$
Proportions in Sample...	.2	.2	.2	.2	.2

13. Use Part I, Chapter VI, § 2. What is the formula for the probable error of the probable error? Find the probable error of the answer to Example 14, § 9 (pp. 82, 274).

14. Refer to Part I, Chapter VIII, Problems 6 and 7, page 146. A truly random sample was supposedly drawn from a universe in which the proportions were as in Problem 6, to obtain Problem 7. Compute δ and $1 - P_\delta$ for the coefficient of correlation. Comment on the value of $1 - P_\delta$.

15. Refer to Part I, Chapter VIII, Problem 4, page 145. (*a*) Require a probability as great as .995 for security and find out how closely we know the value of r_{LB} for the whole library if the data here given constitute a random sample. Suppose in the whole library $r = 0.8$. Would it have been almost impossible to have chosen a random sample of 900 books for which $r \geqq .875$? (*b*) Answer the same question, omitting the word "random."

16. A correlation coefficient is sometimes said to be "significant" if it differs significantly from zero. Let us interpret this as implying a δ so great that $P_\delta \geqq .99$, and so, by the table on page 258, that $\delta \geqq 3.5$. Now how large must r be to be significant

(*a*) if $n = 100$? (*b*) if $n = 1000$? (*c*) if $n = 10{,}000$?

17. A recent book on statistics gave as the correlation between honesty and judgment, as measured by certain methods in 136 children, $r = -.2064$. (*a*) What was the probability that, in sampling from a universe in which the correlation was zero, a correlation as far from zero as $-.2064$ might have been obtained? (*b*) Requiring $P \geqq .999$ for security, and using the table[1] on page 258,

[1] A more satisfactory answer could be obtained by more advanced methods.

are we sure that $-.2064$ could not have been obtained from a universe in which r was $+.2064$?

18. (Illustrating the necessity of having a truly random sample if one is to apply sampling theory.) The student should fill the blanks below. Given a universe like that of Problem 15, one selects 579 books in order to obtain the mean breadth for the universe, but instead of obtaining a truly random sample one actually selects those books of that table for which the lengths were 20 cm. or greater. The mean of these 579 is , instead of the true breadth 14.714 desired. If the true mean breadth were as small as 14.714, the chance of obtaining a random sample in which the mean would be as great as that just observed would be less than . Therefore, if one applied sampling theory to this case, one would be very certain that the true mean breadth was not as small as 14.714, and this conclusion would be false.

19. In each of the following cases find the probability that r is in error by as much as 1% of itself (in the sense of Problem 4): (*a*) Part I, Chapter VIII, Problem 3, page 144; (*b*) same, Problem 4; (*c*) same, Problem 12, page 147.

20. Use the χ-test to discover whether the following samples were likely ones from their respective universes:

(*a*) The first sample of Example 1.

(*b*) The sum of the fifty samples in Example 1.

(*c*) The distribution of lengths in Problem 14.

(*d*) The distribution of lengths in Problem 8.

(*e*) The correlation table of Problem 14.

(*f*) The following 1000 shots actually fired[1] at a target. The universe is a normal ladder of dispersion with 7 rungs, as indicated:

Belt	*f Observed*	*f Expected*	*Belt*	*f Observed*	*f Expected*
A	5	7	D	397	397
B	99	94	E	95	96
C	402	404	F	2	2

21. In Example 1, what was the likelihood of obtaining two samples of 27 each, such that their means would differ by as much as

[1] Merriman.

(*a*) the means of the first two? (*b*) the means of the last two? (*c*) the mean of the first and the mean of the fifth? (*d*) the two means in which the observed difference was greatest?

22. The average grades of two college fraternities numbering 60 and 35 men, respectively, were 79.98 and 78.22. Was the difference significant? Use $\tilde{\sigma} = 10$.

23. Data from the *United States Life Tables*, mortality of native white male children by months. Did the children who died in 1910 die at a mean age significantly different from those who died in 1909? For $\tilde{\sigma}$ use the σ of the data for 1909 and 1910 added together.

Months	*1909*	*1910*	*Months*	*1909*	*1910*
0–1	14 199	14 808	6–7	1780	2034
1–2	3 642	3 937	7–8	1591	1767
2–3	2 880	3 229	8–9	1590	1616
3–4	2 639	2 764	9–10	1425	1487
4–5	2 247	2 445	10–11	1269	1198
5–6	1 985	2 144	11–12	1206	1271

Totals: 1909, 36,453; 1910, 38,700.

24. Is there a significant difference in the mean age during the adolescent period? (Mortality by years of age.)

Years	*1909*	*1910*	*Years*	*1909*	*1910*
5–6	1020	1016	11–12	456	478
6–7	782	860	12–13	424	418
7–8	632	732	13–14	439	476
8–9	553	603	14–15	444	479
9–10	459	534	15–16	476	495
10–11	477	437			

Totals: 1909, 6162; 1910, 6528.

25. Data of Problem 11. Is the mean age of death for infants significantly different in New York and New Jersey?

26. In the following data, *Biometrika*, vol. 2, p. 444, is there evidence that younger brothers are on the average of different stature from elder brothers? For $\tilde{\sigma}$ compare Problem 23.

FREQUENCY TABLE

Stature	*Elder*	*Younger*	*Stature*	*Elder*	*Younger*
60–61	0	.5	70–71	41.5	47.5
61–62	0	1.5	71–72	30	23
62–63	2.5	3.5	72–73	14.5	15
63–64	6	5.5	73–74	13.5	10.5
64–65	13.5	14.5	74–75	5	4.5
65–66	24	12.5	75–76	1	3
66–67	31.5	38.5	76–77	.5	2.5
67–68	50.5	44.5	77–78	.5	.5
68–69	55	62.5	78–79	.5	.5
69–70	38	36.5	79–80	0	1

27. The problem of Example 13 for age 80, given the data:

White Population	d_x	l_{x+1}
Males.....	1932	12 302
Females.....	2221	15 433

28. Is the female death rate for negroes at age 60 significantly different from the female death rate for whites?

Females	d_x	l_{x+1}
Negro.....	101	2 110
White.....	2597	97 959

29. (*Fisher, R. A.*) Comment on the significance of the following results of inoculation for typhoid: Number inoculated, 6815, of whom 56 were attacked; not inoculated, 11,668, of whom 272 were attacked.

30. Prove Corollary 1 of Theorem IX.

31. Prove the Corollary of Theorem X.

CHAPTER V

CORRELATION — FURTHER TOPICS

1. Regression Curve. In Part I, Chapter IX, regression lines were discussed. The regression line y on x was defined as that line which was the best fit to the data in the sense that

$$\frac{1}{N}\Sigma\delta^2 f(x, y) \tag{1}$$

was a minimum, where δ was the distance, measured parallel to the y-axis, from the line to the point (x, y) of the table, $\delta = y - (A + Bx)$. Also we found, equation (11), § 2, that

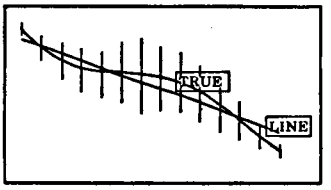
$$\frac{1}{N}\Sigma\delta^2 f(x, y) = \sigma_y^2(1 - r^2), \tag{2}$$

a formula which we shall use again presently. We learned too that we could have arrived at the same result had we chosen, before computing the regression line, to replace all the points that lay in each column by a single point at the mean of that column, but, in so doing, it had to be remembered that this single point was to have a weight equal to the total frequency of the column. The mean of the column at x was written:

$$\bar{y}(x) = \frac{1}{f(x)}\sum_y yf(x, y). \tag{3}$$

Let us now consider the totality of these mean points, $(x, \bar{y}(x))$. They may be connected by a broken line or by a curve. Any such curve, which connects all the mean points, is called a regression curve — more specifically, *the true regression* or *the regression*. Of course, strictly speaking, there are an infinity of such curves, but when we think of such a curve as *the* regression we are thinking, not of the given

correlation table in which the cells have widths h, but of an ideal table in which there are an infinity of cells of width 0. Or we may put it this way: we are thinking, not of the solid histogram bounded by a broken surface, but of some solid which approximates this histogram and is bounded by a smooth surface. In this case there is but one such curve. It is the locus of the mean points $(x, \bar{y}(x))$. What we proved in Chapter IX of Part I shows that the regression *line* is a good linear approximation to this true regression *curve*. Now, instead of approximating the true regression by a straight line, we might approximate it by a parabola, or by an exponential, or by some other type of curve. We should then speak of the parabolic, exponential, or some other kind of regression. This curve which may be chosen to approximate the true regression must not be confounded with the true regression. Both are regression curves but one is an approximation to, and the other is the true regression curve. Occasionally, of course, in ideal cases, the true regression curve is exactly an exponential, or a parabola, or even a straight line. Then we say that *the* regression is exponential, or parabolic, or linear, as the case may be. A similar set of statements may be made for the regression of x on y, so that, in summary, we may say:



The True Regression and the Regression Line

(*a*) *The true regression of y on x is the locus of the mean points of the columns.*

(*b*) *The true regression of x on y is the locus of the mean points of the rows.*

EXERCISES § 1

1. Glance at the data of Problem 12, Chapter VIII, of Part I. Does the true regression of y on x appear to be linear, or

perhaps parabolic, or exponential? What about the regression of x on y (p. 147)?

2. Same for Problem 4 of the same set.

2. Errors of Estimate. The square root of the expression in (1),

$$\sigma_y\sqrt{1 - r^2}, \tag{2a}$$

is called the *standard error of estimate*, and sometimes, analogously, $.6745\ \sigma_y\sqrt{1 - r^2}$ is called the *probable error of estimate*. We learned in Part I that these quantities measure the closeness with which the dots cluster about the regression line. We shall now seek a similar measure with respect to the true regression curve. Again we need to evaluate the expression (1), but now δ shall stand for the distance, measured parallel to the y-axis from the *true* regression curve to the point (x, y),

$$\delta = y - \bar{y}(x).$$

But, with this understanding, (1) is precisely the second moment, about the mean, of the column at x. This we shall designate by

$$\mu_{y^2 \cdot x} = \sigma^2_{y \cdot x}, \tag{4}$$

since it is the second moment in the y direction of the column at x. Similarly, we should write for the second moment of a row at y

$$\mu_{x^2 \cdot y} = \sigma^2_{x \cdot y}. \tag{4a}$$

We have, therefore,

$$\sigma^2_{y \cdot x} = \frac{1}{f(x)} \sum_y f(x, y)[y - \bar{y}(x)]^2. \tag{5}$$

Now, by Part I, Chapter II, § 3, the second moment about the mean point equals the second moment about any given origin minus the square of the distance between the two; so (5) may as well be written thus:

$$\sigma^2_{y \cdot x} = \frac{1}{f(x)} \sum_y y^2 f(x, y) - \bar{y}^2(x). \tag{6}$$

What we need to obtain as a measure of error is the (weighted) average value of $\sigma_{y \cdot x}^2$ for all the x's of the table, and as a unit we should prefer to choose the σ_y of the whole table. This average value is usually designated by $1 - \eta_y^2$ and η is called the *correlation ratio.*

DEFINITIONS. $$1 - \eta_y^2 = \frac{1}{N\sigma_y^2} \sum_x f(x)\sigma_{y \cdot x}^2 \tag{7}$$

Similarly, $$1 - \eta_x^2 = \frac{1}{N\sigma_x^2} \sum_y f(y)\sigma_{x \cdot y}^2 \tag{7a}$$

The square root $\sigma_y\sqrt{1 - \eta_y^2}$ will be a measure of error of estimate corresponding to the standard error of estimate (2a). This may be called the correlation ratio error of estimate, and so, in summary, we may write:

$$\left.\begin{aligned} \textit{standard error} &= \sigma_y\sqrt{1 - r^2}, \\ \textit{correlation ratio error} &= \sigma_y\sqrt{1 - \eta_y^2}, \end{aligned}\right\} \tag{8}$$

and each equals $$\sqrt{\frac{1}{N} \Sigma \delta^2 f(x, y)}, \tag{8a}$$

but in the first case δ is measured from the regression line, and in the second case δ is measured from the true regression curve. This relationship shows why the notation $1 - \eta^2$, instead of η^2, was chosen to denote the average value in (7). It was desired to have a letter η which for this case would correspond to r in the earlier case. If η^2 is nearly 1, η is nearly $\pm$ 1, and the error or dispersion about the true regression is small, just as, when r is nearly $\pm$ 1, the dispersion about the regression line is small. There are, however, two η's, corresponding to the two regression curves. There is in this case no single number available for both curves. The reason for using η^2 instead of η was again because of the expected analogy with r^2, but nevertheless we ought not to have done it unless our definition was to be such that η^2 would be always a positive number. Theorem I shows that this is the case.

Theorem I. $\eta_y^2 = \frac{1}{N\sigma_y^2}\sum_x \bar{y}^2(x)f(x).$

Proof. By (7) and (5)

$$\eta_y^2 = 1 - \frac{1}{N\sigma_y^2}\sum_x\sum_y f(x, y)[y - \bar{y}(x)]^2$$

$$= 1 - \frac{1}{N\sigma_y^2}\sum_x\sum_y f(x, y)[y^2 - 2y\bar{y}(x) + \bar{y}^2(x)]$$

$$= 1 - \frac{1}{\sigma_y^2}\Big[\sigma_y^2 - \frac{2}{N}\sum_x \bar{y}(x)\sum_y yf(x, y) + \frac{1}{N}\sum_x f(x)\bar{y}^2(x)\Big]$$

The second term in the square brackets is minus twice the third term, so that the entire expression simplifies to

$$\eta_y^2 = 1 - 1 + \frac{1}{N\sigma_y^2}\sum_x \bar{y}^2(x)f(x),$$

as desired.

COROLLARY. (*a*) $\eta_x^2 = \frac{1}{N\sigma_x^2}\sum_y \bar{x}^2(y)f(y),$

(*b*) $0 \leqq \eta_y^2 \leqq 1,$

(*c*) $0 \leqq \eta_x^2 \leqq 1.$

To prove (*b*), we note first that by the theorem $\eta^2 \geqq 0$, being equal to an essentially positive (or zero) expression. But, similarly, by (7), $1 - \eta^2 \geqq 0$, *i.e.*, $\eta^2 \leqq 1$.

It would appear from (8*a*) that, if the regression line and the true regression curve are identical, that is, if the true regression is linear, then η^2 equals r^2, because then the δ's are the same in the two cases. Hence the expression,

$$|\eta^2 - r^2| \tag{9}$$

is a measure of linearity of regression.[1] Since there are two η's there are two such measures. It may happen that one regression is linear and the other is not.

[1] Note the criticism in Problem 3.

3. Computation of η. Only the case where N is large will be given special attention, for it is only when N is large that it is important to know the value of η. The computation can best be performed by the use of formulae similar to those used in Part I for computing r. As there, let

$$U = \sum_u uf(u, v), \quad V = \sum_v vf(u, v), \tag{10}$$

where, as before, U and V are in class interval units h and k, and are referred to arbitrary origins, and, as in Part I,

$$\sigma_x = h\sigma_u, \quad \sigma_y = k\sigma_v, \tag{11}$$

and

$$x = h(u - \bar{u}), \quad y = k(v - \bar{v}). \tag{12}$$

Substituting in (3),

$$\bar{y}(x) = \frac{1}{f(u)}\sum_v k(v - \bar{v})f(u, v) = \frac{kV}{f(u)} - k\bar{v}.$$

Therefore, by Theorem I,

$$\eta_y^2 = \frac{k^2}{Nk^2\sigma_v^2}\sum_u\left[\frac{V}{f(u)} - \bar{v}\right]^2 f(u) = \frac{1}{N\sigma_v^2}\left[\sum_u \frac{V^2}{f(u)} - 2\bar{v}\sum_u V + N\bar{v}^2\right]. \tag{13}$$

Now, since

$$\bar{v} = \frac{1}{N}\sum_u\sum_v vf(u, v), \quad \sum_u V = \sum_u\sum_v vf(u, v) = \bar{v}N,$$

and so, inserting this in the second term in the last square brackets of (13), we have

$$\left.\begin{aligned} \eta_y^2 &= \frac{1}{\sigma_y^2}\left[\frac{1}{N}\sum_u \frac{V^2}{f(u)} - \bar{v}^2\right], \\ \text{and, similarly,} \\ \eta_x^2 &= \frac{1}{\sigma_x^2}\left[\frac{1}{N}\sum_v \frac{U^2}{f(v)} - \bar{u}^2\right]. \end{aligned}\right\} \tag{14}$$

Example 1. Find η_y for the data of Example 7, Part I, Chapter VIII, page 140.

u	-2	-1	0	1	2	3	*Sums*
$f(u)$	5	82	400	367	37	9	900
V	-10	-143	-416	-201	23	11	
V^2	100	20449	173056	40401	529	121	
$V^2/f(u)$	20	249.38	432.64	110.08	14.30	13.44	839.84

Down to the double lines the material is copied from Example 7. From that example we know also that $\bar{v} = -.8178$, $\sigma_v^2 = 0.4335$. Hence,

$$\eta_y^2 = \frac{1}{.4335}\left[\frac{1}{900}(839.84) - .6688\right] = .6145; \quad \eta_y = 0.784.$$

EXERCISES § 3

1. For the data of the preceding example find η_x. Note the value of $|\eta^2 - r^2|$ in each case.

2. Find η_y and η_x for the data of Problem 3, Part I, Chapter VIII (p. 144).

4. Common Elements. We shall [1] now find the values of the correlation coefficient in certain chance distributions. The first problem of this nature leads one to an interpretation of correlation which may be expressed, at first not very precisely, as follows: The coefficient of correlation between two characters is equal to the relative number of elements they have in common. To illustrate, suppose we imagine that two characters, say general intelligence and ability in mathematics, were exclusively inherited traits. Suppose further that the process of inheriting were as if the child were given a certain number, say 50, of character elements chosen at random from a large assortment in which the proportion of elements producing intelligence to those not producing intelligence was some fixed number, and suppose the amount of intelligence to appear in the child depended on how large a proportion of

[1] Sections 4 and 5 have no bearing on what follows and may be omitted if desired.

the favorable elements he got. Suppose a similar thing with regard to ability in mathematics, using the same fixed ratio as before, but suppose also that the number of additional elements drawn here were only 30 because it is to be assumed further that 20 of the first type are also of the second type. Thus the child is tq have finally 30 exclusively intelligence determinants, 30 exclusively mathematical determinants, and 20 of the combined type. If all this rather fanciful program were carried out, then it would happen that when children of a fixed parentage were examined the correlation between general intelligence and ability in mathematics would be $20/50: r = 2/5$, the proportion of elements these two characters have in common. Conversely, if the observed correlation is 2/5, then the situation is "as if" some such theory as that suggested were the dominating cause. In our demonstration, we shall use the simpler simile of colored balls drawn at random from an urn.[1] The simplicity of the result is striking and is helpful in aiding one to form a conception of the meaning of correlation, although the student should understand that the illustration is introduced merely for its value as a picture, not as biology.

Theorem II. *Given an infinite set of balls in the ratio, p white to q black ($p + q = 1$); make a pair of drawings from this set as follows: In each drawing there shall be n balls, but before the second drawing is made, select at random m balls from the first group and count them also in the second group, so that for the second group there shall be only $(n - m)$ balls still to be drawn from the universe. Then the correlation between the numbers of white balls in the two groups is m/n.*

COROLLARY.[2] *Suppose n coins are thrown twice, but suppose that before the second throw m coins from the first throw are*

[1] The theorem and various generalizations of it were given by H. L. Rietz, *Annals of Mathematics*, series 2, vol. 21, pp. 306–332. The proof given here differs somewhat from that of Rietz.

[2] This is illustrated very fully by Gavett.

left as they fell, leaving only $(n - m)$ *to be thrown again. The correlation between the numbers of heads on the two counts is* m/n.

The corollary is no different from the theorem, except that in the corollary $p = q = \frac{1}{2}$, but it is perhaps a little easier to hold in mind. We shall therefore give the details of proof in this special case only, leaving the theorem proper to be considered in a problem. It will be found that r is independent of the value of p, and in fact the value of p plays no essential rôle in the proof. Consider first what would happen if none of the m marked coins were counted after the first throw. The correlation table giving the probability of X heads on the first throw and Y on the second would be such that the $f(X)$ marginal totals would be given by the terms of

$$(\tfrac{1}{2} + \tfrac{1}{2})^n,$$

and the $f(Y)$ totals by the terms of

$$(\tfrac{1}{2} + \tfrac{1}{2})^{n-m},$$

and $f(X, Y)$ would equal $f(X) \cdot f(Y)$, and r would be 0. Such a table is given below for $n = 5$, $m = 3$, $n - m = 2$.

Y \ X	0	1	2	3	4	5	$f(Y)$
0	1	5	10	10	5	1	32
1	2	10	20	20	10	2	64
2	1	5	10	10	5	1	32
$f(X)$	4	20	40	40	20	4	128

(The frequencies printed are 128 times the respective probabilities.)

$$(\tfrac{1}{2} + \tfrac{1}{2})^5 = \tfrac{1}{32}(1 + 5 + 10 + 10 + 5 + 1),$$
$$(\tfrac{1}{2} + \tfrac{1}{2})^2 = \tfrac{1}{4}(1 + 2 + 1).$$

Now consider the effect on the means of the successive columns in this table when the m marked coins are counted. On the first column there will be no effect at all, for by the hypothesis that $X = 0$, none of the coins, in particular, therefore, none of the marked coins, were heads; and so on the second count there are now no marked heads to be added in. So the mean of the column at $X = 0$ is as initially

$$\bar{Y}(0) = \frac{n - m}{2}.$$

If $X = 1$, one head was in the first group. There are now two cases:

Case a: When this head was unmarked. Then there is no effect on the count of the second group, for again there are no additional heads to be counted.

Case b: When this head was marked. Then the effect is to add one to each count of the second group. In our example above, the probability of $Y = 0$ at $X = 1$ is no longer 5/128. It is now 0, for we know there is at least 1 head in the second group. But since the probability of no heads was 5/128 before, this is now the probability of 1 head. That is, it is the probability of no additional heads besides the one left down after the first throw. Similarly, 10/128 is the probability, not of 1, but of 2, heads, etc. Let it be clearly understood that the effect of the presence of the marked coin is merely on the number to be counted in the second group, not at all on the number of heads obtained from the two free coins which are thrown the second time. This number is independent of the marked coins, but whenever it is 0, the number of heads to be counted is 1, this 0 plus the head on the marked coin. When it is 1, the number of heads to be counted is 2, etc. The general effect, then, is to move the entire distribution in this column downwards by 1 unit.

The relative frequency with which Case *b* happens is the

relative frequency with which the solitary head will be a marked one. This is

$$\frac{{}_mC_1}{{}_nC_1} = \frac{m}{n},$$

so

$$\bar{Y}(1) = 1 \cdot \frac{m}{n} + \frac{n-m}{2}.$$

Now consider the column at any value of X. The mean will be increased by 1 when exactly 1 of the X heads is marked; by 2 when exactly 2 of these heads are marked; etc. These relative frequencies are, respectively,

$$\frac{{}_mC_1\ {}_{n-m}C_{X-1}}{{}_nC_X}, \quad \frac{{}_mC_2\ {}_{n-m}C_{X-2}}{{}_nC_X}, \text{ etc.},$$

and so

$$\bar{Y}(X) = \sum_{i=1}^{X} \frac{i\ {}_mC_i\ {}_{n-m}C_{X-i}}{{}_nC_X} + \frac{n-m}{2}.$$

The value of Σ in this expression is given by Theorem IV of Chapter II, and is mX/n, for it is the mean of the hypergeometric obtained by drawing a set of m balls from a set of X in which the proportion of marked balls is $q = m/n$. (In that theorem set $m = n$, $n = X$, $t = i$, $p = 1 - q$, $q = m/n$.) Hence,

$$\bar{Y}(X) = \frac{mX}{n} + \frac{n-m}{2}.$$

This, therefore, is the equation of the regression line Y on X in our resultant table. The slope of the regression line is m/n, and we know from the standard equation of a regression line that the slope is $r\sigma_y/\sigma_x$, so that we have finally

$$r = \frac{m\sigma_x}{n\sigma_y}. \tag{15}$$

We now wish to find σ_x and σ_y, and for this purpose let us examine $f(X)$ and $f(Y)$ of the resultant table. Evidently $f(X)$ of the resultant table is the same as $f(X)$ of the initial table, and is given by the terms of $(\frac{1}{2} + \frac{1}{2})^n$. The $f(Y)$ of

the resultant table will not be the same as the initial $f(Y)$ but, like $f(X)$, it will also be given by the successive terms of $(\frac{1}{2} + \frac{1}{2})^n$. This is because our resultant correlation table would not have been different had we asserted that first of all the m marked coins were to be separated out and that then the two throws were to be made in the reverse order. The rôles of the two throws could have been interchanged, and if we had interchanged them, we would have begun with $f(Y)$ as we did begin before with $f(X)$. In each case, therefore, $\sigma = \sqrt{n}/2$, and so $\sigma_x = \sigma_y$, and then by (15), $r = m/n$.

The resultant correlation table in the numerical case outlined on page 294 is as follows:

Y \ X	0	1	2	3	4	5	$f(Y)$
0	1	2	1				4
1	2	7	8	3			20
2	1	8	16	12	3		40
3		3	12	16	8	1	40
4			3	8	7	2	20
5				1	2	1	4
$f(x)$	4	20	40	40	20	4	128

EXERCISES § 4

1. Prove by numerical computation from the correlation table above that: (*a*) the regression Y on X is as stated, (*b*) the correlation $r = m/n$.

2. Throw 5 coins twice as indicated in the corollary just proved, leaving 3 marked coins from the first throw. Repeat this until you have a table in which $N = 128$. Compare with the theoretical table.

5. **Other Probability Distributions.** There are many other interesting types of probability distributions for which one can compute

the correlations. The only other one we shall consider is also well described by a pseudo-biological illustration. First, for simplicity, suppose one has to do with non-sexual reproduction. There is a single grandparent, who produces a number of parents, each of whom produces, independently, a number of children. We want to find the theoretical correlation between the statures of parent and child on the following hypothesis. The initial generating cell from which the adult organism is produced contains n determinants. Some may be called "growth" and others "neutral" determinants. The stature of the adult is determined by the proportion of growth determinants in the cell. Moreover, the adult organism will produce determinants in the same ratio as they were inherited by it. Let the proportion of growth determinants produced by the grandparent be p. The parent selects from this supposedly very large universe of determinants a single cell containing n of them chosen at random. Let the number of growth determinants in that cell be $p'n$. Accordingly, the stature, X, of the parent is proportional to $p'n$ and may as well be taken equal to $p'n$. $X = p'n$. The child now in turn selects a cell containing n determinants chosen at random from a universe in which the proportion of growth determinants is p'. Let the number of its growth determinants be $p''n$. The stature, Y, of the child is then $Y = p''n$.

To find r_{XY}, we have the usual table:

Y \ X		X		
Y		$f(X, Y)$		$f(Y)$
		$f(X)$		$N = 1$

The marginal totals $f(X)$ are given by the point binomial $(p + q)$

$$f(X) = {}_nC_X p^X q^{n-X}, \tag{16}$$

because this is the probability that the parent will obtain exactly X growth determinants. Likewise, confining the attention to a single

column at X (fixed), the numbers $f(X, Y)$ are proportional to the terms of the point binomial $(p' + q')^n$:

$$f(X, Y) = f(X) \cdot {}_nC_Y p'^Y q'^{n-Y}, \tag{17}$$

where $$p' = X/n, \; q' = 1 - p',$$

for the probability $f(X, Y)$ of the child obtaining Y growth determinants from a parent of stature X is the product of the probabilities, first of obtaining such a parent, and secondly of then obtaining Y growth determinants. So

$$f(Y) = {}_nC_Y \sum_X f(X) p'^Y q'^{n-Y}. \tag{18}$$

By (16)

$$\bar{X} = pn, \; \sigma_X^2 = pqn,$$

by Chapter II, Theorem I (b). By (18)

$$\bar{Y} = \sum_Y Y f(Y) = \sum_X f(X) \sum_Y Y \, {}_nC_Y \, p'^Y q'^{n-Y}$$

$$= \sum_X f(X) p'n = \sum_X X f(X) = \bar{X} = pn.$$

Also, using again Chapter II, Theorem I (b), in the places indicated:

$$\sigma_Y^2 = \sum_Y Y^2 f(Y) - \bar{Y}^2 = \sum_X f(X) \sum_Y Y^2 \, {}_nC_Y \, p'^Y q'^{n-Y} - \bar{Y}^2$$

$$= \sum_X f(X)(p'q'n + p'^2n^2) - p^2n^2 \qquad \text{by } (b)$$

$$= \Sigma f(X)(p'n)(q' + p'n) - p^2n^2$$

$$= \Sigma f(X)(X)\left(1 - \frac{X}{n} + X\right) - p^2n^2$$

$$= \Sigma X f(X) + \Sigma X^2 f(X)\left(1 - \frac{1}{n}\right) - p^2n \qquad \text{by } (b)$$

$$= pn + (pqn + p^2n^2)\left(1 - \frac{1}{n}\right) - p^2n^2$$

$$= pn + pqn - pq - p^2n = pn(1 - p) + pq(n - 1)$$

$$= pq(2n - 1).$$

Thus far we have obtained $\bar{X}$, $\bar{Y}$, σ_X, σ_Y. We shall now show that the regression is linear, and then, as in the proof of the preceding

theorem, we shall be able to find r from the slope of the regression line. Fix the attention on a particular column. The mean of the column at X is, relative to the origin of Y, by (17),

$$\bar{Y}(X) = p'n = \dot{X}.$$

This is the equation of the true regression, which is therefore a line of slope 1. So

$$1 = r\frac{\sigma_Y}{\sigma_X} = r\sqrt{\frac{pq(2n-1)}{pqn}},$$

and

$$r = \sqrt{\frac{n}{2n-1}}. \tag{19}$$

COROLLARY. (*a*) *The average stature of the children equals the average stature of the parents and this also equals the stature of the grandparent* ($\bar{X} = \bar{Y} = pn$).

(*b*) *The average stature of the children of a single parent equals the stature of this parent* ($\bar{Y}(X) = X$).

(*c*) *If* $n = 1$, $r = 1$; *if* $n = 2$, $r = \sqrt{\frac{1}{2}} = .82$; *if* $n = \infty$, $r = \sqrt{\frac{1}{2}} = .71$; *so in non-sexual reproduction*[1] *we should expect to find a correlation between* 0.7 *and* 0.8. *The correlation is also independent of* p.

6. Grouping Error in Correlation. In Part I, the reader was warned that there was a large grouping error involved when the number of cells of a correlation table was small. In Chapter VIII, Problem 4, the correlation in a 9×11 table was found to be .875; but earlier, in Example 7, the same data were grouped into a 5×6 table, and the observed correlation was then found to be only 0.581. The error is

[1] In reality, r would be somewhat less than this because stature is not wholly dependent on heredity.

The generalization to the case of sexual reproduction is not presented here, but it is of some interest to know that it may be worked out in a closely similar fashion, and that on the basis of certain reasonable assumptions the correlation comes out between 0.67 and 0.70. This presupposes a fixed grandparentage. If, as in the observable cases, the grandparentage is a mixed population, the effect would be to reduce the correlation still further.

caused by decreasing the number of cells, and is a grouping error like that discussed in Part I, Chapter IV, for one-way frequency distributions, and it is due to the fact that in our computation we assume tacitly that all the material of a cell is concentrated at its central point. A part of this error can be eliminated if we use Sheppard's corrections for σ_u and σ_v, and it would have been a little better to have done this in our earlier work. Also, there are other corrections [1] which may be made instead of these. We shall not consider them here, however, except to remark that they show, first, that the grouping error introduced in reducing a table from one having an infinity of cells to a 10×10 table is usually not important, amounting to about 4%, and, secondly, that when the number of cells is further reduced the error does become increasingly important and commonly amounts to as much as indicated in our illustration. It follows from these considerations that it is better, when feasible, to use the ungrouped material and the longer method of computing r which was indicated for the case when N was small, unless N is really so large that at least a 10×10 grouping is proper. Also, when a 10×10 grouping is proper it is a little better to use Sheppard's corrections for the standard deviations.

But there are cases where these suggestions cannot be carried out, and they occur frequently when we are dealing with unmeasured material. Our method of computing the correlation for such material involved a normalizing of the marginal totals. This process eliminated the grouping error in the same degree as Sheppard's corrections would eliminate it for measured material, but when the number of cells is quite small that amount of correction is far from sufficient. Moreover, for a 2×2 grouping, our method becomes of no value, since it gives the same value for r as would be found if one chose arbitrary numbers as the mid-points of the class

[1] Pearson, K., *Biometrika*, vol. 9, pp. 116 *et seq.*

intervals.[1] This happens because the expression for r is independent both of the origin and of the sizes of the class intervals (h and k). Fortunately, however, an excellent method of finding r for a 2×2 table has been devised by K. Pearson, and this will be considered shortly. It is so good that it can even be employed to advantage when the number of cells is greater than 4. That is, if one had a 3×3 table, one would usually get a value of r closer to the true value by making a coarser, 2×2, grouping and employing Pearson's method than by using any of the methods that we have so far described.[2] In general, however, when the number of cells is greater than 2×2, it is preferable to use the method to be described in the next section.

7. Polychoric Correlation. The word polychoric, as applied to a correlation table, means that the table contains several cells, more than four but not enough to warrant the use of the methods of Part I. We shall consider in this connection — although this is not strictly necessary — only qualitative measures, and so in the following illustration (Table I) the symbols X_1 and Y_1 denote order merely. The problem is to find the true correlation between the characters X and Y. The theory underlying its solution supposes that we are able to construct a normal surface which when cut up into this number of cells (5×6 in our illustration) will contain in the several cells precisely the relative frequencies

[1] The value of r found in this manner can always be computed from the simple formula,

$$r = \frac{a_1 b_2 - a_2 b_1}{\sqrt{(a_1 + b_1)(a_2 + b_2)\,(a_1 + a_2)(b_1 + b_2)}},$$

in the notation of § 8.

[2] Certain more effective methods of handling problems like this are available, but they are complicated. Some authors have proposed defining other coefficients of interrelationship than the coefficient of correlation in cases like this, in order to avoid these difficult computations, but their proposals introduce another difficulty, namely, that of interpreting these other coefficients.

given: $\frac{f_{11}}{N'} \frac{f_{21}}{N'}$, etc. The r that pertains to such a surface is to be the coefficient of correlation sought. In the ideal case there is exactly one such surface. In the practical case there

TABLE I

Y \ X	X_1	X_2	X_3	X_4	X_5	X_6
Y_1	f_{11}	f_{21}	etc.			
A— Y_2						—B
Y_3						
Y_4						
Y_5						
TOTALS	f_1	f_2	f_3	f_4	f_5	f_6

is usually no such surface, and then we are required to find the r of the surface which satisfies these conditions as nearly as possible.

Method.[1] Let AB be any one of the horizontal divisions of the table which cuts [2] all the frequency columns. In the first column let a_1 denote the total frequency above AB, b_1 the total below; in the second column use similarly a_2 and b_2, etc., as indicated in Table II. Now normalize the totals f_1, f_2, etc., so as to obtain the mean points $\bar{x}_1, \bar{x}_2$, etc., by the

[1] The problem is a difficult one and its proper solution belongs to a more advanced text. (Cf. especially K. and E. S. Pearson, *Biometrika*, vol. 14, p. 127.) The simple method presented here is designed to give the student some insight into the nature of the problem and to place in his hands a tool which will yield fairly good results in the less extreme cases.

[2] That is, in no column of the table may the total frequency above or below this line be zero. If this condition cannot be satisfied, then the line is to be taken so that it will be satisfied as nearly as possible, and those columns for which it is not true are to be left out of the reckoning; but it would be possible to include them by modifying the method suitably.

method used in Part I. These means will, of course, be referred to the mean abscissa of the whole table as origin, and the unit of measurement will be σ_x. Now let us consider

TABLE II

Y \ X	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$
FREQUENCIES ABOVE AB	a_1	a_2	a_3	a_4	a_5
FREQUENCIES BELOW AB	b_1	b_2	b_3	b_4	b_5
TOTALS	f_1	f_2	f_3	f_4	f_5

A B

each of these columns individually. Let σ_1, σ_2, etc., be the standard deviations of the columns. By Problem 7, Chapter X, Part I, these standard deviations are all equal [1] to $\sigma_y\sqrt{1-r^2}$, a fortunate circumstance which is essential to the success of this method. Let $\bar{y}_1$, $\bar{y}_2$, etc., be the distances of the means of the several columns below the line AB in terms of σ_1, σ_2, etc., as units. If the columns are normal distributions, these distances can be found from the equations:

$$\int_{-\infty}^{\bar{y}_1}\phi(x)dx = \frac{b_1}{f_1}, \quad \int_{-\infty}^{\bar{y}_2}\phi(x)dx = \frac{b_2}{f_2}, \text{ etc.} \qquad (20)$$

One may thus obtain, relative to a horizontal axis AB and a vertical axis through the general mean point of the whole table, the following coördinates of the mean points of the several columns in σ_x and σ_y units:

$$(\bar{x}_1, \bar{y}_1\sqrt{1-r^2}), (\bar{x}_2, \bar{y}_2\sqrt{1-r^2}), \text{ etc.}$$

Now if there is a normal surface satisfying the conditions laid down at the outset, its regression line Y on X passes through these mean points; its slope is r in the σ_x and σ_y units and $\dfrac{r}{\sqrt{1-r^2}}$ in the σ_x and σ_1 units. If these mean

[1] Exactly equal if the columns are of zero width, otherwise only approximately equal. For very broad columns this approximation is poor and the results are correspondingly affected (cf. § 8).

points do not lie exactly on a line, there is no such surface, and we have to resort to an approximation.[1] Least squares might be used, but, with due allowance for inaccuracies introduced by our other approximations, it is sufficient to use the graphical method, although in drawing our graph it is desirable to accord greater weights to points representative of greater column frequencies. Also, columns in which either a or b is very small should be given very little weight. We may therefore proceed thus: Plot the points $(\bar{x}_1, \bar{y}_1)$, $(\bar{x}_2, \bar{y}_2)$, etc., found from (20), and draw by eye an approximate trend line. Let (x', y'), (x'', y'') be any two conveniently chosen points (preferably far apart) on this line. Then

$$m = \frac{y'' - y'}{x'' - x'}, \quad r = m\sqrt{1 - r^2}, \tag{21}$$

and hence [2]

$$r = \frac{m}{\sqrt{1 + m^2}} = \sin \tan^{-1} m. \tag{22}$$

Example 2. Find polychoric r for the following data. The table pertains to a distribution in which the true correlation is 0.518. It was divided in this peculiar manner by Pearson to illustrate certain difficulties introduced by the use of other measures of interrelation than r. (Statures of father and son with "eye color groupings." *Biometrika*, vol. 9, p. 220.)

	203	117	9	6	
A	125	214	36	46	*B*
	18	53	11	23	
	12	60	13	54	
	358	444	69	129	

[1] If the points do not lie *approximately* on a line, there is no normal surface which approximately fits the data and so the method cannot be used.

[2] The relation $m/\sqrt{1 + m^2} = \sin \tan^{-1} m$ is not needed but it is convenient. This part of the computation can be performed with a slide rule in a single position.

This table corresponds to Table I of the text: $f_1 = 358$, $f_2 = 444$, $N = 1000$, etc. Taking AB as indicated, we next compute Table II:

A	328	331	45	52	B
	30	113	24	77	
	358	444	69	129	

$a_1 = 328$; $b_1 = 30$; etc.

COMPUTATION OF $\bar{x}_1$, $\bar{x}_2$, *etc.*

f/N	.358	.444	.069	.129
Cum f/N *to End Points*	.358	.802	.871	
x, *End Points*	−.364	.849	1.131	
$\phi(x)$	.3734	.2782	.2105	
Differences	−.3734	.0952	.0677	.2105
$\bar{x}$	−1.043	.2144	.9812	1.6318

$$\int_{-\infty}^{x_1} \phi(x)dx = .358; \text{ so } x_1 = -.364.$$

$$\phi(-.364) = .3734.$$

$$\bar{x}_1 = \frac{-.3734}{.358} = -1.043.$$

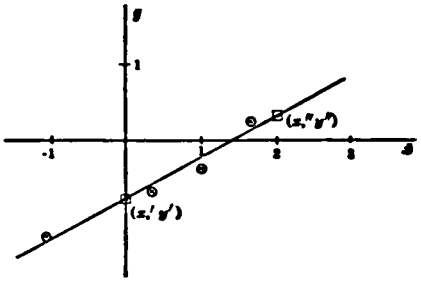
COMPUTATION OF $\bar{y}_1$, $\bar{y}_2$, *etc.*

b/f	.0838	.2545	.3478	.5969
$\bar{y}$	−1.380	−.660	−.391	.245

$$\int_{-\infty}^{\bar{y}_1} = .0838; \text{ so } \bar{y}_1 = -1.380.$$

We now plot the points (− 1.043, − 1.380), (.2144, − .660), etc., and then draw the best-fitting line by eye. On this line we choose

two points, and estimate the values of their coördinates (x', y') and (x'', y'').



By (21) and (22),

$$m = \frac{.35 + .80}{2.0 - 0} = .575; \quad r = \sin \tan^{-1} (.575) = \sin 29.9^\circ = 0.50.$$

EXERCISES § 7

1. Find polychoric r for the table in Problem 4, Chapter VIII, Part I, page 145, taking AB between $B = 12 -$ and $B = 14 -$.

2. Find polychoric r for the following tables taken from the problems of Chapter X, Part I: 2*d*; 2*e* (*Pearson's* answer is .52); 2*f* (*Pearson's* answer is .52); 2*g*.

8. Tetrachoric Correlation ($|r| < 0.8$). The term tetrachoric refers to a 2×2 fold classification. Let the absolute frequencies be denoted as in Table III and the ratios indicated below as in Table IV.

TABLE III

a_1	a_2	
b_1	b_2	
f_1	f_2	N

TABLE IV

A_1	A_2	
B_1	B_2	
F_1	F_2	1

$$f_1 = a_1 + b_1; \quad A_1 = \frac{a_1}{f_1}; \quad B_1 = \frac{b_1}{f_1}; \quad F_1 = \frac{f_1}{N}; \text{ etc.}$$

The problem here is a special case of that of the preceding section: to find a normal surface which when divided into

four cells will present in these cells exactly the relative frequencies observed. This case usually admits of an exact solution. The r that pertains to this surface is to be called tetrachoric r.

K. Pearson has given a method of evaluating tetrachoric r which is perfect, except that often the burden of computation involved is so great that practical investigators are loath to use it. He and Alice Lee have published some tables which lift this burden when $|r| \geqq 0.8$. These will be explained in the next section; in such cases his method should always be used in preference to the one about to be presented. Likewise, when $|r| \leqq .2$,[1] or when a high degree of accuracy is required, his method is better; but otherwise, and certainly in preliminary studies, the rapid, approximate method about to be described will be preferred. It cannot be guaranteed to give r accurately to more than one or two places, but it is better to have a very short method which will do this than rely wholly on a very accurate[2] method which often becomes so tedious that most investigators will resort to some other coefficient rather than use it. For in theory tetrachoric r is simple and has a clean-cut meaning: it is the coefficient of correlation of that normal surface which exactly fits the data.

Applying now the method of § 7 to Table IV we may develop (cf. Problem 6) the following rules of procedure. Find x, y_1, y_2 from the relations:

$$\int_{-\infty}^{x} \phi(x)dx = F_1, \int_{-\infty}^{y_1} \phi(x)dx = A_1, \int_{-\infty}^{y_2} \phi(x)dx = B_2. \quad (23)$$

[1] When $|r|$ is small, the practical problem is not, usually, to find the value of r, but to find out whether or not there exists a significant relation between the two variables. This is better solved by the method of Chapter IV, § 8.

[2] Since the hypothesis, that the data belong to a truly normal distribution, is seldom satisfied, Pearson's high degree of accuracy is usually spurious.

Then

$$m = F_1F_2\frac{y_1 + y_2}{\phi(x)}, r = \frac{m}{\sqrt{1 + m^2}} = \sin \tan^{-1} m, \quad (24)$$

and, as indicated, r has the same sign as m. This formula rests on three assumptions: first, that for such a division of a normal surface the mean of each column would lie on the regression line; second, that its standard deviation would equal $\sigma_y\sqrt{1 - r^2}$; and third, that, considered as a one-way distribution in the y direction, it would be normal. The first of these assumptions is true exactly, the other two approximately only. One could partially compensate for the error due to the second by the use of a more complex formula,

$$r = \frac{F_1F_2}{\phi(x)}\left[y_2\sqrt{1 - r^2\frac{\phi(x)}{F_2}\left(\frac{\phi(x)}{F_2} - x\right)}\right.$$
$$\left. + y_1\sqrt{1 - r^2\frac{\phi(x)}{F_1}\left(\frac{\phi(x)}{F_1} + x\right)}\right], \quad (25)$$

but it is not simply solvable for r. In the special case where $F = .5$, however, it simplifies to

$$r = \frac{m}{\sqrt{1 + \frac{2m^2}{\pi}}}, \text{ and } \frac{2}{\pi} = 0.6366.$$

Investigation of other special cases leads to the somewhat more general approximation:

$$r = \frac{m}{\sqrt{1 + \theta m^2}}, \quad (26)$$

where m is as in (24) and θ is related to F_1 thus:

F_1	.5	.6	.7	.8	.9
θ	.637	.63	.62	.60	.56

This is the formula we shall use, always orienting the table initially so that $F_1 \geqq 0.5$. The conditions under which it

has been derived are such that it should not be employed in certain extreme cases: when one of the given frequencies is less than one per cent of N, when $F_1 > 0.9$, and as stated above when $|r| \geq 0.8$. It should work best when F_1 is about one-half, and y_1 and y_2 are nearly equal.

Example 3. Find tetrachoric r for the material of Example 2 divided as follows:

659	97	
143	101	
802	198	1000

The table of ratios becomes:

.822		
	.510	
.802	.198	1

Thence $x = .849$, $\phi(x) = .2782$, $y_1 = .923$, $y_2 = .025$, $m = .541$, $F_1 = .802$, $\theta = .60$, and $r = .50$, the true value being 0.518.

9. Tetrachoric Tables ($|r| > 0.8$). To handle this case at all satisfactorily the tables [1] of K. Pearson and Lee must be consulted. We proceed merely to show how these tables are to be used. They are reproduced in very condensed form on page 311, but though this is sufficient to permit the student to experiment with the method, it will not permit him to obtain results of value. The method is very simple. From Table III (on page 307) construct Table V by dividing each of the entries in Table III by N: $\alpha_1 = a_1/N$, $\beta_1 = b_1/N$, etc. Also, write

$$G_1 = \frac{a_1 + a_2}{N}, \quad G_2 = \frac{b_1 + b_2}{N}.$$

Further, the initial table should have been so arranged that both F_1 and G_1 would be as large as 0.5.

[1] K. Pearson, *Tables*, XXX. Alice Lee, *Biometrika*, vol. 11, p. 284.

TABLE V

α_1	α_2	G_1
β_1	β_2	G_2
F_1	F_2	1

Let x and x' be determined from the equations,

$$\int_{-\infty}^{x} \phi(x)dx = F_1, \quad \int_{-\infty}^{x'} \phi(x)dx = G_1; \tag{27}$$

x is as before. (In the notation of Pearson and Lee $h = x$, $k = x'$, $d/N = \beta_2$.) Then r can be estimated by interpolation from such tables as the following, in which β_2 is tabulated for various values of x, x', and r.

VALUES OF $\beta_2 = \frac{b_2}{N}$

r	.80						.90						.95					
x' \ x	0	.5	1.0	1.5	2.0	2.5	0	.5	1.0	1.5	2.0	2.5	0	.5	1.0	1.5	2.0	2.5
0	.40	.28	.15	.07	.02	.01	.43	.30	.16	.07	.02	.01	.45	.31	.16	.07	.02	.01
.5	.28	.22	.14	.06	.02	.01	.30	.25	.15	.07	.02	.01	.31	.26	.16	.07	.02	.01
1.0	.15	.14	.10	.05	.02	.01	.16	.15	.12	.06	.02	.01	.16	.16	.13	.07	.02	.01
1.5	.07	.06	.05	.03	.02	.01	.07	.07	.06	.04	.02	.01	.07	.07	.07	.05	.02	.01
2.0	.02	.02	.02	.02	.01	.00	.02	.02	.02	.02	.01	.01	.02	.02	.02	.02	.02	.01
2.5	.01	.01	.01	.01	.00	.00	.01	.01	.01	.01	.01	.00	.01	.01	.01	.01	.01	.00

r	-.80					-.90				-.95			
x' \ x	0	.2	.4	.6	.8	0	.2	.4	.6	0	.2	.4	.6
0	.10	.07	.04	.02	.01	.07	.04	.02	.01	.05	.02	.01	.00
.2	.07	.04	.02	.01	.01	.04	.02	.01	.00	.02	.01	.00	.00
.4	.04	.02	.01	.01	.00	.02	.01	.00	.00	.01	.00	.00	.00
.6	.02	.01	.01	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00
.8	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Example 4. Find the correlation between the ages of husband and wife from Table I, page 312. It is condensed from the material of Example 1, Chapter X, Part I, in which $r = 0.91$.

TABLE I (GIVEN)

451	2429	2880
2324	113	2437
2775	2542	5317

TABLE V (COMPUTED)

		.5417
	.0213	.4583
.5219	.4781	1

Thence $x = .055$, $x' = .105$, $\beta_2 = 0.0213$. The tables indicate that r is between $-.97$ and $-.95$. In order to make both F_1 and G_1 as large as .5, it was necessary to arrange the material initially so that the ages of wives decreased to the right instead of increasing as in the earlier example. So a negative correlation now implies a positive correlation then.

EXERCISES §§ 8–9

Find tetrachoric r in the following cases:

1. The following division of the material of Example 4:

W \ H	60–89	15–59
15–59	222	4631
60–89	406	58

2. Dullness and developmental defects in children (K. Pearson, *Tables*, p. li). *Pearson's ans.*, .652.

	WITHOUT DEFECTS	WITH DEFECTS	TOTALS
NOT DULL	22,793	1,420	24,213
DULL	1,186	888	2,074
TOTALS	23,979	2,308	26,287

3. Mothers' and fathers' habits, Bradford parents (K. Pearson, *Tables*, p. lii). *Pearson's ans.*, $-.081$. (Data on page 313.)

Mothers' Habits

Fathers' Habits	Good	Bad	Totals
Good	994	67	1061
Bad	159	476	635
Totals	1153	543	1696

4. Strength to resist smallpox when incurred and degree of effective vaccination (*W. P. Elderton*, p. 164). *Elderton's ans.*, .7692.

Cicatrix	Recoveries	Deaths	Totals
Present	3951	200	4151
Absent	278	274	552
Totals	4229	474	4703

5. Thyroid anomalies and cancer (Stocks, *Biometrika*, vol. 16, p. 385). *Stocks's ans.*, .4587.

	Thyroid Anomalous	Thyroid Normal	Totals
Cancer Present	93	407	500
Cancer Absent	177	4470	4647
Totals	270	4877	5147

6. Same as Exercise 5 for males only. *Stocks's ans.*, .4905.

	Thyroid +	Thyroid −	Totals
Cancer +	22	143	165
Cancer −	33	1613	1646
Totals	55	1756	1811

PROBLEMS CHAPTER V

1. Compute η_y and η_x and $|\eta^2 - r^2|$ in both cases for the data of Problem 4, Chapter VIII, Part I, page 145.

2. Are the η's, like r, independent of the choice of origin and units? Why?

3. (*a*) Show from equation (8*a*) of § 2 that $\eta = r$ for every 2×2 fold table. (*b*) Prove this also by the use of Theorem I. (*c*) Illustrate by a numerical example of your own choosing. (*d*) Does this indicate that the measure of linearity of regression has the fault that it depends partly on the number of cells used?

4. (*a*) Do the remarks on the cause of grouping errors at the beginning of § 6 apply to η as well as to r? (*b*) Why would it be undesirable to try to obtain a "tetrachoric η" by a method similar to that used for tetrachoric r?

5. Derive both parts of (22) from (21).

6. Derive (23) and (24).

7. Derive the equation following (25).

8. Show that it is always possible to arrange Table V, § 9, so that both F_1 and G_1 are as large as 0.5.

9. Experiment with the data of Example 7, Chapter VIII, Part I, page 140, as follows: (*a*) Find polychoric r. (*b*) Divide into a tetrachoric table in three different ways and find tetrachoric r in each. (*c*) Compare your answers with the corrected value of r obtained in that example (.696).

10. Experiment in a similar manner with the data of Example 1, Chapter X, Part I, page 166.

11. Use (7*a*), (8), and (8*a*) to prove that $\sigma_{y \cdot x} = \sigma_y \sqrt{1 - r^2}$ if $\sigma_{y \cdot x}$ is constant and if the regression y on x is linear. (Cf. also Problem 7, Chapter X, Part I.)

CHAPTER VI

MULTIPLE CORRELATION

1. Notation. We have been studying simple or bivariate correlation. The frequency $f(X, Y)$ has been a function of two variables, X and Y. We now study multiple correlation, in which the frequency, $f(X, Y, Z, \cdots)$, is a function of three or more variables, $X, Y, Z, \cdots$. For the most part, we shall confine ourselves to the three-variable case only. The case of more than three variables is not difficult, once the three-variable case is understood, but the three-variable case permits of a somewhat simpler notation. We formed a physical picture of a bivariate, or, as we shall say, a two-way distribution, by means of a thin sheet of metal of variable density. We marked it off into rectangular cells. Then we had an analogue of the correlation table. The amount of matter in one of these cells corresponded to the frequency in a cell of the table. Instead of a thin sheet we could have thought of a slab of unit thickness if we had preferred to. Now let us think of a solid piece of metal of more than unit thickness and suppose its density to vary from point to point. Mark it off into box cells, small rectangular parallelepipeds. The amount of material in one of these cells corresponds to the frequency in the three-way correlation table we are about to describe. Honey in the honeycomb is a perfect illustration, the number of drops in each cell being the frequency. Another picture is a modern office building. The cells are the offices, situated on various floors, and in various positions on these floors. The number of individuals in a given office corresponds to the frequency in the cell. In this picture, all the offices on any given floor constitute a

two-way table. In the honeycomb, the corresponding two-way table would be indicated by a plane slab of unit thickness cut from the solid. One of our earlier examples of two-way correlation tables was obtained from the measurements of the lengths and breadths of books. If we had taken account of their thickness also, we should have had to use a three-way table. The idea of multiple correlation is very important in practical work because usually the quantities we try to measure are dependent on more than one variable. The stature of a child is dependent not only on the stature of the father but also on the stature of the mother and also to a less degree on the statures of the grandparents.

The framework of our three-way table is a rectangular parallelepiped divided into cells by slicing planes. Our given coördinates are denoted by X, Y, Z; our arbitrary coördinates, of which the given coördinates are a special case, are u, v, w; and x, y, z are the coördinates relative to the general mean of the table. The frequency in the cell at (x, y, z) is $f(x, y, z)$.

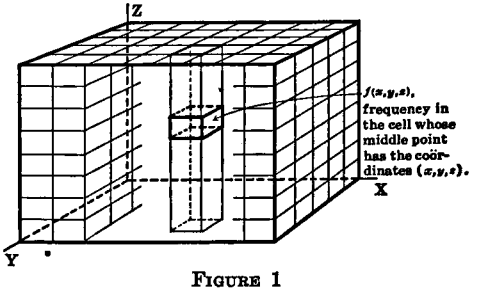


FIGURE 1

Marginal totals are of two kinds, and to avoid ambiguity we shall not use the term at all. If we add all the frequencies in any column parallel to OZ, keeping (x, y) fixed, we get the "column total":

$$f(x, y) = \sum_z f(x, y, z). \qquad (1)$$

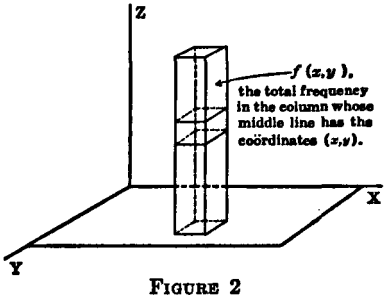


FIGURE 2

All these column totals $f(x, y)$ constitute a two-way set of frequencies whose numerical values may be written if desired on the rectangles of the XY plane.

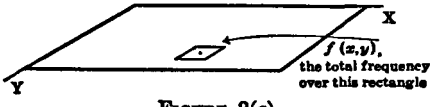


FIGURE 2(a)

They constitute a two-way correlation table. Now if we add all those columns which have the same x, we shall get the total frequency in a slab whose thickness is the same as that of one cell. This slab total is called

$$f(x) = \Sigma f(x, y). \tag{2}$$

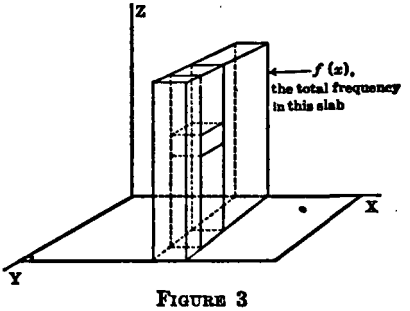


FIGURE 3

Finally, if we add all the slab totals we get the total frequency in the whole table,

$$N = \sum_x f(x). \tag{3}$$

This notation is quite consistent with that used earlier: f signifies "the total frequency at" the place or places indicated by the letter following it. It is always a function of those letters, but not "the f" function in the sense that $f(x)$ would be the same function of x as $f(y)$ is of y. Also, whenever we sum in a given direction, say z, the result is a function which is independent of z and so that letter drops out. Thus, in equation (1), $f(x, y, z)$ is the total frequency at the point, more precisely in the cell at the point, (x, y, z). We sum with respect to z and get $f(x, y)$, the total frequency at the line, more precisely in the column at the line, (x, y). In equation (2) we sum again, now with respect to y, and this letter drops out, leaving $f(x)$, the total slab frequency at x. The notation is summarized in Figure 4, which is a

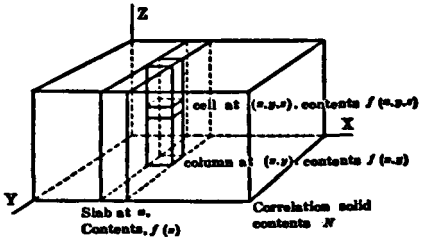


FIGURE 4

composite picture of Figures 1, 2, and 3. Of course we might write (2) as a double sum and (3) as a triple sum:

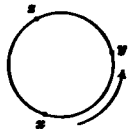
$$f(x) = \sum_{y\,z} f(x, y, z), \quad N = \sum_{x\,y\,z} f(x, y, z).$$

Now, further, instead of beginning with a column parallel to OZ, we might have begun with one parallel to OX or to OY. Another name for a column, running either vertically or

horizontally, is *array*. We may think of these other possibilities geometrically, as pictured for the first case, or, more easily if the notation has been understood, we may write down the equations without thinking of the geometry. Repeating (1), (2), (3), and writing down two more analogous sets, we have:

$$\left.\begin{aligned} f(x,y) &= \sum_z f(x,y,z), \quad f(x) = \sum_y f(x,y), \quad N = \sum_z f(x), \\ f(y,z) &= \sum_x f(y,z,x), \quad f(y) = \sum_z f(y,z), \quad N = \sum_y f(y), \\ f(z,x) &= \sum_y f(z,x,y), \quad f(z) = \sum_x f(z,x), \quad N = \sum_z f(z). \end{aligned}\right\} \tag{4}$$

The relations among these three sets may also be described thus: In going from the first to the second, we make what is called a cyclical permutation of the letters x, y, z; that is, x becomes y, y becomes z, and z becomes x, as if one were going around a circle as in the figure. Do the same in going from the second set to the third, and again the same in going from the third back to the first. This has been done consistently even to the point of writing $f(y, z, x)$ in the second set instead of the usual $f(x, y, z)$. Inasmuch as these two expressions mean the same, it is really immaterial which way they are written. Hereafter, we shall follow the usual practice and write them always in the order (x, y, z). In all the formulae of this section we might have used (X, Y, Z) or (u, v, w) coördinates if desired.



2. **Moments.** In the (u, v, w) system we would have the following definitions of the mean u, mean v, and mean w:

MEANS:

$$\left.\begin{aligned} \bar{u} &= \frac{1}{N}\sum_u\sum_v\sum_w uf(u,v,w) = \frac{1}{N}\sum_u\sum_v uf(u,v) = \frac{1}{N}\sum_u uf(u), \\ \bar{v} &= \frac{1}{N}\sum_v vf(v), \quad \bar{w} = \frac{1}{N}\sum_w wf(w). \end{aligned}\right\} \tag{5}$$

That is, analogous to the two-way case, the general mean point (center of gravity) of the correlation solid is given by $(\bar{u}, \bar{v}, \bar{w})$, and each of these letters represents the mean of a one-way frequency distribution, which consists of all the slab totals in a given direction. Taking this point as the origin of the (x, y, z) system, we have the moments about the general mean defined also in a manner analogous to the two-way case:

MOMENTS:

$$\left.\begin{aligned} \mu_{x^r} &= \frac{1}{N}\sum_{x,y,z} x^r f(x, y, z) = \frac{1}{N}\sum_{x} x^r f(x), \\ \mu_{y^r} &= \frac{1}{N}\sum_{y} y^r f(y), \quad \mu_{z^r} = \frac{1}{N}\sum_{z} z^r f(z). \end{aligned}\right\} \tag{6}$$

The (a, b, c) product moment is

$$p_{x^a y^b z^c} = \frac{1}{N}\sum_{x,y,z} x^a y^b z^c f(x, y, z). \tag{7}$$

There are three important special cases of this:

Let $a = 1, b = 1, c = 0$.

$$p_{xy} = \frac{1}{N}\sum_{x,y,z} xyf(x, y, z) = \frac{1}{N}\sum_{x,y} xyf(x, y) = r_{xy}\sigma_x\sigma_y; \tag{7a}$$

where r_{xy} is the correlation between x and y. The last equation follows from the definition of r_{xy} in Part I, Chapter VIII, page 137.

Let $a = 0, b = 1, c = 1$; then

$$p_{yz} = r_{yz}\sigma_y\sigma_z. \tag{7b}$$

Let $a = 1, b = 0, c = 1$; then

$$p_{xz} = r_{xz}\sigma_x\sigma_z. \tag{7c}$$

Note that equations (7b) and (7c) may be obtained from (7a) by cyclically permuting the letters.

Example 1. As a first example we take an artificial one containing only 12 cells and having a total frequency of only 100. These

numbers are much too small to yield practical results of value, but they are preferable to large ones for an illustration of method. In Figure 5 we have an office building of four floors; $Z = 0, 1, 2, 3$. The partitions have been removed, leaving only their traces on the floors. The number of individuals in each office is the number marked on the floor. This is actually the way we have to write out a solid correlation table, except that usually we do not take the trouble to make a solid picture, using instead four plane two-way pictures and indicating the level (Z) of each just as an architect would use four plans. We will now compute some of the totals and moments. Let the X and Y units be equal to the widths of the cells. Beginning with the column where $X = Y = 0$, and adding upwards, we have

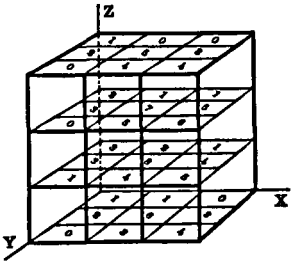


FIGURE 5

$$f(X, Y) = 1 + 3 + 2 + 1 = 7, \text{ at } X = 0, Y = 0;$$
$$f(X, Y) = 1 + 2 + 1 + 0 = 4, \text{ at } X = 1, Y = 0;$$
$$f(X, Y) = 0 + 1 + 1 + 0 = 2, \text{ at } X = 2, Y = 0;$$

etc.; producing finally the set of column totals $f(X, Y)$ in Table 1, and thence by addition the two slab totals $f(X)$, $f(Y)$.

TABLE 1

	Column Totals, $f(X, Y)$			*Slab Totals*
Y \ X	0	1	2	$f(Y)$
0	7	4	2	13
1	10	26	15	51
2	1	16	19	36
Slab Totals $f(X)$	18	46	36	100

In constructing $f(X, Y)$ we added upwards. Let us now begin by adding to the right instead, in arrays parallel to OX:

Coördinates of Column		$f(Y, Z)$
Y	Z	
0	0	1 + 1 + 0 = 2
1	0	2 + 6 + 3 = 11
2	0	0 + 3 + 4 = 7
0	1	3 + 2 + 1 = 6
1	1	3 + 8 + 4 = 15
2	1	1 + 4 + 5 = 10
0	2	2 + 1 + 1 = 4
1	2	3 + 7 + 5 = 15
2	2	0 + 5 + 6 = 11
0	3	1 + 0 + 0 = 1
1	3	2 + 5 + 3 = 10
2	3	0 + 4 + 4 = 8

We thus obtain the array totals $f(Y, Z)$ of Table 2. Then we sum $f(Y, Z)$ first in the Y and then in the Z direction to get $f(Z)$ and $f(Y)$ slab totals:

TABLE 2

	Array Totals, $f(Y, Z)$			*Slab Totals*
Z \ Y	0	1	2	$f(Z)$
0	2	11	7	20
1	6	15	10	31
2	4	15	11	30
3	1	10	8	19
$f(Y)$	13	51	36	100

EXERCISE. Write out the details involved in obtaining the table containing the $f(Z, X)$ totals of Table 3:

TABLE 3

	Array Totals, $f(Z, X)$				*Slab Totals*
X \ Z	0	1	2	3	$f(X)$
0	3	7	5	3	18
1	10	14	13	9	46
2	7	10	12	7	36
$f(Z)$	20	31	30	19	100

In the example just used, we have arrived at three tables of array totals from which, by the methods of Part I, also recapitulated in formulae (5) – (7*c*), page 319, we may obtain the two-way correlations r_{zy}, r_{yx}, r_{zx}. These are called total correlations, to distinguish them from the partial correlations to be described later. Notice that the $f(X, Y)$ table and r_{xy} are exactly what they would be if initially we had paid no attention to what floor the offices were on, but had treated them all the same, irrespective of the floors. Suppose now one had a three-way table showing the relative length, breadth, and thickness of books. The total correlation between length and breadth in such a table would be precisely the correlation we obtained in Part I between length and breadth, no account having been taken there of thickness. Similarly, the total correlation between length and thickness would be what one would obtain if he did not measure or consider the breadth, and the total correlation between thickness and breadth would be obtained by omitting consideration of the length.

The student will have recognized by now that multiple correlation tables of the sort we need to have in order to get

valuable practical results, containing say $10 \times 10 \times 10$ cells, present an enormous amount of detail, which at first sight is confusing. It is the purpose of the mathematical analysis to substitute for this mass of detail a certain few important numbers which will give a simple and effective description of its major characteristics; and we have already seen what most of these numbers are, *viz.*, $\bar{X}$, $\bar{Y}$, $\bar{Z}$; σ_x, σ_y, σ_z; and the total correlations.

Example 2. For Example 1, find the general mean, the σ's, and the total correlations. These are obtained from Tables 1, 2, and 3 after the manner of Part I, and the results are:

$\bar{X} = 1.18$, $\bar{Y} = 1.23$, $\bar{Z} = 1.48$; $\sigma_x = .71$, $\sigma_y = .66$, $\sigma_z = 1.015$; $r_{xy} = .40$, $r_{yz} = .074$, $r_{zx} = 0.033$.

EXERCISES § 2

1. Obtain the results of Example 2.

2. (*a*) For the following multiple correlation table obtain the three tables of array totals and their corresponding slab totals. (*b*) Obtain the general mean, the σ's, and the total correlations.

$(Z = 1)$

Y \ X	4	5	6
5	1	1	
7	1	2	1
9		4	1
11		1	
13			

$(Z = 2)$

Y \ X	4	5	6
5	1		
7		1	2
9	1	3	2
11		2	1
13			1

$(Z = 3)$

Y \ X	4	5	6
5			1
7		2	1
9	2	4	4
11		3	4
13		2	2

3. Regression. The true regression surface, z on xy, is the locus [1] of the mean points of all the columns parallel to

[1] Strictly, this is not a continuous surface, but a set of isolated points through which, of course, many different continuous surfaces might be passed.

OZ. Using a notation analogous to that of Chapter V, § 1, we may write the mean of the column at (x, y), parallel to OZ,

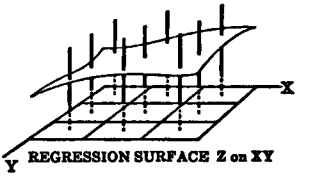
$$\bar{z}(x, y) \equiv \frac{1}{f(x, y)} \sum_z zf(x, y, z); \tag{8}$$

and, similarly, the means of the columns parallel to OX and OY,

$$\bar{x}(y, z) \equiv \frac{1}{f(y, z)} \sum_x xf(x, y, z); \tag{9}$$

$$\bar{y}(z, x) \equiv \frac{1}{f(z, x)} \sum_y yf(x, y, z). \tag{10}$$

So the equations of the three regression surfaces, z on xy, x on yz, and y on zx, are, referred to axes through $(\bar{X}, \bar{Y}, \bar{Z})$:



$$z = \bar{z}(x, y), \quad x = \bar{x}(y, z), \quad y = \bar{y}(z, x) \tag{11}$$

In the figure the continuous vertical lines represent columnar distributions of frequency.

Example 3. Plot the regression z on xy in Example 1. To do this, we must find the mean of each of the columns, *i.e.*, obtain $\bar{z}(x, y)$ for each pair of values of (x, y). The totals of the several columns have, of course, already been computed and are in Table 1, Example 1. Using the larger letters as coördinates instead of the small ones of (8), the mean of the column at $(X = 0, Y = 0)$, whose total is 7, is $\frac{1}{7}(0 \cdot 1 + 1 \cdot 3 + 2 \cdot 2 + 3 \cdot 1) = 1\frac{3}{7}$, and the others are as indicated:

MEANS OF COLUMNS AT (X, Y)

Y \ X	0	1	2
0	$1\frac{3}{7} = 1.43$	$\frac{4}{4} = 1.00$	$\frac{3}{2} = 1.50$
1	1.50	1.42	1.53
2	1.00	1.62	1.53

These numbers are Z coördinates, and the points should be graphed as in the figure.

EXERCISES § 3

1. Obtain results similar to those of Example 3 for: (*a*) the regression x on yz, (*b*) the regression y on zx.

2. Obtain results similar to those of Example 3 and Exercise 1 for the multiple table of Exercise 2 of § 2.

4. **Regression Plane.** The regression plane z on xy is that plane which best fits the true regression surface. Sometimes there is a plane which contains all the mean points of the columns. Then we say the regression is linear, but when this is not the case, and of course it seldom is exactly, we can still find the plane which most nearly contains them all. As in the two-way case, we use the method of moments to determine the equation of this plane.

Let the desired equation be

$$z = A + Bx + Cy. \tag{12}$$

Equate the 0*th* and 1*st* moments of the functions on the two sides of this equation, in the x direction, also in the y direction:

$$\left.\begin{aligned}
&(0th \text{ moment, } x \text{ or } y \text{ direction})\\
&\frac{1}{N}\sum_{x,y,z} zf(x,y,z) = \frac{1}{N}\sum_{x,y,z}(A + Bx + Cy)f(x,y,z),\\
&(1st \text{ moment, } x \text{ direction})\\
&\frac{1}{N}\sum_{x,y,z} xzf(x,y,z) = \frac{1}{N}\sum_{x,y,z} x(A + Bx + Cy)f(x,y,z),\\
&(1st \text{ moment, } y \text{ direction})\\
&\frac{1}{N}\sum_{x,y,z} yzf(x,y,z) = \frac{1}{N}\sum_{x,y,z} y(A + Bx + Cy)f(x,y,z).
\end{aligned}\right\} \tag{13}$$

These equations simplify because of the relations:

$$\frac{1}{N}\Sigma f(x,y,z) = 1, \qquad \frac{1}{N}\Sigma xf(x,y,z) = 0,$$

$$\frac{1}{N}\Sigma yf(x,y,z) = 0, \qquad \frac{1}{N}\Sigma zf(x,y,z) = 0,$$

$$\frac{1}{N}\Sigma x^2 f(x,y,z) = \sigma_x^2, \qquad \frac{1}{N}\Sigma xyf(x,y,z) = p_{xy} \text{ etc.}$$

So we have

$$\left.\begin{aligned} 0 &= A, \\ p_{zz} &= B\sigma_z^2 + Cp_{zy}, \\ p_{yz} &= Bp_{zy} + C\sigma_y^2. \end{aligned}\right\} \tag{14}$$

But $p_{zy} = r_{zy}\sigma_z\sigma_y$, etc., and therefore

$$\begin{cases} r_{zz}\sigma_z = B\sigma_z + Cr_{zy}\sigma_y, \\ r_{yz}\sigma_z = Br_{zy}\sigma_z + C\sigma_y, \end{cases}$$

whence

$$B = \frac{r_{zz} - r_{zy}r_{yz}}{1 - r_{zy}^2} \cdot \frac{\sigma_z}{\sigma_z}, \quad C = \frac{r_{yz} - r_{zy}r_{zz}}{1 - r_{zy}^2} \cdot \frac{\sigma_z}{\sigma_y}, \tag{15}$$

and equation (12) becomes the *Regression Plane*,[1]

$$\frac{z}{\sigma_z}(1 - r_{zy}^2) = \frac{x}{\sigma_z}(r_{zz} - r_{zy}r_{yz}) + \frac{y}{\sigma_y}(r_{yz} - r_{zy}r_{zz}). \tag{16}$$

It is pleasant to notice that this important equation involves only such quantities as can be found from the three two-way tables of column totals indicated in Example 1, and that all the necessary computations were considered in Part I. This result illustrates well the remark made earlier about the objective of the mathematical analysis, for we have here an effective summary of the manner in which, on the average, changes in x and y affect z, without the use of anything more complicated than two-way correlations. There are two analogous equations for the regression of x

[1] Remember this formula thus:

$$\frac{z}{\sigma_z}(1 - a^2) = \frac{x}{\sigma_z}(b - \quad) + \frac{y}{\sigma_y}(c - \quad),$$

where a is the correlation between the two independent variables, not z; b is the correlation between z and the same letter x as appears on the other side of its parenthesis; c is the correlation between z and the same letter y as appears on the other side of its parenthesis. After each minus sign we now put the product of the other two correlations: after $(b -$ $)$ we put ac; after $(c -$ $)$ we put ab. It is useful in numerical applications to remember the equation in this form, not thinking of the precise letters x, y, and z.

on yz and y on zx, which can be obtained easily from (16) by cyclically permuting the letters. In the X, Y, Z letters, (16) would have been:

$$\frac{Z - \bar{Z}}{\sigma_z}(1 - r^2_{xy}) = \frac{X - \bar{X}}{\sigma_x}(r_{xz} - r_{xy}r_{yz}) + \frac{Y - \bar{Y}}{\sigma_y}(r_{yz} - r_{xy}r_{xz}). \quad (16a)$$

EXERCISE § 4

Using the results of the preceding examples and exercises, write the equation of each regression plane for: (*a*) the data of Example 3; (*b*) the data of Exercise 2, § 3.

Ans., (*a*): $Z = .0058X + .11Y + 1.34$,
$X = .43Y + .0025Z + .65$,
$Y = .040Z + .37X + .73$.

5.[1] **Extension to m Dimensions.** Equation (16) can be written in a very simple form by the use of determinants. Let the variables be x_1, x_2, and x_3, and suppose (always) that x_1 is the dependent variable (corresponding to z of § 4). Then the regression plane x_1 on x_2x_3 is given by the equation

$$\frac{x_1}{\sigma_1}R_{11} + \frac{x_2}{\sigma_2}R_{12} + \frac{x_3}{\sigma_3}R_{13} = 0, \quad (16b)$$

where

$$R = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix},$$

and R_{hk} is the cofactor of r_{hk}, *i.e.*, the minor of r_{hk}, including the sign which should be prefixed to it when the determinant is expanded. Thus,

$$R_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & r_{33} \end{vmatrix}, \quad R_{13} = \begin{vmatrix} r_{21} & r_{22} \\ r_{31} & r_{33} \end{vmatrix}.$$

Of course, $r_{11} = r_{22} = r_{33} = 1$, and $r_{12} = r_{21}$, etc., so that for

[1] This section should be omitted by those students who have not studied determinants.

this three-dimensional case the determinant might have been written

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}.$$

Now it happens that the equation for m dimensions is exactly analogous to (16b): The regression "plane" x_1 on $x_2 x_3 \cdots x_m$ is given by

$$\frac{x_1}{\sigma_1} R_{11} + \frac{x_2}{\sigma_2} R_{12} + \cdots + \frac{x_m}{\sigma_m} R_{1m} = 0, \qquad (17)$$

where $R = \begin{vmatrix} r_{11} & \cdot & \cdot & \cdot & r_{1m} \\ \cdot & r_{22} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{m1} & \cdot & \cdot & \cdot & r_{mm} \end{vmatrix}$, and R_{hk} is the cofactor of r_{hk}.

Again it is to be noticed that all the computations may be performed by the use of two-way tables only.

6. Applications. We now illustrate the uses to which the regression plane is commonly put, and, in a later section (§ 8), we shall consider some precautions that should be taken in making these applications. Here let us assume that the regression is nearly linear, so that the true regression surface is not far from the regression plane, and that each column is nearly a symmetrical distribution, so that the most likely value (mode) of a column is close to its mean value. Then, if we know X and Y, the equation of the regression plane (16a) determines, approximately, the most likely value of Z.

Example 4. For a set of 665 books like those considered earlier, the following correlations, means, and σ's have been computed:

$r_{LB} = .86$, $r_{LT} = .35$, $r_{BT} = .41$; $\bar{L} = 20.76$, $\bar{B} = 14.72$, $\bar{T} = 3.42$; $\sigma_L = 3.35$, $\sigma_B = 2.54$, $\sigma_T = 1.50$; in centimeters.

(*a*) Find the equation which gives approximately the most likely values of the length, when the breadth and thickness are known.

(*b*) Similarly, find the breadth, given the length and thickness.

In (*a*) we let L, B, T play the rôles of Z, Y, X of equation (16*a*): So

$$\frac{L - 20.76}{3.35}(1 - r^2_{BT}) = \frac{T - 3.42}{1.50}(r_{LT} - r_{BT}r_{LB}) + \frac{B - 14.72}{2.54}(r_{LB} - r_{BT}r_{LT}),$$

whence

$$L = 4.06 + 1.14\,B - .007\,T, \text{ in centimeters.}$$

In (*b*) we let B play the rôle of Z. So, using (16*a*) again:

$$\frac{B - 14.72}{2.54}(1 - .1225) = \frac{L - 20.76}{3.35}(.86 - .143) + \frac{T - 3.42}{1.50}(.41 - .301),$$

whence

$$B = 1.15 + .619\,L + .210\,T, \text{ in centimeters.}$$

Example 5. My forearm measures 18¼ inches in length; my wife's is 16¾. (*a*) What is the most probable length of our son's forearm? From *Biometrika*, vol. 2, pp. 376 ff., we obtain the following auxiliary data. Let F refer to the forearm of the father, M of the mother, and S of the son. One inch is the unit.

$$r_{FM} = .198, \qquad \bar{F} = 18.3, \qquad \sigma_F = .96,$$
$$r_{FS} = .421, \qquad \bar{M} = 16.5, \qquad \sigma_M = .86,$$
$$r_{MS} = .406, \qquad \bar{S} = 18.5, \qquad \sigma_S = .98.$$

By (16*a*) the regression equation is

$$\frac{S - 18.5}{.98}(1 - .039) = \frac{F - 18.3}{.96}(.421 - .0805) + \frac{M - 16.5}{.86}(.406 - .0835).$$

$$S = 5.56 + .362\,F + .383\,M.$$

Putting $F = 18.25$ and $M = 16.75$, this gives 18.58 for S.

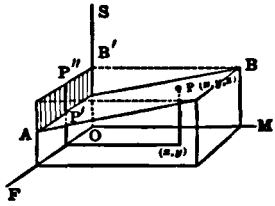
(*b*) What would be the most probable length of my son's forearm if I did not know the length of my wife's? This is an inexact statement of the question at issue, for the most probable length of my

son's forearm is not truly affected by my knowledge of my wife's. The exact question at issue is: What is the mean length of forearms of sons of all those fathers whose own forearms measure 18¼ inches? This is given, approximately, by the regression (in the two-way table of column totals) of S on F:

$$\frac{S - \bar{S}}{\sigma_S} = r_{SF} \frac{18.25 - \bar{F}}{\sigma_F},$$

and so $S = 18.49$.

It may be asked, why have both these answers, (a) and (b), to substantially the same question? In practice we do not use both. If one wished to find the value of S, one would use (a) in preference to (b) if one had the data, because, as we shall prove analytically later, (a) is more likely to be correct. In the column at (F, M) the dispersion about the mean value of S is less than it is in the slab at F. This fact to be established later may be appreciated now by the use of a geometrical diagram. Consider an ideal case, where all the points of the solid, three-way table lie exactly in one regression plane AB. If we know that $F = y$, $M = x$, we can find that $S = s$ exactly, as pictured: we come up from the point (x, y) to a *single* point P in the regression plane. Let AB' be the projection of AB on the plane FS. This is a representation of the two-way correlation table of the totals of the columns which are parallel to OM. If, now, we knew only that $F = y$, we would only know that the value of S is such that the corresponding point in AB' is somewhere in the line segment $P'P''$. So, in this case, by using the solid figure we can find the height of P exactly, but if we only use the plane figure, FS, its height is quite uncertain. The total dispersion is zero in the column at (F, M); it equals $P'P''$ in the slab at F. In less ideal cases the difference is not so marked, but it is real and may be interpreted geometrically in the same manner.



The discussion just given indicates that, did we know also other pertinent conditions, such as the lengths of the forearms of the grandparents, or of the brothers or sisters, and the intercorrelations, our answer could be even more exact. This is true, and this is why

it was desirable to extend, in § 5, the formula for regression in three-way space to m-way space.

EXERCISES § 6

1. (*a*) Using the data of Example 4, find the equation which gives the thickness if the length and breadth are known. (*b*) Which is the more important factor?

2. (*a*) Using the data of Example 5, find the most probable length of my forearm if my son's measures 18.58 inches and my wife's 16.75 inches. (*b*) Explain the meaning of the result. Why may it differ from 18.25 inches? (*c*) Similarly, find the length of my forearm if it be known merely that my son's measures 18.58 inches. Why does the answer differ but slightly from that of (*a*)?

3. A student's grades are: mathematics, 80%; history, 60%. What is his probable grade in chemistry if, in the percentage unit,

$$\bar{m} = 68, \quad \sigma_m = 10, \quad r_{mh} = .3,$$
$$\bar{h} = 72, \quad \sigma_h = 8, \quad r_{mc} = .4,$$
$$\bar{c} = 75, \quad \sigma_c = 7, \quad r_{hc} = .3?$$ *Ans.*, 75.8.

4. What would be the results in Exercise 3 if

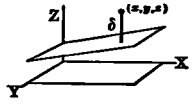
(*a*) his grades in mathematics were unknown? *Ans.*, 71.8.

(*b*) his grades in history were unknown? *Ans.*, 78.4.

7. Multiple Correlation Coefficient. In Chapter IX of Part I, § 2, we learned that

$$\frac{1}{N}\sum_{x,y}\delta^2 f(x, y) = \sigma_y^2(1 - r^2), \tag{18}$$

where δ was the distance, measured parallel to the y-axis, from the point (x, y) to the regression line y on x. This was



obviously a measure of the closeness with which the data clustered about this line, and its square root was for this reason called a standard error of estimate. Also, it was evident from equation (18) that the greater r^2, the smaller was this standard error of estimate. In a similar way, we may now form the analogous expression,

$$\frac{1}{N}\sum_{x,y,z}\delta^2 f(x, y, z), \tag{19}$$

where δ is the distance, measured parallel to the Z-axis, between the regression plane and (x, y, z); and the square root of this expression will be called a standard error of estimate. Furthermore, we shall prove that this expression is such that, if we so define ρ_z that (19) may be set equal to

$$\sigma_z^2(1 - \rho_z^2), \tag{20}$$

then $0 \leqq \rho^2 \leqq 1$. The quantity ρ, thus defined, is called the *multiple coefficient of correlation of z on xy*, and it has other properties analogous to those of r in the two-way case. Unlike r in the two-way case, it must be accompanied by its subscript, for the multiple coefficient of correlation of z on xy is not usually the same as that of x on yz or of y on zx. If, in (19), δ were the distance between (x, y, z) and the true regression surface, the expression could be put equal to

$$\sigma_z^2(1 - \eta_z^2), \tag{21}$$

where η_z is the *multiple correlation ratio of z on xy* and is analogous to η_y in the two-way case. In particular, the smallness of

$$|\eta^2 - \rho^2| \tag{22}$$

is again a measure of linearity of regression.

In order to prove these statements, we proceed in a manner quite analogous to that of Chapter V. First we prove

Theorem I. $\eta_z^2 = \dfrac{1}{N\sigma_z^2} \sum_{x,y} \bar{z}^2(x, y) f(x, y).$

Proof. By (19) and (21),

$$\eta_z^2 = 1 - \frac{1}{N\sigma_z^2} \sum_{x,y,z} f(x, y, z)[z - \bar{z}(x, y)]^2$$

$$= 1 - \frac{1}{N\sigma_z^2} \sum_{x,y,z} f(x, y, z)[z^2 - 2z\bar{z}(x, y) + \bar{z}^2(x, y)]$$

$$= 1 - \frac{1}{\sigma_z^2}\left[\sigma_z^2 - \frac{2}{N} \sum_{x,y} \bar{z}(x, y) \sum_z z f(x, y, z) + \frac{1}{N} \sum_{x,y} \bar{z}^2(x, y) f(x, y)\right].$$

The second term in the square brackets is minus twice the third term, so that the entire expression simplifies to

$$\eta_z^2 = 1 - 1 + \frac{1}{N\sigma_z^2} \sum_{x,y} z^2(x, y)f(x, y),$$

as desired.

COROLLARY. $0 \leq \eta_z^2 \leq 1$. The proof is exactly analogous to that of Corollary (*b*) to Theorem I of Chapter V.

Theorem II. $\rho_z^2 = \frac{1}{N\sigma_z^2} \sum_{x,y} (Bx + Cy)^2 f(x, y)$.

Proof. By (19) and (20),

$$\rho_z^2 = 1 - \frac{1}{N\sigma_z^2} \sum_{x,y,z} f(x, y, z)[z - (Bx + Cy)]^2$$

$$= 1 - \frac{1}{N\sigma_z^2} \sum_{x,y,z} f(x, y, z)(z^2 + B^2x^2 + C^2y^2 - 2Bxz - 2Cyz + 2BCxy)$$

$$= 1 - \frac{1}{\sigma_z^2} (\sigma_z^2 + B^2\sigma_x^2 + C^2\sigma_y^2 - 2Bp_{xz} - 2Cp_{yz} + 2BCp_{xy}). \quad (23)$$

We may simplify the expression in parentheses by the substitution of the values of p_{xz} and p_{yz} given in (14). Then (23) becomes

$$\rho_z^2 = 1 - 1 + \frac{1}{\sigma_z^2} (B^2\sigma_x^2 + C^2\sigma_y^2 + 2BCp_{xy}). \quad (24)$$

But this is exactly what we get if we expand

$$\frac{1}{N\sigma_z^2} \sum_{x,y} (Bx + Cy)^2 f(x, y). \quad (25)$$

Hence (25) also equals ρ_z^2 and the theorem is proved.

COROLLARY 1. $0 \leq \rho_z^2 \leq 1$. By the theorem $\rho_z^2 \geq 0$, being equal to an essentially positive quantity. By definition (20) = (19), and therefore

$$1 - \rho_z^2 \geq 0,$$

since this also equals an essentially positive quantity. So

$$\rho_z^2 \leq 1.$$

COROLLARY 2. If the regression is linear,

$$\rho_z^2 = \eta_z^2.$$

This follows at once from a comparison of Theorems I and II. When the regression is linear,

$$\bar{z}(x, y) \equiv Bx + Cy.$$

All the statements made above about ρ_z and η_z have now been proved. Of course, similar definitions and statements apply to ρ_x, η_x, and to ρ_y, η_y.

Theorem III. $\rho_z^2 = \dfrac{r_{xz}^2 + r_{yz}^2 - 2r_{xy}r_{yz}r_{xz}}{1 - r_{xy}^2}$

Proof. Add equations (23) and (24) and simplify:

$$\sigma_z^2\rho_z^2 = Bp_{xz} + Cp_{yz} = Br_{xz}\sigma_x\sigma_z + Cr_{yz}\sigma_y\sigma_z.$$

Insert the values of B and C given in (15):

$$\rho_z^2 = \frac{(r_{xz} - r_{xy}r_{yz})r_{xz}}{1 - r_{xy}^2} + \frac{(r_{yz} - r_{xy}r_{xz})r_{yz}}{1 - r_{xy}^2}$$

$$= \frac{r_{xz}^2 + r_{yz}^2 - 2r_{xy}r_{yz}r_{xz}}{1 - r_{xy}^2}.$$

COROLLARY. (*a*) If $r_{xy} = 1$, the standard error of estimate for z on xy is the same as it would be if the simple correlation z on x or z on y were used.

(*b*) If $r_{xy} = 0$, the standard error of estimate is found from the relation

$$\rho_z^2 = r_{xz}^2 + r_{yz}^2.$$

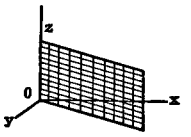
Proofs. Consider first the limiting case, when $r_{xy} = 1$. Apparently, by Theorem III, then ρ_z^2 is infinite because the denominator is zero, but this cannot be the case because by definition $\rho_z^2 \leqq 1$. When $r_{xy} = 1$, it will also happen that $r_{xz} = r_{yz}$, and so the formula of Theorem III really becomes 0/0, a meaningless expression. We can see what is really taking place as r_{xy}^2 approaches 1 by letting $r_{xz} = r_{xy}$ first. Then Theorem III becomes

$$\rho_z^2 = \frac{2r_{xz}^2(1 - r_{xy})}{1 - r_{xy}^2} = \frac{2r_{xz}^2}{1 + r_{xy}},$$

and now if we let r_{xy}^2 approach 1 we see that ρ_z^2 becomes equal to r_{xz}^2. So in this case

$$\rho_z^2 = r_{xz}^2 = r_{yz}^2.$$

Therefore, as stated in (*a*), the standard error of estimate, $\sigma_z\sqrt{1 - \rho_z^2}$, is exactly what it would be if we had used the correlation between x and z or y and z only, without adding the third variable. This would be a case, then, when no advantage would be gained by using multiple correlation instead of simple correlation. Geometrically (see figure), all the data lie in a plane perpendicular to xoy.



Now consider the other limiting case, when $r_{xy} = 0$. The formula of Theorem III becomes

$$\rho_z^2 = r_{xz}^2 + r_{yz}^2.$$

Hence, in this case, we shall get a much closer fit by using multiple regression than if we employ either of the simple regressions, z on x, or z on y; very much closer if also r_{xz} and r_{yz} are nearly equal in value (see also Problem 6).

Example 6 (*Data of Problem* 5). One could determine the eye color of a son from the eye colors of his two parents more accurately than one could determine the eye color of a father from the eye colors of his two sons. For, in the first instance the intercorrelation of the independent variables (father-mother) is nearly zero (.10), while in the second this intercorrelation (brother-brother) is considerable (.517). The actual values of ρ in the two cases turn out to be .667 and .567, respectively. The first case approximates the ideal, which would have given $\rho = \sqrt{(.4947)^2 + (.4947)^2} = 0.700$.

The formula of Corollary (*b*) shows that in the ideal case, when the xy correlation is zero, correlations are combined like forces in mechanics. It may be extended to m-way space:

$$\rho_1^2 = r_{12}^2 + r_{13}^2 + \cdots + r_{1m}^2.$$

In this connection, Pearson makes the interesting remark that this means that grandparental correlation *could* not be more than $\frac{1}{2}$. He assumes that the intercorrelation among one's four grandparents is about zero, and that all four are equally potent in determining the character of the grandson. Then this equation becomes, for $m = 4$,

$$\rho_1^2 = 4r_{12}^2,$$

where r_{12} is the correlation with any chosen grandparent. Since $\rho_1^2 \leq 1$, $r_{12} \leq \sqrt{\frac{1}{4}} = \frac{1}{2}$. Similarly, of course, parental correlation could not be greater than $\sqrt{\frac{1}{2}} = .707$ (cf. Chapter V, § 5, Corollary *c*, and footnote).

EXERCISES § 7

1. Compute each of the three multiple correlation coefficients for the data of Example 2, page 324.

2. Same for the data of Example 4.

3. Same for the data of Example 5. *Ans.*, $\rho_P = .422$, $\rho_M = .407$, $\rho_S = .534$.

4. Using part of the results of Exercise 3, answer the question raised in the discussion of Example 5 (*b*).

5. Compare the accuracy expected in finding (by linear regression methods) the length of a book, knowing both its breadth and thickness, and that obtainable when the thickness only is known.

6. Similarly, compare the accuracy of the prediction of the grade in chemistry from a knowledge of grades in both history and mathematics with that obtainable from a knowledge of one only. (Data of Exercise 3, § 6.)

8. Size of *N*. It is important to know how large a sample we should have in order to use the method of multiple correlation with some confidence. For this purpose one should properly consider the probable errors of the parameters of multiple correlation, but this will not be done here. We shall, however, mention some general considerations which apply to the three-way case. The number of data required

will depend, of course, on just what we want to find out from the data.

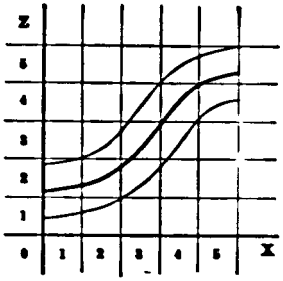
(*a*) *Complete Information.* First suppose we wish complete information about our distribution; that is, as nearly complete information as we sometimes desired with one-way frequency distributions. To solve the one-way case we needed, say, about 10 cells and at least an average of 10 data in each cell. We were hardly satisfied with less than 100 data. To solve with similar completeness the two-way case we should need 10^2 cells and a total of 10^3 data. For the three-way case we should need 10^3 cells and 10^4 data. This is a good many more than we are likely to have, and so we can seldom obtain complete information about the multiple solid.

(*b*) *The True Regressions.* Suppose next we desire to establish the true regressions of z on xy. We need these to get η's. Let us begin with the two-way case, z on x. To establish the form of a regression *curve*, we can hardly get along with less than five well-spaced points of that curve. Each is the mean point of a set of data. To find the mean of a set of data, we are not usually content with less than 10 data. This implies the need of 50 data to establish a regression curve, but they must be properly spaced. It will not do to choose 50 data at random, for they will most probably be bunched in such a way as to allow one to determine very well the points at the center of our curve, but not so well those at the ends. To allow for this tendency in random data, let us double the number and say that at least 100 random data are required to establish, even very approximately, a true regression *curve*. Now, if we are to establish a true regression *surface*, we should have at least five properly spaced curves, that is, apparently 500 data. But again, with random data, even though a certain five curves may thus be well determined, they will not usually be properly spaced unless we allow for bunching toward the

center in the y direction as well as in the x direction. To allow for this, we may perhaps again double the number of data. Finally, then, to establish a true regression surface using random data, and making no a priori assumptions as to the form or even as to the smoothness of that surface, we need something like 1000 data, at least.

If we needed one true regression surface only, say z on xy, and could control the (x, y) spacing of the data, we could manage with a smaller number, perhaps $5 \times 10 \times 5 = 250$, but it should be noticed that if this were done we would not then be in a position to determine also the other regressions, for the special spacing of our data for the z on xy case is a particularly bad spacing for the other cases. This can be seen from a consideration of the two-way cases and the diagram.

The data tend to be grouped within a belt about the true regression curve. This means that in the horizontal rows at $Z = 1$ and $Z = 5$ there are not enough data to determine well the means of these rows. This difficulty is greater if the regression curve is more nearly parallel to the X-axis.



(c) *The Regression Planes.* Suppose now that the regression planes only are required. For this we need the means, standard deviations, and two-way correlations. Fewer data are required for the means than for the standard deviations, and, since also fewer are required for the standard deviations than for the correlations, it is only necessary to consider how many are needed for the correlations. We know approximately the standard deviation of a coefficient of correlation, $\frac{(1 - r^2)}{\sqrt{N}}$, and this will always be less than .05 if $N > 400$. We may, therefore, accept this number as a

proximate figure for this case, but we must be particularly on our guard against using correlations obtained by coarse groupings, for then the grouping error may be so large as to give an entirely erroneous value to r, independently of the fluctuations of random sampling. Finally, if one knows that the regressions are exactly, or very closely, linear, and one requires their equations, he can obtain them with even fewer data. Especially, if one wants only one regression plane and can control the spacings, a very small number of data will suffice, for, to determine this plane, it is only necessary to establish well three widely spaced points which lie on it. But one seldom does know this, and it is a common error to assume that it is true in cases where there are not enough data to prove it. A perfunctory computation of η_s may agree closely enough with the computed ρ_s, but if the number of data is so small that not even the true regression surface (case *b*) can be established, let alone η_s, which measures the average fluctuation about that surface, such an agreement is not a valid argument to establish linearity of regression.

9. Partial Correlation. Let us look back at Figure 3, § 1. Here we have a picture of one slab of our solid, having the thickness of one cell only. It is a two-way correlation table involving the relations between z and y which hold for a particular value of x. The correlation between z and y found from this table is called the *partial correlation* between z and y for the given x and is denoted by

$$r_{zy \cdot x}, \text{ or by } r_{zy}(x).$$

When there are only three variables, as here, we may without ambiguity write it more simply $r(x)$, the correlation at x, analogous to $f(x)$, the frequency at x. Like $f(x)$, $r(x)$ is a function of x but not "the r function" of x. The true regression curves and the regression lines found from this table are called also the *partial true regressions* and the *partial*

regression lines. Since the partial regression curve at x of z on y is the locus of the means of those columns parallel to OZ which lie in the slab at x, it must be a section of the true regression surface z on xy. But in general the partial regression *line* z on y is not a section of the regression *plane* z on xy. Indeed, if for all x's all the partial regression lines z on xy were drawn, these lines would not in general all lie in a plane, but would together constitute the elements of a "ruled" curved surface. Likewise, in general, $r_{zy}(x)$ cannot be found in a simple fashion from the total correlations. In a certain special case about to be considered this is true, but it would be a serious error to assume that it is true in general.

Theorem IV. *Let the true regression, z on xy, and the total regression, z on x, be linear, and let the standard deviations of all the columns parallel to OZ be the same.*[1] *Let also $\sigma_z(x)$ be constant in the two-way table of totals $f(x, z)$. Then (a) the partial regression line, z on y at x, is a section of the regression plane, and (b) the partial correlation is*

$$r_{zy \cdot x} = \frac{r_{yz} - r_{xz}r_{xy}}{\sqrt{(1 - r_{xy}^2)(1 - r_{xz}^2)}}.$$

Proof. Part (*a*) is immediately obvious geometrically, for in this case both the partial regression line and the regression plane contain the mean points of all the z columns in the slab at x. To prove part (*b*), let $\sigma_z(x, y)$ be the constant standard deviation of the column at (x, y). By definition, if δ goes to the regression plane,

$$\sigma_z^2(x, y) = \frac{1}{f(x, y)} \sum_z \delta^2 f(x, y, z), \tag{26}$$

[1] The distribution is then said to be homoscedastic (equally scattered) in the z direction. The conditions of this theorem are satisfied when the distribution is "normal," as defined by *Pearson*.

and so, using also (19) and (20),

$$(1 - \rho_z^2)\sigma_z^2 = \frac{1}{N}\sum_{x,y,z}\delta^2 f(x, y, z) = \frac{1}{N}\sum_{x,y}\sigma_z^2(x, y)f(x, y)$$
$$= \sigma_z^2(x, y), \quad (27)$$

since $\sigma_z^2(x, y)$ is constant by hypothesis. Similarly, in the two-way table at x, we have

$$(1 - r_{zy\cdot x}^2)\sigma_z^2(x) = \frac{1}{f(x)}\sum_y\sum_z\delta^2 f(x, y, z)$$
$$= \frac{1}{f(x)}\sum_y\sigma_z^2(x, y)f(x, y) = \sigma_z^2(x, y). \quad (28)$$

Furthermore, in the two-way table of totals $f(x, z)$ of those columns which are parallel to OY:

$$(1 - r_{zx}^2)\sigma_z^2 = \sigma_z^2(x), \quad (29)$$

by Problem 11 of Chapter V. Now putting together (27), (28), and (29), we have

$$(1 - r_{zy\cdot x}^2)(1 - r_{zx}^2) = 1 - \rho_z^2. \quad (30)$$

Insert in this the value of ρ_z^2 given in Theorem III and we have

$$r_{zy\cdot x}^2 = \frac{r_{yz}^2 - 2r_{xy}r_{yz}r_{xz} + r_{xz}^2 r_{xy}^2}{(1 - r_{xy}^2)(1 - r_{xz}^2)}.$$

Take the square root of both sides, and obtain the result desired.

COROLLARY. *When the conditions of the theorem are satisfied, the partial correlation between y and z for one value of x is the same as for any other value of x.*

This is obvious from the formula, which does not involve the value of x. It means that, not only is the distribution homoscedastic in the z direction, but also the correlations within all slabs perpendicular to the x-axis are equal. One's intuition rebels at accepting this in practical cases. One would be unwilling to grant, for example, that the correlation between mother and son would be truly independent of

the father. So, in practical cases, one must accept the theorem as pertaining to an idealized condition which may approximate but does not exactly represent the true state of affairs.

Example 7. Use the formula of Theorem IV to find the correlation between forearm lengths of mother and son for those cases where the father's forearm measures 18.25 inches.

Let $X = 18.25, x = 18.25 - \bar{F} = -.06, y = M - \bar{M}, z = S - \bar{S}$:

$$r_{SM}(-.06) = \frac{r_{SM} - r_{FS}\, r_{FM}}{\sqrt{(1 - r^2_{FM})(1 - r^2_{FS})}} = 0.363.$$

As stated in the corollary, the result does not make use of the value of x.

10. Application. The expression derived in § 9 for the partial correlation coefficient may be used to eliminate the effect of one of the variables. Suppose, as before, that x, y, and z are the three variables, relative to the general mean as origin, and that we would like to know what the correlation between z and y would be if it were not for the presence of x. It is not always true that this problem has a solution, but when it does have one we might try to find it somewhat as follows: First subtract from the z of each point that part of z which is indicated as due to the influence of x by the equation of the regression line, z on x. Then subtract from the y of each point that part of y which is indicated as due to x by the regression line, y on x. Finally, find the ordinary correlation between what is left of z and what is left of y. The result, we shall prove, is precisely the $r(x)$ of § 9. When interpreted in this manner, $r(x)$ is sometimes called the "net correlation" between z and y.

Proof. Let z' be what is left of z, and y' what is left of y; *i.e.*, let

$$z' = z - r_{zx}\frac{\sigma_z}{\sigma_x}x, \quad y' = y - r_{yx}\frac{\sigma_y}{\sigma_x}x.$$

Then

$$r_{x'y'} = \frac{1}{N\sigma_{x'}\sigma_{y'}} \Sigma z'y'f$$

$$= \frac{1}{N\sigma_{x'}\sigma_{y'}}\left(\Sigma zyf - r_{yz}\frac{\sigma_y}{\sigma_z}\Sigma zxf - r_{xz}\frac{\sigma_x}{\sigma_z}\Sigma yxf + r_{xz}r_{yz}\frac{\sigma_x\sigma_y}{\sigma_z^2}\Sigma x^2 f\right)$$

$$= \frac{1}{\sigma_{x'}\sigma_{y'}}(\sigma_x\sigma_y r_{xy} - \sigma_x\sigma_y r_{yz}r_{xz} - \sigma_x\sigma_y r_{xz}r_{yz} + \sigma_x\sigma_y r_{xz}r_{yz})$$

$$= \frac{\sigma_x\sigma_y}{\sigma_{x'}\sigma_{y'}}(r_{xy} - r_{xz}r_{yz}). \qquad (31)$$

Now, computing

$$\sigma_{x'}^2 = \frac{1}{N}\left(\Sigma z^2 f - 2r_{xz}\frac{\sigma_z}{\sigma_z}\Sigma zxf + r_{xz}^2\frac{\sigma_z^2}{\sigma_z^2}\Sigma x^2 f\right)$$

$$= \sigma_x^2(1 - r_{xz}^2),$$

and

$$\sigma_{y'}^2 = \sigma_y^2(1 - r_{yz}^2),$$

and inserting them in (31), we have

$$r_{x'y'} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}},$$

as desired.

Example 8. In Example 7 we may now say that .363 is the net correlation between the lengths of mother's and son's forearms, after elimination in the aforesaid manner of the effect of the father.

The insertion of the words *in the aforesaid manner* in this example was needed in order to insure that too much would not be read into the result. We know from § 9 that, if the distribution satisfies certain special conditions, then the true partial correlation is constant, and is therefore independent of the father, and in that case the formula in question does truly eliminate the effect of the father. But otherwise it may well happen that the true partial correlation for one father is quite different from what it is for another, and in that case we do not know what the mother-son correlation with the effect of the father eliminated would be. What we have actually found is the net amount of correlation left over

after subtracting the *average* effect [1] of the father on the mother and also his *average* effect on the son. This method of elimination is not ideal, and we must regard our process as affording an approximate rather than a clean-cut and final solution of the problem.

The further study of partial correlation is omitted because it can be treated much more easily by the use of determinants.

PROBLEMS CHAPTER VI

1. The following data, in addition to some previously used, are taken from a paper by Pearson and Lee, in *Biometrika*, vol. 2, pp. 357–462. The measurements are in inches.

	Stature		Span		Correlation Stature-Span
	Mean	σ	*Mean*	σ	
Father = *F*	67.68	2.70	68.67	3.14	.783
Mother = *M*	62.48	2.39	61.80	2.81	.756
Son = *S*	68.65	2.71	69.94	3.11	.802
Daughter = *D*	63.87	2.61	63.40	2.94	.828

Family Correlations

	F–M	*F–S*	*F–D*	*M–S*	*M–D*	*S–D*
Stature	.2804	.514	.510	.494	.507	.553
Span	.1989	.454	.454	.457	.452	.525

Stature-Span Correlations

	F–S	*F–D*	*M–S*	*M–D*	*S–S*	*D–D*	*S–D*
Stature-Span	.418	.423	.424	.431	.444	.471	.478
Span-Stature	.399	.407	.390	.385	.444	.471	.456

In all these cases the correlation solids may be assumed approximately normal. Find the following regression equations (Pearson and Lee), and the corresponding multiple correlation coefficients:

[1] Throughout the section we are using the word "effect" in the sense of relationship as indicated by the data, not in the sense of a causal relationship.

Stature,

$$S = 14.08 + .409\,F + .430\,M,\ D = 10.82 + .386\,F + .431\,M;$$

Span,

$$S = 18.04 + .375\,F + .423\,M,\ D = 14.70 + .355\,F + .395\,M.$$

2. Use the data of Problem 1. (*a*) How accurately can the stature of a man (father) be derived from his span? from his span and the span of his daughter also? (*b*) Which is better to know if one wishes to predict the stature of a daughter, the statures of her brother and father, or the spans of her mother and father? That is, compare the coefficients in the two cases.

3. Use the data of Problem 1. (*a*) Find the correlation in stature between sons and daughters of those mothers who are 5 feet tall; who are $5\frac{1}{2}$ feet tall. Why are the two answers related as they appear to be? (*b*) Find the correlation between stature and span for sons of fathers of given stature. How is this related to the correlation given in the table?

4. Prove the corollary to Theorem I.

5. (Data from *Biometrika*, vol. 2, pp. 221, 390, 482.) Correlations of eye color in man: son-parent = .4947, son-grandparent = .3166, son-great-grandparent = .1879, brother-brother = .517, father-mother = .10 (Galton's family records).

(*a*) Find the multiple correlation, son on parent and grandparent. Assume that the correlation between son and parent is the same as that between father and grandparent.

(*b*) Find the multiple correlation, son on four grandparents, supposing the intercorrelations among the grandparents to be zero.

(*c*) Find the multiple correlation, grandparent on parent and son.

(*d*) Supposing the intercorrelations within one generation to be zero, how well could one estimate one's own eye color from a knowledge of the eye colors of eight great-grandparents? of four grandparents? of two parents? of one parent?

(*e*) Draw a curve through the points indicated by the correlations given at the outset, $r_{S \cdot P}$, $r_{S \cdot G}$, $r_{S \cdot G-G}$, and estimate $r_{S \cdot G-G-G}$. Use ratio paper because Galton and Pearson believed the law to be exponential.

(*f*) From the result of (*e*) compute the accuracy with which one could derive one's eye color by the use of sixteen great-great-grandparents. Does this also indicate the accuracy with which one might estimate the eye color of one's great-great-grandparent by the use of sixteen of his descendants in his own generation?

6. Prove that $|\rho_s| \geq |r_{ss}|$.

CHAPTER VII

FINITE DIFFERENCES

1. Notation. In this chapter we shall consider certain elementary parts of the calculus of finite differences. This subject is closely allied to the calculus of infinitesimal differences, the ordinary "calculus" that is more commonly studied; and some of the interrelations will be mentioned in footnotes. Already we have had occasion to use some of the results and a part of the notation of the calculus of finite differences. It is a subject which deserves, and commonly has, much more space than we can give it in this one chapter, but we can give here the parts which the student will most often find it convenient to use.

Our notation will be explained by the use of an example.

Example 1. Let u be the following function of x: $u(x) = x^3 + 3x^2 + 10$, and suppose it tabulated for $x = 0, 1, \cdots$, as indicated in the first two columns. In the third column, headed Δ, we have the "first differences" of u: $4 = 14 - 10$, $16 = 30 - 14$, etc. In the second column, headed Δ^2, we have the "second differences" of u, which are defined as the first differences of Δ: $12 = 16 - 4$, $18 = 34 - 16$, etc. In the third column, headed Δ^3, we have the "third differences" of u, which are defined as the first differences of Δ^2: $6 = 18 - 12$, $6 = 24 - 18$, etc. Strictly,

x	u	Δ	Δ^2	Δ^3	Δ^4
0	10	4	12	6	0
1	14	16	18	6	0
2	30	34	24	6	
3	64	58	30		
4	122	88			
5	210				

the Δ's are to be thought of as lying on the horizontal lines between the successive u's, the Δ^2's on the lines between the Δ's, etc.

In our theoretical work we shall be thinking either of Table A or of Table B: in the first the difference, h, between the successive x's, is always unity; in the second it is always the same, but not necessarily unity.

TABLE A ($h = 1$)

x	u	Δ	Δ^2	Δ^3
0	u_0	Δ_0	Δ_0^2	
1	u_1	Δ_1	Δ_1^2	Δ_0^3
2	u_2	Δ_2	Δ_2^2	Δ_1^3
3	u_3	Δ_3		
4	u_4			

TABLE B

x	u	Δ	Δ^2	Δ^3
x	$u(x)$	$\Delta u(x)$	$\Delta^2 u(x)$	
$x + h$	$u(x + h)$	$\Delta u(x + h)$	$\Delta^2 u(x + h)$	$\Delta^3 u(x)$
$x + 2h$	$u(x + 2h)$	$\Delta u(x + 2h)$	$\Delta^2 u(x + 2h)$	$\Delta^3 u(x + h)$
$x + 3h$	$u(x + 3h)$	$\Delta u(x + 3h)$		
$x + 4h$	$u(x + 4h)$			

DEFINITIONS: $\Delta u(x) = u(x + h) - u(x)$

$\Delta^2 u(x) = \Delta\Delta u(x)$

$\Delta^3 u(x) = \Delta\Delta\Delta u(x) = \Delta\Delta^2 u(x) = \Delta^2\Delta u(x)$, etc.

The "principal" differences are: Δ_0, Δ_0^2, Δ_0^3, etc.

Example 2. We shall prove later that all the Δ^4's of Example 1 are zero. Assuming this to be true now, find $u(6)$ by the process of addition only. Add one more 0 to the Δ^4 column. This zero must be added to the third 6 in Δ^3 to get the next number in Δ^3, which is also 6. This 6 must be added to 30 to get the fifth number in Δ^2, 36. Then $36 + 88 = 124$, and finally $124 + 210 = 334 = u(6)$.

2. Errors. For most functions, the fourth differences are not all zero, nor is there any value of m such that all the Δ^m's are zero, but for most functions the differences do

become small ultimately. In making tables of functions, one very common method of checking the numerical work is by inspection of the differences.

Example 3. There is an error in one of the following logarithms. Locate it by inspection of the third differences.

x	$\log x$	Δ	Δ^2	Δ^3
5.00	.698 9700			
		8677		
5.01	.699 8377		− 17	
		8660		0
5.02	.700 7037		− 17	
		8643		− 1
5.03	.701 5680		− 18	
		8625		5
5.04	.702 4305		− 13	
		8612		− 11
5.05	.703 2917		− 24	
		8588		11
5.06	.704 1505		− 13	
		8575		− 5
5.07	.705 0080		− 18	
		8557		2
5.08	.705 8637		− 16	
		8541		− 1
5.09	.706 7178		− 17	
		8524		0
5.10	.707 5702		− 17	
		8507		1
5.11	.708 4209		− 16	
		8491		
5.12	.709 2700			

When the table of differences is written in this way, Δ's on the lines between the u's, etc., a single error will manifest itself by producing unusually large differences on the same horizontal line with the error. This will be proved later in a problem. The two 11's in the column Δ^3 indicate that the error lies on the line between them, and it does, for log 5.05 = .7032914 instead of .7032917. It is

almost always true that, when the differences are about equal numerically, but fluctuating in sign, they are as small as they can be made; it is useless to compute higher differences. If it were not for this error, the third differences would have been 0, − 1, 2, − 2, 2, − 2, 2, − 1, 0, 1.

EXERCISES § 2

1. Find the fourth differences of the table of $u = 3x^4 - 20$, using the values at $x = 0, 1, 2, 3, 4, 5$. *Ans.*, 72.

2. Find the error in the following table by using differences: 1065, 1079, 1094, 1109, 1120, 1139, 1154, 1170, 1185, 1201, 1217.

3. Show that

$$\Delta^2 u(x + h) = u(x + 3h) - 2u(x + 2h) + u(x + h).$$

4. By use of Exercise 3 show that

$$\Delta^3 u(x + h) = u(x + 4h) - 3u(x + 3h) + 3u(x + 2h) - u(x + h).$$

5. Difference four times the values of ϕ^{IV} given in the first column of Table IV.

3. Difference Formulae. The following theorems show us the relation between certain functions of x and their differences.

Theorem I. *If $f(x)$ and $g(x)$ are two functions of x, the nth difference of their sum equals the sum of their nth differences:*

$$\Delta^n(f + g) = \Delta^n f + \Delta^n g, \; n = 1, 2, \cdots.$$

Proof.

$$\begin{aligned}\Delta(f + g) &= [f(x + h) + g(x + h)] - [f(x) + g(x)] \\ &= [f(x + h) - f(x)] + [g(x + h) - g(x)] \\ &= \Delta f + \Delta g.\end{aligned}$$

Repeating the process,

$$\Delta\Delta(f + g) = \Delta(\Delta f + \Delta g) = \Delta^2 f + \Delta^2 g; \text{ etc.}$$

Theorem II. *If c is a constant, and f is a function of x,*

$$\Delta^n(cf) = c\,\Delta^n f, \; n = 1, 2, \cdots.$$

Proof.

$$\Delta cf = cf(x + h) - cf(x) = c[f(x + h) - f(x)] = c\Delta f.$$

The process may be repeated.

Theorem III. $\Delta^n(a_0x^n) = a_0h^n n!$, *where* a_0 *is any constant, and* $h = \Delta x$.

Proof. $\Delta(a_0x^n) = a_0[(x + h)^n - x^n] = a_0nx^{n-1}h$ plus terms of lower degree than $(n - 1)$. Differencing a second time, $\Delta^2(a_0x^n) = a_0n(n - 1)x^{n-2}h^2$ plus terms of lower degree than $(n - 2)$. Each repetition of the process lowers the degree by unity and increases the exponent of h by unity and adds one factor to the succession, $n(n - 1)(n - 2) \cdots$. So, after the process has been used n times, $\Delta^n(a_0x^n) = a_0h^n n!$, as desired.

COROLLARY 1. *The* nth *difference of* x^n *is factorial* nh^n.

COROLLARY 2. *The* $(n + 1)$*th difference of* x^n *is zero.*

COROLLARY 3. *The* nth *difference of a polynomial of the* nth *degree in* x *is the same as the* nth *difference of the term in* x^n*:*

$$\Delta^n(a_0x^n + a_1x^{n-1} + \cdots + a_n) = a_0h^n n!.$$

For $\Delta^n(a_1x^{n-1} + \cdots + a_n) = 0$, by Corollary 2.

DEFINITION. *The product* $x(x - h)(x - 2h) \cdots (x - mh + h)$ *is called "factorial* x *to* m *factors" and is written* $x^{(m)}$. *In the particular case where* x *is an integer and* $h = 1$, *this is obviously the same as* ${}_xP_m$; *and if in addition* $m = x$, *it becomes* $x!$.

Theorem[1] **IV.** $\Delta x^{(m)} = mx^{(m-1)}h$.

The proof is left for Exercise 4.

Theorem[2] **V.** (*Newton's formula.*) *If* $f(x)$ *is a polynomial of the* nth *degree in* x, *it may be written in the form of a series of factorials:*

$$f(x) = f(0) + x^{(1)}\Delta_0 + \frac{x^{(2)}}{\lfloor 2}\Delta_0^2 + \frac{x^{(3)}}{\lfloor 3}\Delta_0^3 + \cdots + \frac{x^{(n)}}{\lfloor n}\Delta_0^{(n)},$$

if $\Delta x = 1$.

[1] This is analogous to the calculus formula $\frac{dx^m}{dx} = mx^{m-1}$. In general, $x^{(m)}$ plays in the calculus of finite differences a rôle similar to that of x^m in the calculus of infinitesimal differences. The first is simpler to deal with by finite differences, the second by infinitesimal differences.

[2] This is analogous to Maclaurin's series in the differential calculus.

The proof of the theorem that we shall give depends on an assumption which can be justified, but we shall not give the reasons for it here. We shall assume that $f(x)$ may be written as a series of factorials:

$$f(x) \equiv a_0 + a_1 x^{(1)} + a_2 x^{(2)} + \cdots + a_m x^{(m)}, \text{ where } m \geq n, \quad (1)$$

but that, initially, we do not know what the a's are. We shall now determine what values the a's must have in order that this may be true. Difference (1) m times:

$$\left.\begin{aligned} \Delta f &\equiv a_1 + 2a_2 x + 3a_3 x^{(2)} + \cdots + m a_m x^{(m-1)}, \\ \Delta^2 f &\equiv 2 \cdot 1 a_2 + 3 \cdot 2 a_3 x + \cdots \\ &\qquad + m(m-1) a_m x^{(m-2)}, \\ \Delta^3 f &\equiv 3 \cdot 2 \cdot 1 a_3 + \cdots + m(m-1)(m-2) a_m x^{(m-3)}, \\ &\cdots \\ &\cdots \\ \Delta^m f &\equiv \qquad\qquad \lfloor m a_m. \end{aligned}\right\} \quad (2)$$

Since these identities are all true for all values of x, they must hold when in particular $x = 0$. Put $x = 0$ in (1) and (2); it follows that

$$a_0 = f(0),\ a_1 = \frac{\Delta_0}{\lfloor 1},\ a_2 = \frac{\Delta_0^2}{\lfloor 2},\ \cdots,\ a_n = \frac{\Delta_0^n}{\lfloor n};\ a_m = 0 \text{ if } m > n.$$

These are the values desired.

COROLLARY. *If $\Delta x \neq 1$, let $s = x/\Delta x$. Then $f(x) = u(s)$, and $u(s)$ is a polynomial of the nth degree in s and $\Delta s = 1$, so that*

$$u(s) = u(0) + s^{(1)}\Delta_0 + \cdots + \frac{s^{(n)}}{\lfloor n}\Delta_0^n.$$

This is a very important formula.

EXERCISES § 3

1. Find Δe^x.
2. Find $\Delta(fg)$. *Ans.*, $g(x+h)\Delta f + f(x)\Delta g$.
3. Find $\Delta\left(\frac{1}{g}\right)$.

4. Prove Theorem IV.

5. The symbol $x^{(-m)}$ is used to denote the "inverse factorial":

$$\frac{1}{x(x+h)(x+2h)\cdots(x+mh-h)}.$$

Show that $\Delta x^{(-m)} = (-m)hx^{(-m-1)}$.

6. Write in the form of Newton's series the following polynomials:

(*a*) $x^2 + 2$. *Ans.*, $2 + x + x(x-1)$.

(*b*) $3x^3 + 4x^2 - 1$.

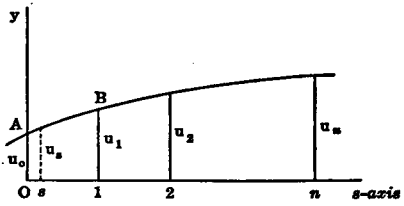
Ans., $-1 + 7x + 13x(x-1) + 3x(x-1)(x-2)$.

(*c*) $4x^4 - 5x^3 + x$.

Ans., $13x(x-1) + 19x(x-1)(x-2) + 4x(x-1)(x-2)(x-3)$.

4. Interpolation. Consider the following geometrical problem: Given the ordinates, u_0, u_1, u_2, $\cdots$, u_n, of $(n+1)$ points of a curve, to pass a parabolic curve of the nth degree through all these points, and to find its ordinate at a point whose abscissa is s $(0 < s < 1)$. By a parabolic curve of the nth degree is meant a curve having the equation

$$y = a_0 + a_1 s + a_2 s^2 + \cdots + a_n s^n.$$



Newton's formula furnishes us with an immediate solution of this problem, provided the given ordinates are equispaced, for it tells us how to find the a's in terms of the differences. An example will make the method clear.

Example 4. Find the ordinate at $s = .2$, given the points indicated in the first two columns of the table at the top of page 355.

s	$u(s)$	Δ	Δ^2	Δ^3
0	.30	.18	− .06	.04
1	.48	.12	− .02	
2	.60	.10		
3	.70			

By Theorem V, for $s = .2$,

$$u(s) = .30 + \left(\frac{.18}{1}\right)(.2) - \left(\frac{.06}{\lfloor 2}\right)(.2)(-.8) + \left(\frac{.04}{\lfloor 3}\right)(.2)(-.8)(-1.8)$$

$$= .30 + .036 + .0048 + .00192 = 0.3427.$$

The graph is the one on page 354. We should now notice the relation of this geometrical problem to the one of interpolation in a table. Heretofore, if we wished to interpolate in the $u(s)$ table opposite $s = .2$, we should have used the principle of proportional parts, and found that

$$u(s) - .30 = (.18)(.2),\ u(s) = .30 + .036 = 0.336.$$

This is equivalent to using the first two terms of our longer equation, and this in turn is the same as finding the ordinate of the straight line which would pass through the points A and B of our given curve. This is also called interpolation by means of first differences. Now it is geometrically evident that it would be better usually to use a smooth curve going through several of the points rather than a line through two of them, and this is what we have done in Example 4.

Example 5. Find $e^{1.7543}$ from the following table, using third differences.

t	e^t	Δ	Δ^2	Δ^3
1.7	5.474	.576	.060	.007
1.8	6.050	.636	.067	.007
1.9	6.686	.703	.074	.008
2.0	7.389	.777	.082	.008
2.1	8.166	.859	.090	
2.2	9.025	.949		
2.3	9.974			

In this case we have to make use of the corollary to Theorem V. Take a new variable, say s, with origin at $t = 1.7$, and let $s = 1$ at $t = 1.8$, so that when $t = 1.7 + .0543$, $s = .0543/.1 = .543$, $u(s = 0) = 5.474$, $\Delta_0 = .576$, $\Delta_0^2 = .060$, $\Delta_0^3 = .007$. So our formula gives us

$$u(.543) = 5.474 + \frac{.576}{1}(.543) + \frac{.060}{2}(.543)(-.457) + \frac{.007}{6}(.543)(-.457)(-1.457) = 5.7797.$$

The true value is 5.7794.

5. Backward Interpolation. The successful use of Newton's formula depends in part on the table actually being given for points beyond $s = 1$. When we get to the end of a table and wish to interpolate, of course there are no such points given. But, since a table does not need to run down the page instead of up it, we may in such a case simply reverse the table so that the end becomes the beginning and proceed as before. If we have already computed the differences for the table as given initially, we do not need to compute them over again, for a little reflection will assure us that the new Δ, Δ^3, Δ^5, etc., will be the same as before except that their signs will be reversed, and that the new Δ^2, Δ^4, etc., will be exactly the same as before.

Example 6. Find the ordinate at $s = 2.3$ in Example 4.

Old s	*New s*	*u(s)*	*New Δ*	*New Δ²*	*New Δ³*
3	0	.70	− .10	− .02	− .04
2	1	.60	− .12	− .06	
1	2	.48	− .18		
0	3	.30			

We interpolate at new $s = .7$:

$$u(.7) = .70 + \frac{.7}{1}(-.10) + \frac{(.7)(-.3)}{2}(-.02) + \frac{(.7)(-.3)(-1.3)}{6}(-.04) = 0.6303.$$

EXERCISES § 5

1. Find $e^{1.65}$ from the table in Example 5, § 4. *Ans.*, 6.3600.

2. Find $\phi(.31)$ from Table I(a), using only the values given in the first column. *Ans.*, .3802.

3. Similarly, find $\phi(1.732)$. *Ans.*, .0890.

4. Similarly, find from the first column of Table II: $\phi^{(2)}(.73)$, also $\phi^{(2)}(2.175)$.

5. From Table IV find $\phi^{(4)}(.42)$, $\phi^{(4)}(3.335)$.

Ans., .7206, .09198.

6. The amounts which \$1000 will be worth after 11 years if put at compound interest at varying rates are:

Per Cent	5	6	7	8
Value	\$1710.34	\$1898.30	\$2104.85	\$2331.64

Find the amount after 11 years at 5.2%. *Ans.*, \$1746.52.

7. Use Newton's formula to find the equation of the parabola which will go through the points (0, 2), (2, 5), (4, 6), and check your result.

8. A parabola goes through the points (10, 2), (13, 5), (16, 4). Find the ordinate of that point whose abscissa is 10.5. *Ans.*, $2\frac{1}{8}$.

9. Find $e^{2.27}$ from the table in Example 5, § 4. *Ans.*, 9.679.

10. Find the ordinate of the point (a) at $x = 13.3$ in Exercise 8, (b) at $x = 3.8$ in Exercise 7.

11. Find the value of \$1000 after 11 years at compound interest at $7\frac{1}{2}$% from the table in Exercise 6. *Ans.*, \$2215.61.

6. Inverse Interpolation. In "coming out" of an ordinary table, in which the argument changes by equal increments but the function changes by unequal increments, we usually have to interpolate. In doing so we interchange the rôles of the argument and function, so that our new argument does not change by equal increments although the new function does. This is called inverse interpolation. We cannot use our previous methods (except linear inter-

polation), and the problem is often a difficult one. Let us indicate its nature by the use of a numerical example.

Example 7. Find the value of x for which $u = 20$ in Example 1.

x	u
0	10
1	14
	←20
2	30
3	64

Linear interpolation gives $x = 1\frac{3}{8}$, but this is not close to the correct value, for $u(1\frac{3}{8}) = (1\frac{3}{8})^3 + 3(1\frac{3}{8})^2 + 10 = 18.271 \neq 20$.

Of the many methods that have been devised to solve this problem, we shall describe here only one. It can be used quickly and will give a much better result than can be obtained by linear interpolation.[1] It consists in dividing the interval, over which interpolation must be made, into four equal parts and then using linear interpolation.

COROLLARY TO THEOREM V. *If the argument and function are displayed as indicated below, formulae for interpolation at the points* $x = 1\frac{1}{4}$, $x = 1\frac{1}{2}$, $x = 1\frac{3}{4}$ *are:*

(*a*) *at* $x = 1\frac{1}{4}$, $u = \frac{1}{128}[-7u_0 + 105u_1 + 35u_2 - 5u_3]$;

(*b*) *at* $x = 1\frac{1}{2}$, $u = \frac{1}{16}[-u_0 + 9u_1 + 9u_2 - u_3]$;

(*c*) *at* $x = 1\frac{3}{4}$, $u = \frac{1}{128}[-5u_0 + 35u_1 + 105u_2 - 7u_3]$.

ARGUMENT AND FUNCTION

x	u	Δ	Δ^2	Δ^3
0	u_0			
		u_1-u_0		
1	u_1		$u_2-2u_1+u_0$	
		u_2-u_1		$u_3-3u_2+3u_1-u_0$
2	u_2		$u_3-2u_2+u_1$	
		u_3-u_2		
3	u_3			

[1] But too much must not be expected of it. Longer but more accurate methods are given in *Tracts for Computers*, No. 2, K. Pearson. The present method was suggested by this tract, but is not included in it.

These results follow immediately from the theorem by substitution of the proper values. We prove in detail (*a*) only:

$$u\left(\frac{5}{4}\right) = u_0 + \frac{5}{4}(u_1 - u_0) + \frac{5}{4}\cdot\frac{1}{4}\cdot\frac{1}{\underline{2}}(u_2 - 2u_1 + u_0)$$

$$+ \frac{5}{4}\cdot\frac{1}{4}\cdot\frac{-3}{4}\cdot\frac{1}{\underline{3}}(u_3 - 3u_2 + 3u_1 - u_0)$$

$$= \frac{1}{128}(-7u_0 + 105u_1 + 35u_2 - 5u_3).$$

With the aid of a machine, these formulae can be evaluated rapidly, and it is a little better to write $\frac{1}{128} = 0.0078125$, $\frac{1}{16} = 0.0625$. The student should prepare a small card as indicated in the figures. This should be placed alongside the table in which the interpolation is being made.

FRONT OF CARD

	− 7
$(\frac{1}{4})\rightarrow$	105
	35
	− 5
	.0078125

BACK OF CARD

	− 1
$(\frac{1}{4})\longrightarrow$	9
	9
	− 1
	.0625

Continuing Example 7, now, we first find the value of u that would come opposite $x = 1\frac{1}{4}$. Place the front of the card alongside the u's, multiply and add and then multiply the sum by .0078125, as indicated. Of course, all the work is

FRONT OF CARD

	− 7
$(\frac{1}{4})\longrightarrow$	105
	35
	− 5
	.0078125

u	Products
10	− 70
14	+ 1470
30	+ 1050
64	− 320
	Sum = 2130

(2130)(.0078125) = 16.6406.

done on the machine, nothing appearing on the paper except the final answer. Now if our answer were greater than 20 we should not need to obtain also the value of u corresponding to $x = 1\frac{1}{2}$. As it is less than 20 we must do so, using the back of the card. This value is 20.1250. We now have the following table in which we use linear interpolation to find the x required:

x	u	
1.25	16.6406	
		←—20.0000
1.50	20.1250	

$x = 1.4910$. As a check, we may interpolate directly in the original table opposite this value of x, or in this case substitute the value of x in the formula from which the table was made. We find that $u = 19.98$. Thus our required value of x is now known to be a little greater than the one we have found. By successive trials we may obtain it to as great accuracy as desired. To six decimal places it is 1.492033.

Example 8. Find x if $\log x = .715000$ from the following six-place table.

x	$\log x$	$\log x - .698$
5.0	.698 970	970
5.1	.707 570	9570
5.2	.716 003	18003
5.3	.724 276	26276

This differs from the previous example in that the x's as given do not exactly fit the formulae, but all that was really necessary was that they should be equi-spaced, and they are equi-spaced. It is simpler to use the numbers ($\log x - .698$) and to omit the decimal points. In the third column, then, we are required to interpolate to find the position of 17,000. This is so close to 18,003 that we try first the point at 5.175. By formula (c) it is 15,910. We have then the table:

5.175	15,910	
	←	—17,000.
5.200	18,003	

Interpolating linearly in this we find $x = 5.18802$, and the true value to six figures is 5.18800. Linear interpolation if used initially would have given 5.18811.

Example 9. Solve the equation $x^3 + 3x^2 + 10 = 0$, so as to find all the real roots.

Let $f(x) = x^3 + 3x^2 + 10$. We are required to find those values of x for which $f(x)$ is zero. First we make a table (A) of $f(x)$.

TABLE A

x	$f(x)$
− 5	− 40
− 4	− 6
− 3	10
− 2	14
− 1	12
0	10
1	14
2	30

From this table we find that $f(x) = 0$ probably for some value of x between − 4 and − 3, and we are required to interpolate so as to find that x. We now confine our attention to the first four rows of the table and proceed as before, obtaining Table B.

TABLE B

x	$f(x)$	
− 4	− 6	
− 3.75	− 0.54687	
	←	—0
− 3.50	3.87500	

Interpolating so as to make $f(x) = 0$, we find $x = -3.719$. The true value to four figures is − 3.7219. This indicates an error of

about 3 in the fourth figure. Better accuracy may now be obtained by substituting the value found in the equation and other values near it, so as to form a new table in the closer vicinity of the root. However, this is not recommended, because, if a high degree of accuracy is required, there are other methods which should be used instead. The root found above is the only real root this equation has.

EXERCISES [1] § 6

1. In Example 7, find x if $u = 29$. *Ans.*, 1.954.

2. In Example 8, find x if $\log x = 0.710{,}000$. *Ans.*, 5.1286.

3. In Example 9, find x if $f(x) = 40$. *Ans.*, 2.360.

4. In Table IV, find the smallest value of x for which $\phi^{(4)}(x) = 0.8000$. *Ans.*, .3797.

5. In Table IV, find x if $\phi^{(4)}(x) = -0.7000$. *Ans.*, 1.213.

6. In Table IV, find the smallest value of x for which $\phi^{(4)}(x) = 0.1000$.

7. Find the median in the frequency table of Example 1, Part I, Chapter III, page 36, using a more refined method than simple interpolation.

8. In Exercise 6 of § 5, find the rate of interest for which the value of \$1 will be \$2 after 11 years.

7. Summation of Series. The problem of this section is one of finding a compact formula which will express the sum

x	U	u	Δu	$\Delta^2 u$
0	U_0	u_0	Δ_0	Δ_0^2
1	U_1	u_1	Δ_1	Δ_1^2
2	U_2	u_2	Δ_2	Δ_2^2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	U_n	u_n	Δ_n	Δ_n^2

[1] The answers given are not necessarily the true values. They are the values that should be found by the methods to be used. The true value of x in Exercise 1 is about 1.958.

of the first n terms of a given series. We had to use such a formula in Part I, Chapter VII, § 3, for the sum $1^2 + 2^2 + \cdots + \left(\frac{n-1}{2}\right)^2$, and we asserted that it was $\frac{n(n+1)(n-1)}{24}$. In the general case, we will insert in our difference table (page 362) a column of U's of which the u's are the first differences, and we have the following theorem: [1]

Theorem VI. *The sum of the first n terms of the u series is* $U_n - U_0$; *in symbols:*

$$\sum_{x=0}^{n-1} u = U_x\Big]_0^n.$$

The proof is very simple. Since $u_0 = U_1 - U_0$, etc.,

$$u_0 + u_1 + \cdots + u_{n-1} = (U_1 - U_0) + (U_2 - U_1) + \cdots + (U_n - U_{n-1}),$$

and in this latter sum all the terms cancel in pairs except $-U_0$ and U_n.

Application. To apply this theorem as it stands, it is necessary to know what U_x is,[2] and unfortunately we have at our disposal but a few simple cases in which we do know U_x. These are indicated by difference formulae similar to those of § 3. The most important of these is (Theorem IV) $\Delta x^{(m)} = mx^{(m-1)}h$. Hence, if $u_x = x^{(m-1)}$,

$$U_x = \frac{x^{(m)}}{mh}.$$

We may apply this to an example like the following.

Example 10. Find the sum of 20 terms of the series,

$$20 \cdot 19 \cdot 18 + 21 \cdot 20 \cdot 19 + \cdots + 39 \cdot 38 \cdot 37.$$

[1] It is analogous to the familiar relation between the definite and the indefinite integral in calculus. U_x plays the part of the indefinite integral. In fact, $U_x\Big]_0^n$ may be called the definite sum and U_x the indefinite sum.

[2] Just as in integration, it is necessary to know the indefinite integral.

This series is of the form $20^{(m-1)} + 21^{(m-1)} + \cdots + 39^{(m-1)}$, h being 1 and $m - 1$ being 3; and if the origin be taken so that $u_0 = 20^{(m-1)}$, the sum of the first ($n = 20$) terms is $U_{20} - U_0$. Since

$$U_0 = \frac{20^{(m)}}{m}, \text{ and } U_{20} = \frac{40^{(m)}}{m},$$

the desired sum is

$$\tfrac{1}{4}[40^{(4)} - 20^{(4)}] = 519{,}270.$$

Whenever the u function is a polynomial in x, the sum can be found by a neat application of Newton's formula. The idea[1] is: Express the polynomial in a series of factorials and then, by the method of the theorem, sum each individual term.

COROLLARY. *If $u(x)$ is a polynomial of the kth degree in x,*

$$u_0 + u_1 + \cdots + u_{n-1} = u_0 \frac{n^{(1)}}{\underline{|1}} + \Delta_0 \frac{n^{(2)}}{\underline{|2}} + \cdots + \Delta_0^k \frac{n^{(k+1)}}{\underline{|k+1}}$$

Proof. By Newton's formula,

$$u_x = u_0 + \Delta_0 \frac{x^{(1)}}{\underline{|1}} + \Delta_0^2 \frac{x^{(2)}}{\underline{|2}} + \cdots + \Delta_0^k \frac{x^{(k)}}{\underline{|k}},$$

and so, by the theorem,

$$\sum_0^{n-1} u_x = nu_0 + \frac{\Delta_0}{\underline{|1}} \sum_0^{n-1} x^{(1)} + \frac{\Delta_0^2}{\underline{|2}} \sum_0^{n-1} x^{(2)} + \cdots + \frac{\Delta_0^k}{\underline{|k}} \sum_0^{n-1} x^{(k)}.$$

But each individual sum on the right is of the form

$$\sum_0^{n-1} x^{(r)} = \frac{x^{(r+1)}}{r+1}\Big]_0^n = \frac{n^{(r+1)}}{r+1},$$

and so

$$\sum_0^{n-1} u_x = u_0 n + \frac{\Delta_0}{\underline{|2}} n^{(2)} + \frac{\Delta_0^2}{\underline{|3}} n^{(3)} + \cdots + \frac{\Delta_0^k}{\underline{|k+1}} n^{(k+1)},$$

the form desired.

Example 11. Find the sum of $1^2 + 2^2 + \cdots + n^2$. To apply the corollary we notice that the general term is of the form x^2, a

[1] This is analogous to the following method of integrating: Expand the integrand in a power series, and integrate each term separately.

simple polynomial in x; but the sum does not begin with 0^2. We might prefix a 0^2 term, but in that case we should have $(n+1)$, not n terms. Rather than make the necessary substitutions it is a little simpler, both in this case and in most cases, to think of the difference table, and to use the principal differences, omitting the x altogether:

$u_0 = 1^2$	$\Delta_0 = 3$	$\Delta_0^2 = 2$	$\Delta_0^3 = 0$
$u_1 = 2^2$	$\Delta_1 = 5$	$\Delta_1^2 = 2$	
$u_2 = 3^2$	$\Delta_2 = 7$		
.			
.			
.			
$u_{n-1} = n^2$			

Then,

$$u_0 + u_1 + \cdots + u_{n-1} = n \cdot 1 + 3\frac{n(n-1)}{2} + 2\frac{n(n-1)(n-2)}{6}$$
$$= \frac{n(n+1)(2n+1)}{6}.$$

Example 12. Find the sum of the first n terms of the series of Example 1.

There we had $u_0 = 10,\ \Delta_0 = 4,\ \Delta_0^2 = 12,\ \Delta_0^3 = 6,$ and so

$$u_0 + u_1 + \cdots + u_{n-1} = 10n + \frac{4}{2}n^{(2)} + \frac{12}{\lfloor 3}n^{(3)} + \frac{6}{\lfloor 4}n^{(4)}$$
$$= n[10 + 2(n-1) + 2(n-1)(n-2) + \tfrac{1}{4}(n-1)(n-2)(n-3)].$$

E.g., if $n = 100$, the sum is easily found to be 25,488,550.

An incidental value of Theorem VI is that it affords a numerical check on the work of differencing. The sum of a set of consecutive differences of the $(k+1)$th order should equal the difference between the first and the last differences of the kth order.

EXERCISES § 7

1. Find the sum, $1 + 2 + 3 + \cdots + n$.

2. Find the sum, $1^3 + 2^3 + \cdots + n^3$. *Ans.*, $\frac{n^2(n+1)^2}{4}$.

3. Find the sum, $1^2 + 3^2 + 5^2 + \cdots + (2n-1)^2$.

Ans., $\frac{n}{3}(4n^2 - 1)$.

4. Find the sum, $2^2 + 4^2 + 6^2 + \cdots + (2n)^2$.

Ans., $\frac{2n}{3}(1+n)(1+2n)$.

5. Use Theorem VI to prove that

$$a^{(3)} + (a+1)^{(3)} + \cdots + (a+n-1)^{(3)} = \frac{(a+n)^{(4)} - a^{(4)}}{4}.$$

6. Find the sum of 100 terms of the sequences:

(*a*) $-2, -1, 2, 7, 14, \cdots$. *Ans.*, 328,150.

(*b*) 11, 30, 67, 128, 219, $\cdots$. *Ans.*, 26,533,100.

7. Check the answers to Exercises 3 and 4 by adding them together and comparing with Example 11.

8. Prove formula (10) of Chapter VII, Part I, page 108, both for the case where n is odd and where n is even.

9. (*a*) Find the sum, $1^4 + 2^4 + \cdots + n^4$.

Ans., $\frac{n(n+1)(2n+1)(3n^2+3n-1)}{30} = \frac{1}{5}(1^2 + 2^2 + \cdots + n^2)(3n^2 + 3n - 1)$.

(*b*) Find the sum, $1^4 + 3^4 + \cdots (2n-1)^4$.

Ans., $\frac{1}{5}[1^2 + 3^2 + \cdots + (2n-1)^2](12n^2 - 7)$.

10. Prove formula (20) of Chapter VII, Part I, page 123.

11. Use the method of Example 12 to check the values of the differences you found for Table IV in Exercise 5 of § 2.

PROBLEMS CHAPTER VII

1. Show that the table of binomial coefficients may be looked upon as a table of successive differences. Express the essential fact here as a special case of Theorem IV.

TABLE OF BINOMIAL COEFFICIENTS

n	${}_nC_0$	${}_nC_1$	${}_nC_2$	${}_nC_3$	${}_nC_4$	.	.	.
0	1							
1	1	1						
2	1	2	1					
3	1	3	3	1				
4	1	4	6	4	1	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

2. (*a*) Show that, if $u(x)$ is any polynomial of the second degree, and if, opposite some particular value x' of x, an error is introduced so that $u(x')$ is written $u(x') + \epsilon$, then Δ^3 will be 0 except for a set of numbers, ϵ, -3ϵ, 3ϵ, $-\epsilon$, symmetrically situated with respect to the line $x = x'$.

(*b*) Prove the corresponding theorem for a polynomial of the nth degree.

3. Find $\Delta({}_nP_x)$ with respect to x. *Ans.*, ${}_nP_x(n - x - 1)$.

4. Find $\Delta({}_nC_x)$ with respect to x. *Ans.*, ${}_nC_x\left(\frac{n - 2x - 1}{x + 1}\right)$.

5. Find $\Delta({}_nP_x)$ with respect to n. *Ans.*, $x\,{}_nP_{x-1}$.

6. Find $\Delta({}_nC_x)$ with respect to n. *Ans.*, ${}_nC_{x-1}$.

7. Find $\Delta({}_nC_t p^{n-t}q^t)$ with respect to t.

Ans., ${}_nC_t q^t p^{n-t}\left(\frac{n - t}{t + 1} \cdot \frac{q}{p} - 1\right)$.

8. Show that $\Delta u(x)v(x) = u(x)\Delta v(x) + v(x + 1)\Delta u(x)$, if $\Delta x = 1$.

9. Show that $\Delta \sin(x - \frac{1}{2})a = 2 \sin \frac{a}{2} \cos xa$, if $\Delta x = 1$.

10. Hence, prove that $\frac{1}{2} + \sum_{x=1}^{n} \cos xa = \frac{\sin(n + \frac{1}{2})a}{2 \sin \frac{a}{2}}$.

This is a useful formula in the theory of trigonometrical series.

11. From (8) show that,

$$\sum_{x=0}^{n-1} u(x)\Delta v(x) = u(x)v(x)\Big]_0^n - \sum_{x=0}^{n-1} v(x+1)\Delta u(x).$$

This is analogous to integration by parts in the calculus of infinitesimal differences.

12. From Exercise 5 of § 3, find by the use of a summation formula the sum

$$\frac{1}{10\cdot 11\cdot 12} + \frac{1}{11\cdot 12\cdot 13} + \cdots + \frac{1}{50\cdot 51\cdot 52}.$$

13. The calculated ranges and times for a high angle gun at various elevations are given below. (K. Pearson, *Tracts for Computers*, No. 2.)

Elevation	*Range in Yards*	*Time in Seconds*
30°	32 935	40.06
35°	34 226	45.02
40°	34 843	49.69
45°	34 764	54.07
50°	33 963	58.12

(*a*) Find the range and time of flight for an elevation of 32°.

(*b*) Find the elevation and range when the time of flight is 50.00 seconds.

(*c*) Find the ranges for a table constructed from the following elevations: 37.50°, 38.75°, 40.00°, 41.25°, 42.50°, 43.75°.

(*d*) Estimate graphically the maximum range for this gun, using the results of (*c*).

14. Tables of the moon's positions (declinations) for March 25 are given as follows:

Hour	*Declination (South)*
0	10° 07′ 07.9″
1	9 58 56.4
2	9 50 41.8
3	9 42 24.3
4	9 34 03.7

An astronomer observes the declination to be 9° 55′ 0″. What time is it?

15. Solve approximately the equation, $x \sin x = 1$, by tabulating the function at the points $x = 1.0, 1.1, 1.2, 1.3$ radians, and interpolating.

16. A portion of Pearson's χ-test table for P is (to four places):

χ^2	$n' = 6$
1	.9626
2	.8491
3	.7000
4	.5494
5	.4159

Find P if $n' = 6$ and $\chi^2 = 2.42$.

17. From the table in Problem 16, find χ^2 such that $P = 0.8$.

18. Find the sum to n terms of the series:

$$1^3 + 4^3 + 7^3 + 10^3 + \cdots.$$

19. Use Problem 3 to find $\sum_{x=0}^{n-1} {}_nP_x(n - x - 1)$.

20. (*a*) Show that

$${}_nC_i - {}_nC_{i-1} + {}_nC_{i-2} - {}_nC_{i-3} + \cdots \pm {}_nC_0 = {}_{n-1}C_i$$

by successive use of the relation that

$${}_nC_r = {}_{n-1}C_r + {}_{n-1}C_{r-1}.$$

(*b*) What is the value of

$${}_nC_n + {}_nC_{n-1} + \cdots + {}_nC_0?$$

(Cf. page 201.)

(*c*) What is the value of ${}_nC_n - {}_nC_{n-1} + \cdots \pm {}_nC_0$?

PART III: FOUR-PLACE TABLES OF PROBABILITY FUNCTIONS

FOUR-PLACE TABLES OF PROBABILITY FUNCTIONS

1. Preface. The purpose of these tables is to enable the student, or the practicing statistician, to obtain quickly approximate answers to certain simple problems in the theory of probability. They have been compiled, partly by compressing certain longer tables, partly by independent computation. The method of tabulation follows that of Huntington's *Four-Place Tables of Logarithms and Trigonometric Functions*. Interpolation in such tables is extremely easy. Thus, instead of using both Table I and Table II of Pearson's *Tables for Statisticians* in order to find the partial area of a normal curve when the abscissa is given, and also the abscissa when the partial area is given, it is here necessary to use but one table; for interpolation is just as easy if one is coming out of these tables as it is if one is entering them. In the introductory material preceding the tables, the exact formulae to be tabulated are given, and also, in each instance, an illustration of the use to which the table may be put. In particular, it is shown (p. 377) that the tables are sufficient for the approximate solution of the very important problem of the skew point binomial.

2. Explanation of the Tables. Table I. This is a tabulation of the area,

$$\tfrac{1}{2}(1 + \alpha) = \int_{-\infty}^{x} \phi(x)dx, \tag{1}$$

where $\phi(x)$ is the normal function,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \tag{2}$$

in such units that the standard deviation and the total area under the curve are both unity. Here, in the notation used by Sheppard and Pearson, α stands for the area,

$$\alpha = \int_{-x}^{x} \phi(x)dx. \tag{3}$$

For the explanation of the column headed "e," see page 378.

Application. The point binomial $(p + q)^n$ may be written in the form:

$$(p + q)^n = p^n + \cdots + {}_nC_s p^s q^t + \cdots + q^n = u(o) + \cdots + u(t) + \cdots + u(n). \tag{4}$$

Its mean t, mean s, and standard deviation are:

$$\bar{t} = nq,\ \bar{s} = np,\ \sigma = \sqrt{npq}. \tag{5}$$

If p is not very far from $\frac{1}{2}$, and if n is large, the sum of those terms in which s differs from $\bar{s}$ (or, what is the same thing, t differs from $\bar{t}$) by k or less is, approximately,

$$2\int_o^x \phi(x)dx, \quad \text{where } x = (k + \tfrac{1}{2})/\sigma. \tag{6}$$

Also, the sum of those terms in which s differs from $\bar{s}$ by k or more is, approximately,

$$2\int_x^\infty \phi(x)dx, \quad \text{where } x = (k - \tfrac{1}{2})/\sigma. \tag{6a}$$

In both (6) and (6a), k must be the differences between $\bar{s}$ and two actually occurring exponents s. These exponents are integers, but usually $\bar{s}$ and k are not integers.[1] For numerical illustrations see pages 214–216.

Table I(a). This is a tabulation of the function called $\phi(x)$ in equation (2).

Application. If $p = q$, the term $u(t)$ of the point binomial (4) is

$$u(t) = \frac{n!}{s!t!} \cdot \frac{1}{2^n}. \tag{7}$$

This may be computed outright by the use of Table V, but, if n is fairly large, a close approximation is:

$$u(t) = \frac{1}{\sigma}\phi(x), \text{ where } \sigma = \tfrac{1}{2}\sqrt{n}, \text{ and } x = \frac{|t - n/2|}{\sigma}. \tag{8}$$

Tables II, III, and IV. These are tabulations of the second and third derivatives of $\phi(x)$, $\phi^{(2)}(x) = (x^2 - 1)\phi(x)$, $\phi^{(3)}(x) = (3x - x^3)\phi(x)$, and $\phi^{(4)}(x) = (x^4 - 6x^2 + 3)\phi(x)$.

[1] In practice the condition is seldom exactly satisfied, but often sufficiently nearly satisfied to permit the formula to be a good approximation.

Application. Let the sum of the first $(t+1)$ terms of the point binomial (4) be indicated by

$$S_{t+1} = u(0) + \cdots + u(t). \tag{4a}$$

Approximately,

$$S_{t+1} = \int_x^\infty \phi(x)dx + \frac{q-p}{6\sigma}\phi^{(2)}(x) - \frac{1}{24}\left(\frac{1}{\sigma^2} - \frac{6}{n}\right)\phi^{(3)}(x),$$

where $$x = \frac{s - \frac{1}{2} - np}{\sigma},\ s = n - t. \tag{9}$$

Also, the sum of the last $(s+1)$ terms is approximately equal to

$$\int_x^\infty \phi(x)dx - \frac{q-p}{6\sigma}\phi^{(2)}(x) - \frac{1}{24}\left(\frac{1}{\sigma^2} - \frac{6}{n}\right)\phi^{(3)}(x),$$

where $$x = \frac{t - \frac{1}{2} - nq}{\sigma}. \tag{9a}$$

The sum of those terms in which t lies in the interval $a \leqq t \leqq b$, a and b being integers, is approximately equal to

$$\int_{x_1}^{x_2} \phi(x)dx + \left[\frac{q-p}{6\sigma}\phi^{(2)}(x) + \frac{1}{24}\left(\frac{1}{\sigma^2} - \frac{6}{n}\right)\phi^{(3)}(x)\right]_{x_1}^{x_2}, \tag{9b}$$

where the bracket has the same meaning as in the calculus, thus,

$$\Big[f(x)\Big]_{x_1}^{x_2} = f(x_2) - f(x_1),$$

and $$x_1 = \frac{a - \frac{1}{2} - nq}{\sigma},\ x_2 = \frac{b + \frac{1}{2} - nq}{\sigma}.$$

Equations (9), (9*a*), and (9*b*) may all be used to solve any problem for which any one of them may be used, but sometimes one form is more convenient than the others. For numerical illustrations see pages 219-221.

Table IV is used in § 3, Chapter III, Part II, of the text for the purpose of plotting a Gram-Charlier curve.

Table V. This is a tabulation of the logarithm of factorial n, from $n = 0$ to $n = 509$. If $n > 509$, one may use the approximate formula, $$\log n! = (n + \tfrac{1}{2})\log n - n \log e + \log \sqrt{2\pi}, \tag{10}$$

where $\log e = 0.43429\ 4482$, and $\log \sqrt{2\pi} = 0.39909$. This formula yields also the following approximation to the logarithm of a binomial coefficient:

$$\log {}_nC_r = (n + \tfrac{1}{2}) \log n - (n - r + \tfrac{1}{2}) \log (n - r) - (r + \tfrac{1}{2}) \log r - \log \sqrt{2\pi}. \tag{11}$$

Application. If t is considerably less than nq in the point binomial formula (4a), the following is often a close approximation:

$$S_{t+1} = u(t)/(1 - Q), \text{ where } Q = tp/(s + 1)q. \qquad (12)$$

Also, it is known that the value of S_{t+1} found by this formula is never less than the true value. The computation of $u(t)$ requires the use of Table V, or else of formula (11). If the sum of the *last* r terms of the binomial is desired, it is best to restate the problem, interchanging the values of p and q, so that it reads: Find the sum of the first r terms of the point binomial, $(new\ p + new\ q)^n$. Then a *new* t is obtained from the relation that $r = new\ t + 1$, and a *new* Q from the relation,

$$Q = \frac{new\ t}{new\ s + 1} \cdot \frac{new\ p}{new\ q},$$

and, finally, $u(t)$ is to be computed from the new values of s, t, p, and q.

Table VI. If the area under the normal curve $\phi(x)$ be divided into N equal portions by means of $N - 1$ vertical lines, the mean x of each portion is given by Table VI. This table is therefore useful in giving the normalized position of each number in an ordered series.

Table VII. R_x is defined as the ratio, area under the normal curve $\phi(x)$ from x to infinity, divided by the ordinate at x:

$$R_x = \frac{1}{\phi(x)} \int_x^\infty \phi(x)dx. \qquad (13)$$

Application 1. The mean value of the area under the normal curve from x to infinity equals $1/R_x$. This is the quantity called z/q in the Kelley-Wood table.[1]

[1] For most of the purposes for which the Kelley-Wood table is used, sufficient accuracy will be obtained by the use of Tables I and I(a) of this set and a slide rule. Kelley's argument p is our $(1 + \alpha)/2$ of Table I. His alternative arguments, q and I, are, respectively, $q = 1 - p$ and $I = p - 0.5$. His x is our x, and his z is our ϕ. So, given his argument p, come out of Table I to obtain x. Then enter Table I (a) to get his z. Then obtain the ratios, z/p and z/q, by a slide rule. On the other hand, neither his tabulation of z/q nor Sheppard's tabulation of z and q is sufficient for the purpose of Application 2 of Table VII. But see also footnote, page 101.

Application 2. If $t < nq$, in the notation for the point binomial used in (4) and (4*a*),

$$S_{t+1} = \frac{n!\, R_x}{(s-1)!\, t!\, d} \cdot p^{s-1} q^t, \text{ approximately,} \qquad (14)$$

where, in the order suited to computation, $a = (s-1)/p$, $b = t/q$, $c = a - b$, $d^2 = a/p + b/q$, $x = c/d$. It is important that the value obtained for x be accurate to the third place of decimals. To obtain the sum of the last r terms by this formula, it is best to restate the problem, as in the Application of Table V, interchanging the values of p and q. In the new problem, one will be required to find the first r terms of the new point binomial, $(new\ p + new\ q)^n$, and $t + 1$ will be equal to r. If $n > 509$, the formula (10) for computing the factorials yields us, instead of (14),

$$\log S_{t+1} = (n + \tfrac{1}{2}) \log n - (s-1) \log a - t \log b - \tfrac{1}{2} \log (s-1) - \tfrac{1}{2} \log t - \log d + \log R_x - 0.83338. \qquad (14a)$$

For numerical illustrations see *Biometrika*, vol. 16, pages 169, 170.

3. Rules for the Skew Binomial. Sometimes one of the formulae given in § 2 for the sum of a number of consecutive terms of a point binomial gives the best result, sometimes another. Practice indicates that the following is a good rule:

RULE: *Case* I ($t \leqq nq$, *bounding ordinate to the left of mean*).

(*a*) Use (12) when $Q \leqq 1/10$, or when $x \geqq 8$, where $x = (nq - t)/\sigma$.

(*b*) Use (14) or (14*a*) when $1/10 < Q \leqq \frac{1}{2}$, or when $5 \leqq x < 8$.

(*c*) Use (9), or (9*a*), or (9*b*) elsewhere.

(*d*) If $n < 20$, enough terms may be computed outright without excessive labor, provided the work is properly organized and a computing machine is available.

Case II ($t > nq$). In this case the problem may be restated, with interchanged values of p and q. Then the *new* t will be found to be less than the *new* nq. A *new* Q is then obtained, and a *new* x. For these new quantities the rules of Case I hold.

Summary. In general, therefore, if the bounding ordinate of the tail area to be found is far from the mean, use (12); if at a moderate distance, use (14) or (14*a*); if near the mean, use (9), or (9*a*), or

(9*b*). The last type of formula may be used over a greater range on fairly symmetrical functions than on very skew ones.[1]

4. Accuracy of the Tables. In these tables there is a possible error of about .55 in the last place kept. This is a departure from the best practice, in which the error allowed in the last place kept is 0.50. The reasons for this slight departure are that these tables are essentially abbreviated forms of longer tables which would be available usually if really needed, and that the extra accuracy cannot be maintained uniformly in a table without long computations at occasional points. In addition to this small error, there is a larger possible error which the user can make when he interpolates by means of the tenths of the tabular difference given at the right of his line. These are the tenths of the average tabular difference on that line, not necessarily of the actual tabular difference which he should use. The maximum total error that can result from a combination of both these causes has been computed and is given in the column headed "*e*." If one is not satisfied with the accuracy thus obtainable, he always has the option of interpolating in the usual way by finding his own tabular difference. Further, if he

[1] The author has devised these rules and chosen these tables as the most feasible and generally applicable method of obtaining a good approximation quickly to the sum of a group of terms of the skew point binomial, after much experimenting with various formulae and methods. Certain other formulae have not been chosen because their ranges of applicability were not sufficiently great, or because their use would require the tabulation of special functions which would not be so useful in an elementary course.

The most obvious choice of such special functions would be the partial area under Pearson's Type III curve, for this curve was originally devised to fit the point binomial. This function has been tabulated in part by Pearson (*Tables of the Incomplete Gamma Function*), and, in a form better suited to this use, by Salvosa (*The Annals of Mathematical Statistics*, vol. 1 (1930), p. 191). The accuracy and speed of Salvosa's table is comparable with the use of formula (9); sometimes one method is a little better, sometimes the other. For the types of cases to which it is suggested that it be applied, formula (14) is clearly superior in accuracy to either, although the computation is quite a little longer. The tables of the incomplete Gamma function are particularly useful when p or q is very small and n is very large.

is to be sure of keeping within the limit of error indicated by e, he must not interpolate over a range greater than five-tenths of the difference between two consecutive tabular values. Thus, in entering Table I, instead of going to the right and adding a correction for seven-tenths of the tabular difference, one should go to the left and subtract the correction for three-tenths.

The values given in the tables have been checked by differencing, and in other ways.

THREE-PLACE LOGARITHMS

5. Table VIII. This is a three-place table of common logarithms, to be used where slide rule accuracy is sufficient. No interpolation is required or desirable. Both in entering and in leaving the table of logarithms, one chooses the nearest number obtainable without crossing a horizontal line. In simple computations it is often better, as with the slide rule, to omit characteristics altogether and to find the position of the decimal point in the answer by some other method. In extracting roots, however, it is of course very necessary to write the characteristic. This table may be used to find square roots so very readily that it seemed unnecessary to add a table of square roots. *E.g.*, by this table $\sqrt{90.4} = 9.51$ without interpolation. A table of square roots which would give this without interpolation would involve the tabulation of the roots of 9040 numbers.

TABLE I

	$\int_{-\infty}^{x} \phi(x)dx$ = Area under $\phi(x)$ from $-\infty$ to x.										Tenths of Mean Tabular Difference					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e†
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	4	8	12	16	20	0.6
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	4	8	11	15	19	1.1
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	4	8	11	15	19	0.9
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	4	8	11	15	19	0.7
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879	4	7	11	15	18	0.6
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	3	7	10	14	17	0.9
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	3	6	10	13	16	1.1
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852	3	6	9	12	15	0.9
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	3	6	8	11	14	1.0
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	3	5	8	10	13	1.2
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	2	5	7	9	12	1.3
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	2	4	6	8	10	1.1
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015	2	4	5	7	9	1.1
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	2	3	5	6	8	0.9
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	1	3	4	5	7	0.8
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441	1	2	4	5	6	0.7
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545	1	2	3	4	5	0.6
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633	1	2	3	3	4	1.0
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706	1	1	2	3	4	0.5
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767	1	1	2	2	3	0.7

	$\int_{-\infty}^{x} \phi(x)dx$ = Area under $\phi(x)$ from $-\infty$ to x.										Tenths of Mean Tabular Difference					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e†
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817	1	1	2	2	3	1.1
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857	0	1	1	2	2	0.5
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890	0	1	1	1	2	0.7
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916	0	1	1	1	1	0.8
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	0	0	1	1	1	0.1
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	0	0	0	1	1	0.7
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	0	0	0	0	1	0.8
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974	0	0	0	0	1	1.0
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981	0	0	0	0	1	1.0
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986	0	0	0	0	0	0.3
3.0	*9865	9869	9874	9878	9882	9886	9889	9893	9897	9900	0	1	1	2	2	0.6
3.1	9903	9906	9910	9913	9916	9918	9921	9924	9926	9929	0	1	1	1	2	1.1
3.2	9931	9934	9936	9938	9940	9942	9944	9946	9948	9950	0	0	1	1	1	0.6
3.3	9952	9953	9955	9957	9958	9960	9961	9962	9964	9965	0	0	0	1	1	0.6
3.4	9966	9968	9969	9970	9971	9972	9973	9974	9975	9976	0	0	0	0	1	0.8

	Special Values. Deciles and Quartiles										
$\int_{-\infty}^{x} \phi(x)dx$	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	x is negative
x	∓1.6449	∓1.2816	∓1.0364	∓.8416	∓.6745	∓.5244	∓.3853	∓.2533	∓.1257	.0000	
$\int_{-\infty}^{x} \phi(x)dx$	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50	x is positive

† For explanation of the symbol e see Introduction to these tables, page 378.

* For explanation of the symbol * see page 383.

TABLE I (*Continued*)

	$\int_{-\infty}^{x} \phi(x)dx$ = Area under $\phi(x)$ from $-\infty$ to x.										Tenths of Mean Tabular Difference					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e
3.5	**9767	9776	9784	9792	9800						1	2	2	3	4	0.6
					9800	9807	9815	9822	9828	9835	1	1	2	3	4	1.2
3.6	9841	9847	9853	9858	9864	9869	9874	9879	9883	9888	1	1	2	2	3	1.0
3.7	9892	9896	9900	9904	9908	9912	9915	9918	9922	9925	0	1	1	1	2	1.3
3.8	9928	9931	9933	9936	9939	9941	9943	9946	9948	9950	0	0	1	1	1	0.9
3.9	9952	9954	9956	9958	9959	9961	9963	9964	9966	9967	0	0	0	1	1	0.6
4.0	9968	9970	9971	9972	9973	9974	9976	9977	9978	9978	0	0	0	0	1	0.8
4.1	9979	9980	9981	9982	9983	9983	9984	9985	9985	9986	0	0	0	0	1	0.7
4.2	9987 ***	9987	9988	9988	9989	9989	9990	9990	9991	9991	0	0	0	0	0	0.4
4.3	9915	9918	9922	9925	9929	9932	9935	9938	9941	9943	0	1	1	1	2	0.5
4.4	9946	9948	9951	9953	9955	9957	9959	9961	9963	9964	0	0	1	1	1	0.0
4.5	9966	9968	9969	9971	9972	9973	9974	9976	9977	9978	0	0	0	1	1	0.7
4.6	9979	9980	9981	9982	9983	9983	9984	9985	9986	9986	0	0	0	0	1	0.6
4.7	9987 ****	9988	9988	9989	9989	9990	9990	9991	9991	9992	0	0	0	0	0	0.4
4.8	9921	9925	9928	9932	9935	9938	9941	9944	9947	9950	0	1	1	1	2	1.2
4.9	9952	9954	9957	9959	9961	9963	9965	9967	9968	9970	0	0	1	1	1	0.5

TABLE I — NORMAL AREAS

	$\int_{-\infty}^{x} \phi(x)dx$ = Area under $\phi(x)$ from $-\infty$ to x.										Tenths of Mean Tabular Difference					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e
5.0	9971	9973	9974	9975	9977	9978	9979	9980	9981	9982	0	0	0	0	1	0.5
5.1	9983	9984	9985	9986	9986	9987	9988	9988	9989	9989	0	0	0	0	0	0.4
	*****															
5.2	9900	9906	9911	9915	9920	9924					1	1	2	2	3	1.2
						9924	9928	9932	9935	9939	0	1	1	2	2	0.6
5.3	9942	9945	9948	9951	9954	9956	9958	9961	9963	9965	0	0	1	1	1	0.6
5.4	9967	9969	9970	9972	9973	9975	9976	9978	9979	9980	0	0	0	1	1	0.6
5.5	9981	9982	9983	9984	9985	9986	9987	9987	9988	9989	0	0	0	0	1	1.1
5.6	9989	9990	9991	9991	9992	9992	9992	9993	9993	9994	0	0	0	0	0	0.4
	******															
5.7	9940	9944	9947	9950	9953	9955	9958	9960	9963	9965	0	1	1	1	2	1.0
5.8	9967	9969	9971	9972	9974	9975	9977	9978	9979	9981	0	0	0	1	1	1.0
5.9	9982	9983	9984	9985	9986	9987	9987	9988	9989	9990	0	0	0	0	1	1.0

Note

* Prefix .9 to each entry
** " .99 " " "
*** " .999 " " "
**** " .9999 " " "
***** " .99999 " " "
****** " .999999 " " "

TABLE I (a)

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

x	0	1	2	3	4	5	6	7	8	9	Tenths of Mean Tabular Difference 1	2	3	4	5	$e\dagger$
0.0	.3989	.3989	.3989	.3988	.3986	.3984					0	0	0	0	1	1.3
0.0						.3984	.3982	.3980	.3977	.3973	0	1	1	1	2	1.3
0.1	.3970	.3965	.3961	.3956	.3951	.3945					1	1	2	2	3	1.6
0.1						.3945	.3939	.3932	.3925	.3918	1	1	2	3	3	0.7
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825	1	2	3	4	5	1.4
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697	1	3	4	5	7	1.4
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538	2	3	5	7	8	0.8
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352	2	4	6	8	9	1.1
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144	2	4	6	8	10	1.6
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920	2	4	7	9	11	0.6
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685	2	5	7	9	12	0.7
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444	2	5	7	10	12	0.2
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203	2	5	7	10	12	0.5
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965	2	5	7	9	12	0.5
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736	2	5	7	9	11	0.8
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518	2	4	6	9	11	0.8
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315	2	4	6	8	10	0.8
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127	2	4	5	7	9	0.5
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957	2	3	5	7	8	1.3
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804	2	3	5	6	8	0.6
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669	1	3	4	5	7	0.6
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551	1	2	3	5	6	0.7

TABLE I (c) — NORMAL ORDINATES

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

x	0	1	2	3	4	5	6	7	8	9	Tenths of Mean Tabular Difference 1	2	3	4	5	ϵ†
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449	1	2	3	4	5	0.6
2.1	.0440	.0431	.0422	.0413	.0404	.0395	.0387	.0379	.0371	.0363	1	2	3	3	4	0.8
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290	1	1	2	3	4	0.5
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229	1	1	2	2	3	0.6
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180	0	1	1	2	2	0.8
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139	0	1	1	2	2	0.4
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107	0	1	1	1	2	0.6
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081	0	0	1	1	1	0.8
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061	0	0	1	1	1	0.6
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046	0	0	0	1	1	0.6
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034	0	0	0	0	1	0.8
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025	0	0	0	0	0	0.5
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018	0	0	0	0	0	0.5
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013	0	0	0	0	0	0.5
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009	0	0	0	0	0	0.5
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006	0	0	0	0	0	0.5
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004	0	0	0	0	0	0.5
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003	0	0	0	0	0	0.5
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002	0	0	0	0	0	0.5
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001	0	0	0	0	0	0.5
4.0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	0	0	0	0	0	0.5
4.1	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	0	0	0	0	0	0.5
4.2	.0001	.0001	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	0	0	0	0	0	0.5

† For explanation of the symbol ϵ see Introduction to these tables, page 378.

TABLE II

$\phi^{(2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (x^2 - 1) = (x^2 - 1)\phi(x)$

x	0	1	2	3	4	5	6	7	8	9	*Tenths of Mean Tabular Difference* 1	2	3	4	5	e
0.0	−.3989	−.3989	−.3987	−.3984	−.3980	−.3975					0	1	1	1	2	2.1
0.0						−.3975	−.3968	−.3960	−.3951	−.3941	1	2	3	4	4	1.6
0.1	−.3930	−.3917	−.3904	−.3889	−.3873	−.3856	−.3838	−.3819	−.3798	−.3777	2	3	5	7	8	3.3
0.2	−.3754	−.3730	−.3706	−.3680	−.3653	−.3625	−.3596	−.3566	−.3535	−.3504	3	6	8	11	14	2.0
0.3	−.3471	−.3437	−.3402	−.3367	−.3330	−.3293	−.3255	−.3216	−.3176	−.3135	4	7	11	15	19	2.3
0.4	−.3094	−.3051	−.3008	−.2965	−.2920	−.2875	−.2830	−.2783	−.2736	−.2689	4	9	13	17	21	3.5
0.5	−.2641	−.2592	−.2543	−.2493	−.2443	−.2392	−.2341	−.2289	−.2238	−.2185	5	10	15	20	25	3.1
0.6	−.2133	−.2080	−.2027	−.1973	−.1919	−.1865	−.1811	−.1757	−.1702	−.1647	5	11	16	22	27	0.9
0.7	−.1593	−.1538	−.1483	−.1428	−.1373	−.1318	−.1262	−.1207	−.1153	−.1098	5	11	16	22	27	0.8
0.8	−.1043	−.0988	−.0934	−.0880	−.0825	−.0771	−.0718	−.0664	−.0611	−.0558	5	11	16	22	27	1.1
0.9	−.0506	−.0453	−.0401	−.0350	−.0299	−.0248	−.0197	−.0147	−.0098	−.0049	5	10	15	20	25	1.6
1.0	+.0000	+.0048	+.0096	+.0143	+.0190	+.0236	+.0281	+.0326	+.0371	+.0414	5	9	14	18	23	1.6
1.1	+.0458	+.0500	+.0542	+.0583	+.0624	+.0664	+.0704	+.0742	+.0780	+.0818	4	8	12	16	20	2.0
1.2	+.0854	+.0890	+.0926	+.0960	+.0994	+.1027	+.1060	+.1092	+.1123	+.1153	3	7	10	13	17	2.5
1.3	+.1183	+.1211	+.1239	+.1267	+.1293	+.1319	+.1344	+.1369	+.1392	+.1415	3	5	8	10	13	2.3
1.4	+.1437	+.1459	+.1480	+.1500	+.1519	+.1537	+.1555	+.1572	+.1588	+.1604	2	4	6	7	9	2.1
1.5	+.1619	+.1633	+.1647	+.1660	+.1672	+.1683	+.1694	+.1704	+.1714	+.1722	1	2	3	5	6	2.3
1.6	+.1730	+.1738	+.1745	+.1751	+.1757	+.1762	+.1766	+.1770	+.1773	+.1776	1	1	2	2	3	2.5
1.7	+.1778	+.1779	+.1780	+.1780	+.1780	+.1780	+.1778	+.1777	+.1774	+.1772	0	0	1	1	1	0.9
1.8	+.1769	+.1765	+.1761	+.1756	+.1751	+.1746	+.1740	+.1734	+.1727	+.1720	1	1	2	2	3	1.4
1.9	+.1713	+.1705	+.1697	+.1688	+.1679	+.1670	+.1661	+.1651	+.1641	+.1630	1	2	3	4	5	1.3
2.0	+.1620	+.1609	+.1598	+.1586	+.1575	+.1563	+.1550	+.1538	+.1526	+.1513	1	2	4	5	6	0.7
2.1	+.1500	+.1487	+.1474	+.1460	+.1446	+.1433	+.1419	+.1405	+.1391	+.1377	1	3	4	5	7	0.8
2.2	+.1362	+.1348	+.1333	+.1319	+.1304	+.1289	+.1275	+.1260	+.1245	+.1230	1	3	4	6	7	0.6
2.3	+.1215	+.1200	+.1185	+.1170	+.1155	+.1141	+.1126	+.1111	+.1096	+.1081	1	3	4	6	7	0.7
2.4	+.1066	+.1051	+.1036	+.1022	+.1007	+.0992	+.0978	+.0963	+.0949	+.0935	1	3	4	6	7	0.7

TABLE II — $\Phi^{(2)}$

$$\phi^{(2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (x^2 - 1) = (x^2 - 1)\phi(x)$$

											Tenths of Mean Tabular Difference					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e
2.5	+.0920	+.0906	+.0892	+.0878	+.0864	+.0850	+.0836	+.0823	+.0809	+.0796	1	3	4	6	7	0.1
2.6	+.0782	+.0769	+.0756	+.0743	+.0730	+.0717	+.0705	+.0692	+.0680	+.0668	1	3	4	5	6	0.8
2.7	+.0656	+.0644	+.0632	+.0620	+.0608	+.0597	+.0585	+.0574	+.0563	+.0552	1	2	3	5	6	0.8
2.8	+.0541	+.0531	+.0520	+.0510	+.0500	+.0490	+.0480	+.0470	+.0460	+.0451	1	2	3	4	5	0.7
2.9	+.0441	+.0432	+.0423	+.0414	+.0405	+.0396	+.0388	+.0379	+.0371	+.0363	1	2	3	4	4	0.6
3.0	+.0355	+.0347	+.0339	+.0331	+.0324	+.0316	+.0309	+.0302	+.0295	+.0288	1	1	2	3	4	0.6
3.1	+.0281	+.0275	+.0268	+.0262	+.0256	+.0249	+.0243	+.0237	+.0232	+.0226	1	1	2	2	3	0.1
3.2	+.0220	+.0215	+.0210	+.0204	+.0199	+.0194	+.0189	+.0184	+.0180	+.0175	1	1	2	2	3	0.6
3.3	+.0170	+.0166	+.0162	+.0157	+.0153	+.0149	+.0145	+.0141	+.0138	+.0134	0	1	1	2	2	0.7
3.4	+.0130	+.0127	+.0123	+.0120	+.0116	+.0113	+.0110	+.0107	+.0104	+.0101	0	1	1	1	2	1.0
3.5	+.0098	+.0095	+.0093	+.0090	+.0087	+.0085	+.0082	+.0080	+.0078	+.0075	0	1	1	1	1	0.8
3.6	+.0073	+.0071	+.0069	+.0067	+.0065	+.0063	+.0061	+.0059	+.0057	+.0056	0	0	1	1	1	0.3
3.7	+.0054	+.0052	+.0051	+.0049	+.0048	+.0046	+.0045	+.0043	+.0042	+.0041	0	0	0	1	1	0.2
3.8	+.0039	+.0038	+.0037	+.0036	+.0034	+.0033	+.0032	+.0031	+.0030	+.0029	0	0	0	0	1	0.7
3.9	+.0028	+.0027	+.0026	+.0026	+.0025	+.0024	+.0023	+.0022	+.0022	+.0021	0	0	0	0	0	0.7
4.0	+.0020	+.0019	+.0019	+.0018	+.0018	+.0017	+.0016	+.0016	+.0015	+.0015	0	0	0	0	0	0.0
4.1	+.0014	+.0014	+.0013	+.0013	+.0012	+.0012	+.0011	+.0011	+.0011	+.0010	0	0	0	0	0	0.0
4.2	+.0010	+.0009	+.0009	+.0009	+.0009	+.0008	+.0008	+.0008	+.0007	+.0007	0	0	0	0	0	0.0
4.3	+.0007	+.0007	+.0006	+.0006	+.0006	+.0006	+.0005	+.0005	+.0005	+.0005	0	0	0	0	0	0.0
4.4	+.0005	+.0004	+.0004	+.0004	+.0004	+.0004	+.0004	+.0004	+.0003	+.0003	0	0	0	0	0	0.0
4.5	+.0003	+.0003	+.0003	+.0003	+.0003	+.0003	+.0002	+.0002	+.0002	+.0002	0	0	0	0	0	0.0
4.6	+.0002	+.0002	+.0002	+.0002	+.0002	+.0002	+.0002	+.0002	+.0002	+.0001	0	0	0	0	0	0.0
4.7	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	0	0	0	0	0	0.0
4.8	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	+.0001	0	0	0	0	0	0.0
4.9	+.0001	+.0001	+.0001	+.0001	+.0001	+.0000	+.0000	+.0000	+.0000	+.0000	0	0	0	0	0	0.0

TABLE III

	$\phi^{(3)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (-x^3 + 3x) = (-x^3 + 3x)\,\phi(x)$										*Tenths of Mean Tabular Differences*					
x	0	1	.2	3	4	5	6	7	8	9	1	2	3	4	5	e
0.0	.0000	.0120	.0239	.0359	.0478	.0597	.0716	.0834	.0952	.1070	12	24	36	47	59	1.2
0.1	.1187	.1303	.1419	.1534	.1648	.1762	.1874	.1986	.2097	.2206	11	23	34	45	57	2.3
0.2	.2315	.2422	.2529	.2634	.2737	.2840	.2941	.3040	.3138	.3235	10	21	31	41	51	4.0
0.3	.3330	.3423	.3515	.3605	.3693	.3779					9	18	27	36	45	2.4
						.3779	.3864	.3947	.4028	.4107	8	16	24	32	40	2.6
0.4	.4184	.4259	.4332	.4403	.4472	.4539					7	14	21	28	35	3.0
						.4539	.4603	.4666	.4727	.4785	6	12	18	24	30	2.2
0.5	.4841	.4895	.4947	.4996	.5043	.5088					5	10	15	20	25	2.5
						.5088	.5131	.5171	.5209	.5245	4	8	11	15	19	2.7
0.6	.5278	.5309	.5338	.5365	.5389	.5411					3	5	8	11	13	2.9
						.5411	.5431	.5448	.5463	.5476	2	3	5	6	8	3.0
0.7	.5486	.5495	.5501	.5504	.5506	.5505	.5502	.5497	.5490	.5481	see footnote *					
0.8	.5469	.5456	.5440	.5423	.5403	.5381					2	4	5	7	9	2.6
						.5381	.5358	.5332	.5305	.5276	3	5	8	11	14	2.6
0.9	.5245	.5212	.5177	.5140	.5102	.5062					4	7	11	15	18	2.3
						.5062	.5021	.4978	.4933	.4887	4	9	13	18	22	2.2
1.0	.4839	.4790	.4740	.4688	.4635	.4580	.4524	.4467	.4409	.4350	5	11	16	22	27	3.3
1.1	.4290	.4228	.4166	.4102	.4038	.3973	.3907	.3840	.3772	.3704	6	13	19	26	32	2.6
1.2	.3635	.3566	.3496	.3425	.3354	.3282	.3210	.3138	.3065	.2992	7	14	21	29	36	1.6
1.3	.2918	.2845	.2771	.2697	.2624	.2550	.2476	.2402	.2328	.2254	7	15	22	29	37	0.6
1.4	.2180	.2107	.2033	.1960	.1887	.1815	.1742	.1670	.1599	.1528	7	14	22	29	36	0.8
1.5	.1457	.1387	.1317	.1248	.1180	.1111	.1044	.0977	.0911	.0846	7	14	20	27	34	1.8
1.6	.0781	.0717	.0654	.0591	.0529	.0468	.0408	.0349	.0290	.0233	6	12	18	24	30	2.2
1.7	.0176	.0120	.0065	.0011	−.0042	−.0094	−.0146	−.0196	−.0245	−.0294	5	11	16	22	27	3.6
1.8	−.0341	−.0388	−.0433	−.0477	−.0521	−.0563	−.0605	−.0645	−.0685	−.0723	4	9	13	17	21	2.7
1.9	−.0761	−.0797	−.0832	−.0867	−.0900	−.0933	−.0964	−.0994	−.1024	−.1052	3	6	10	13	16	2.3

TABLE III — $\phi^{(3)}$

$$\phi^{(3)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (-x^3 + 3x) = (-x^3 + 3x)\,\phi(x)$$

x	0	1	2	3	4	5	6	7	8	9	Tenths of Mean Tabular Difference 1	2	3	4	5	e
2.0	−.1080	−.1106	−.1132	−.1157	−.1180	−.1203	−.1225	−.1245	−.1265	−.1284	2	5	7	9	11	2.3
2.1	−.1302	−.1320	−.1336	−.1351	−.1366	−.1380	−.1393	−.1405	−.1416	−.1426	1	3	4	6	7	2.1
2.2	−.1436	−.1445	−.1453	−.1460	−.1467	−.1473	−.1478	−.1483	−.1486	−.1490	1	1	2	2	3	2.3
2.3	−.1492	−.1494	−.1495	−.1496	−.1496	−.1495	−.1494	−.1492	−.1490	−.1487	see footnote *					
2.4	−.1483	−.1480	−.1475	−.1470	−.1465	−.1459	−.1453	−.1446	−.1439	−.1432	1	1	2	2	3	1.1
2.5	−.1424	−.1416	−.1408	−.1399	−.1389	−.1380	−.1370	−.1360	−.1350	−.1339	1	2	3	4	5	1.3
2.6	−.1328	−.1317	−.1305	−.1294	−.1282	−.1270	−.1258	−.1245	−.1233	−.1220	1	2	4	5	6	0.7
2.7	−.1207	−.1194	−.1181	−.1168	−.1154	−.1141	−.1127	−.1114	−.1100	−.1087	1	3	4	5	7	0.6
2.8	−.1073	−.1059	−.1045	−.1031	−.1017	−.1003	−.0990	−.0976	−.0962	−.0948	1	3	4	6	7	0.2
2.9	−.0934	−.0920	−.0906	−.0893	−.0879	−.0865	−.0852	−.0838	−.0825	−.0811	1	3	4	5	7	0.6
3.0	−.0798	−.0785	−.0771	−.0758	−.0745	−.0732	−.0720	−.0707	−.0694	−.0682	1	3	4	5	6	0.9
3.1	−.0669	−.0657	−.0645	−.0633	−.0621	−.0609	−.0598	−.0586	−.0575	−.0564	1	2	4	5	6	0.2
3.2	−.0552	−.0541	−.0531	−.0520	−.0509	−.0499	−.0488	−.0478	−.0468	−.0458	1	2	3	4	5	0.8
3.3	−.0449	−.0439	−.0429	−.0420	−.0411	−.0402	−.0393	−.0384	−.0376	−.0367	1	2	3	4	5	1.2
3.4	−.0359	−.0350	−.0342	−.0334	−.0327	−.0319	−.0311	−.0304	−.0297	−.0290	1	2	2	3	4	1.0
3.5	−.0283	−.0276	−.0269	−.0262	−.0256	−.0249	−.0243	−.0237	−.0231	−.0225	1	1	2	3	3	0.9
3.6	−.0219	−.0214	−.0208	−.0203	−.0198	−.0192	−.0187	−.0182	−.0177	−.0173	1	1	2	2	3	0.7
3.7	−.0168	−.0164	−.0159	−.0155	−.0150	−.0146	−.0142	−.0138	−.0134	−.0131	0	1	1	2	2	0.3
3.8	−.0127	−.0123	−.0120	−.0116	−.0113	−.0110	−.0107	−.0104	−.0100	−.0098	0	1	1	1	2	1.0
3.9	−.0095	−.0092	−.0089	−.0086	−.0084	−.0081	−.0079	−.0076	−.0074	−.0072	0	1	1	1	1	0.8
4.0	−.0070	−.0067	−.0065	−.0063	−.0061	−.0059	−.0058	−.0056	−.0054	−.0052	0	0	1	1	1	0.7
4.1	−.0051	−.0049	−.0047	−.0046	−.0044	−.0043	−.0042	−.0040	−.0039	−.0038	0	0	0	1	1	0.8
4.2	−.0036	−.0035	−.0034	−.0033	−.0032	−.0031	−.0030	−.0029	−.0028	−.0027	0	0	0	0	1	0.9
4.3	−.0026	−.0025	−.0024	−.0023	−.0022	−.0022	−.0021	−.0020	−.0019	−.0019	0	0	0	0	0	0.0
4.4	−.0018	−.0017	−.0017	−.0016	−.0016	−.0015	−.0014	−.0014	−.0013	−.0013	0	0	0	0	0	0.0
4.5	−.0012	−.0012	−.0012	−.0011	−.0011	−.0010	−.0010	−.0010	−.0009	−.0009	0	0	0	0	0	0.0
4.6	−.0009	−.0008	−.0008	−.0008	−.0007	−.0007	−.0007	−.0006	−.0006	−.0006	0	0	0	0	0	0.0
4.7	−.0006	−.0006	−.0005	−.0005	−.0005	−.0005	−.0005	−.0004	−.0004	−.0004	0	0	0	0	0	0.0
4.8	−.0004	−.0004	−.0004	−.0003	−.0003	−.0003	−.0003	−.0003	−.0003	−.0003	0	0	0	0	0	0.0
4.9	−.0003	−.0002	−.0002	−.0002	−.0002	−.0002	−.0002	−.0002	−.0002	−.0002	0	0	0	0	0	0.0

* The differences change sign at a point of this line. The usual method of interpolation should not be used.

TABLE IV — $\Phi^{(4)}$

$\phi^{(4)}(x) = (x^4 - 6x^2 + 3)\,\phi(x)$

x	$\phi^{(4)}$	x	$\phi^{(4)}$	x	$\phi^{(4)}$	x	$\phi^{(4)}$
0.0	1.1968	1.0	−0.4839	2.0	−0.2700	3.0	0.1330
.1	1.1671	.1	− .6091	.1	− .1765	.1	.1231
.2	1.0799	.2	− .6925	.2	− .0927	.2	.1107
.3	0.9413	.3	− .7341	.3	− .0214	.3	.0969
.4	0.7607	.4	− .7364	.4	+ .0362	.4	.0829
.5	0.5501	.5	− .7042	.5	.0800	.5	.0694
.6	0.3231	.6	− .6440	.6	.1105	.6	.0570
7	0.0937	.7	− .5632	.7	.1293	.7	.0460
.8	−0.1247	.8	− .4692	.8	.1379	.8	.0365
.9	−0.3203	.9	− .3693	.9	.1385	.9	.0284

(In this table there is a possible error of .0000 55)

TABLE V

Log n!

n	0	1	2	3	4
0	0.0000	0.0000	0.3010	0.7782	1.3802
1	6.5598	7.6012	8.6803	9.7943	10.9404
2	18.3861	19.7083	21.0508	22.4125	23.7927
3	32.4237	33.9150	35.4202	36.9387	38.4702
4	47.9116	49.5244	51.1477	52.7811	54.4246
5	64.4831	66.1906	67.9066	69.6309	71.3633
6	81.9202	83.7055	85.4979	87.2972	89.1034
7	100.0784	101.9297	103.7870	105.6503	107.5196
8	118.8547	120.7632	122.6770	124.5961	126.5204
9	138.1719	140.1310	142.0948	144.0632	146.0364
10	157.9700	159.9743	161.9829	163.9958	166.0128
11	178.2009	180.2462	182.2955	184.3485	186.4054
12	198.8254	200.9082	202.9945	205.0844	207.1779
13	219.8107	221.9280	224.0485	226.1724	228.2995
14	241.1291	243.2783	245.4306	247.5860	249.7443
15	262.7569	264.9359	267.1177	269.3024	271.4899
16	284.6735	286.8803	289.0898	291.3020	293.5168
17	306.8608	309.0938	311.3293	313.5674	315.8079
18	329.3030	331.5607	333.8207	336.0832	338.3480
19	351.9859	354.2669	356.5502	358.8358	361.1236
20	374.8969	377.2001	379.5054	381.8129	384.1226
21	398.0246	400.3489	402.6752	405.0036	407.3340
22	421.3587	423.7031	426.0494	428.3977	430.7480
23	444.8898	447.2534	449.6189	451.9862	454.3555
24	468.6094	470.9914	473.3752	475.7608	478.1482
25	492.5096	494.9093	497.3107	499.7138	502.1186

TABLE V — LOG $n!$

n	5	6	7	8	9
0	2.0792	2.8573	3.7024	4.6055	5.5598
1	12.1165	13.3206	14.5511	15.8063	17.0851
2	25.1906	26.6056	28.0370	29.4841	30.9465
3	40.0142	41.5705	43.1387	44.7185	46.3096
4	56.0778	57.7406	59.4127	61.0939	62.7841
5	73.1037	74.8519	76.6077	78.3712	80.1420
6	90.9163	92.7359	94.5619	96.3945	98.2333
7	109.3946	111.2754	113.1619	115.0540	116.9516
8	128.4498	130.3843	132.3238	134.2683	136.2177
9	148.0141	149.9964	151.9831	153.9744	155.9700
10	168.0340	170.0593	172.0887	174.1221	176.1595
11	188.4661	190.5306	192.5988	194.6707	196.7462
12	209.2748	211.3751	213.4790	215.5862	217.6967
13	230.4298	232.5634	234.7001	236.8400	238.9830
14	251.9057	254.0700	256.2374	258.4076	260.5808
15	273.6803	275.8734	278.0693	280.2679	282.4693
16	295.7343	297.9544	300.1771	302.4024	304.6303
17	318.0509	320.2965	322.5444	324.7948	327.0477
18	340.6152	342.8847	345.1565	347.4307	349.7071
19	363.4136	365.7059	368.0003	370.2970	372.5959
20	386.4343	388.7482	391.0642	393.3822	395.7024
21	409.6664	412.0009	414.3373	416.6758	419.0162
22	433.1002	435.4543	437.8103	440.1682	442.5281
23	456.7265	459.0994	461.4742	463.8508	466.2292
24	480.5374	482.9283	485.3210	487.7154	490.1116
25	504.5252	506.9334	509.3433	511.7549	514.1682

TABLE V (*Continued*)

Log n!

n	0	1	2	3	4
26	516.5832	518.9999	521.4182	523.8381	526.2597
27	540.8236	543.2566	545.6912	548.1273	550.5651
28	565.2246	567.6733	570.1235	572.5753	575.0287
29	589.7804	592.2443	594.7097	597.1766	599.6449
30	614.4858	616.9644	619.4444	621.9258	624.4087
31	639.3357	641.8285	644.3226	646.8182	649.3151
32	664.3255	666.8320	669.3399	671.8491	674.3596
33	689.4509	691.9707	694.4918	697.0143	699.5380
34	714.7076	717.2404	719.7744	722.3097	724.8463
35	740.0920	742.6373	745.1838	747.7316	750.2806
36	765.6002	768.1577	770.7164	773.2764	775.8375
37	791.2290	793.7983	796.3689	798.9406	801.5135
38	816.9749	819.5559	822.1379	824.7211	827.3055
39	842.8351	845.4272	848.0205	850.6149	853.2104
40	868.8064	871.4096	874.0138	876.6191	879.2255
41	894.8862	897.5001	900.1150	902.7309	905.3479
42	921.0718	923.6961	926.3214	928.9478	931.5751
43	947.3607	949.9952	952.6307	955.2672	957.9047
44	973.7505	976.3949	979.0404	981.6868	984.3342
45	1000.2389	1002.8931	1005.5482	1008.2043	1010.8614
46	1026.8237	1029.4874	1032.1520	1034.8176	1037.4841
47	1053.5028	1056.1758	1058.8498	1061.5246	1064.2004
48	1080.2742	1082.9564	1085.6394	1088.3234	1091.0082
49	1107.1360	1109.8271	1112.5191	1115.2119	1117.9057
50	1134.0864	1136.7862	1139.4870	1142.1885	1144.8909

TABLE V — LOG $n!$

n	5	6	7	8	9
26	528.6830	531.1079	533.5344	535.9625	538.3922
27	553.0044	555.4453	557.8878	560.3318	562.7774
28	577.4835	579.9399	582.3977	584.8571	587.3180
29	602.1147	604.5860	607.0588	609.5330	612.0087
30	626.8930	629.3787	631.8659	634.3544	636.8444
31	651.8134	654.3131	656.8142	659.3166	661.8204
32	676.8715	679.3847	681.8993	684.4152	686.9324
33	702.0631	704.5894	707.1170	709.6460	712.1762
34	727.3841	729.9232	732.4635	735.0051	737.5479
35	752.8308	755.3823	757.9349	760.4888	763.0439
36	778.3997	780.9632	783.5279	786.0937	788.6608
37	804.0875	806.6627	809.2390	811.8165	814.3952
38	829.8909	832.4775	835.0652	837.6540	840.2440
39	855.8070	858.4047	861.0035	863.6034	866.2044
40	881.8329	884.4415	887.0510	889.6617	892.2734
41	907.9660	910.5850	913.2052	915.8264	918.4486
42	934.2035	936.8329	939.4633	942.0948	944.7272
43	960.5431	963.1826	965.8231	968.4646	971.1071
44	986.9825	989.6318	992.2822	994.9334	997.5857
45	1013.5194	1016.1783	1018.8383	1021.4991	1024.1609
46	1040.1516	1042.8200	1045.4893	1048.1595	1050.8307
47	1066.8771	1069.5547	1072.2332	1074.9127	1077.5930
48	1093.6940	1096.3806	1099.0681	1101.7565	1104.4458
49	1120.6003	1123.2958	1125.9921	1128.6893	1131.3874
50	1147.5942	1150.2984	1153.0034	1155.7093	1158.4160

TABLE VI

Table of x_R = Deviate of Rank R

(*Note:* R is the rank in an ordered series in which N is the total number of individuals; x_R is the abscissa of the normal curve $\phi(x)$ pertaining to the mean position of R. There is a possible error of 1 in the last place.)*

R \ N	1	2	3	4	5	6	7	8	9	10
1	.000	−.798	−1.091	−1.271	−1.400	−1.499	−1.579	−1.647	−1.704	−1.755
2		.798	.000	− .325	− .532	− .683	− .800	− .895	− .976	−1.045
3			1.091	.325	.000	− .211	− .368	− .491	− .592	− .677
4				1.271	.532	.211	.000	− .158	− .283	− .386
5					1.400	.683	.368	.158	.000	− .126
6						1.499	.800	.491	.283	.126
7							1.579	.895	.592	.386
8								1.647	.976	.677
9									1.704	1.045
10										1.755

TABLE VI — MEAN ABSCISSAE

R \ N	11	12	13	14	15	16	17	18	19	20
1	−1.800	−1.840	−1.876	−1.909	−1.940	−1.968	−1.994	−2.018	−2.041	−2.063
2	−1.105	−1.158	−1.206	−1.250	−1.289	−1.326	−1.360	−1.391	−1.420	−1.447
3	− .751	− .815	− .872	− .924	− .970	−1.013	−1.052	−1.088	−1.122	−1.153
4	− .474	− .550	− .616	− .676	− .729	− .778	− .822	− .863	− .901	− .936
5	− .230	− .319	− .396	− .464	− .525	− .580	− .630	− .675	− .717	− .756
6	.000	− .105	− .194	− .272	− .341	− .403	− .458	− .509	− .556	− .598
7	.230	.105	.000	− .090	− .168	− .237	− .300	− .356	− .407	− .454
8	.474	.319	.194	.090	.000	− .078	− .148	− .211	− .267	− .319
9	.751	.550	.396	.272	.168	.078	.000	− .070	− .132	− .189
10	1.105	.815	.616	.464	.341	.237	.148	.070	.000	− .063
11	1.800	1.158	.872	.676	.525	.403	.300	.211	.132	.063
12		1.840	1.206	.924	.729	.580	.458	.356	.267	.189
13			1.876	1.250	.970	.778	.630	.509	.407	.319
14				1.909	1.289	1.013	.822	.675	.556	.454
15					1.940	1.326	1.052	.863	.717	.598
16						1.968	1.360	1.088	.901	.756
17							1.994	1.391	1.122	.936
18								2.018	1.420	1.153
19									2.041	1.447
20										2.063

* The use of this table is illustrated in the text, pages 97 and 170.

TABLE VI (*Continued*)

Table of x_R = Deviate of Rank R

(*Note:* R is the rank in an ordered series in which N is the total number of individuals; x_R is the abscissa of the normal curve $\phi(x)$ pertaining to the mean position of R. There is a possible error of 1 in the last place.)

R \ N	21	22	23	24	25	26	27	28	29	30
1	−2.083	−2.102	−2.120	−2.138	−2.154	−2.170	−2.186	−2.199	−2.213	−2.227
2	−1.473	−1.497	−1.520	−1.541	−1.562	−1.582	−1.600	−1.619	−1.636	−1.653
3	−1.183	−1.210	−1.236	−1.261	−1.284	−1.306	−1.328	−1.348	−1.367	−1.386
4	− .969	−1.000	−1.029	−1.056	−1.082	−1.106	−1.130	−1.152	−1.173	−1.193
5	− .792	− .826	− .858	− .888	− .916	− .943	− .968	− .992	−1.015	−1.037
6	− .638	− .675	− .710	− .742	− .773	− .802	− .829	− .855	− .880	− .903
7	− .498	− .538	− .575	− .611	− .644	− .675	− .704	− .732	− .759	− .784
8	− .366	− .410	− .451	− .489	− .525	− .558	− .590	− .619	− .648	− .675
9	− .241	− .289	− .333	− .374	− .413	− .449	− .485	− .514	− .545	− .573
10	− .120	− .172	− .220	− .264	− .306	− .344	− .378	− .414	− .447	− .477
11	.000	− .057	− .109	− .157	− .202	− .243	− .282	− .319	− .353	− .385
12	.120	.057	.000	− .052	− .100	− .145	− .187	− .226	− .262	− .297
13	.241	.172	.109	.052	.000	− .048	− .093	− .135	− .174	− .210
14	.366	.289	.220	.157	.100	.048	.000	− .045	− .086	− .126
15	.498	.410	.333	.264	.202	.145	.093	.045	.000	− .042

TABLE VI — MEAN ABSCISSAE

R \ N	21	22	23	24	25	26	27	28	29	30
16	.638	.538	.451	.374	.306	.243	.187	.135	.086	.042
17	.792	.675	.575	.489	.413	.344	.282	.226	.174	.126
18	.969	.826	.710	.611	.525	.449	.378	.319	.262	.210
19	1.183	1.000	.858	.742	.644	.558	.485	.414	.353	.297
20	1.473	1.210	1.029	.888	.773	.675	.590	.514	.447	.385
21	2.083	1.497	1.236	1.056	.916	.802	.704	.619	.545	.477
22		2.102	1.520	1.261	1.082	.943	.829	.732	.648	.573
23			2.120	1.541	1.284	1.106	.968	.855	.759	.675
24				2.138	1.562	1.306	1.130	.902	.880	.784
25					2.154	1.582	1.328	1.152	1.015	.903
26						2.170	1.600	1.348	1.173	1.037
27							2.186	1.619	1.307	1.193
28								2.199	1.636	1.386
29									2.213	1.653
30										2.227

TABLE VI (*Continued*)

Table of x_R = Deviate of Rank R

(*Note:* R is the rank in an ordered series in which N is the total number of individuals; x_R is the abscissa of the normal curve $\phi(x)$ pertaining to the mean position of R. There is a possible error of 1 in the last place.)

R \ N	31	32	33	34	35	36	37	38	39	40
1	−2.240	−2.252	−2.264	−2.275	−2.287	−2.298	−2.308	−2.319	−2.328	−2.338
2	−1.668	−1.683	−1.698	−1.712	−1.726	−1.739	−1.751	−1.764	−1.776	−1.788
3	−1.403	−1.420	−1.437	−1.453	−1.468	−1.482	−1.497	−1.511	−1.524	−1.537
4	−1.213	−1.231	−1.249	−1.266	−1.283	−1.299	−1.314	−1.329	−1.344	−1.358
5	−1.058	−1.078	−1.098	−1.116	−1.134	−1.151	−1.167	−1.184	−1.199	−1.214
6	− .926	− .947	− .968	− .988	−1.007	−1.025	−1.043	−1.060	−1.077	−1.092
7	− .808	− .831	− .853	− .874	− .894	− .914	− .932	− .950	− .968	− .985
8	− .700	− .725	− .748	− .771	− .792	− .812	− .832	− .851	− .870	− .887
9	− .600	− .626	− .651	− .675	− .697	− .719	− .740	− .760	− .779	− .798
10	− .506	− .534	− .560	− .585	− .609	− .632	− .654	− .675	− .695	− .715
11	− .416	− .445	− .473	− .499	− .524	− .549	− .572	− .594	− .615	− .636
12	− .329	− .360	− .390	− .417	− .444	− .469	− .493	− .517	− .539	− .561
13	− .245	− .278	− .309	− .338	− .366	− .393	− .418	− .443	− .466	− .489
14	− .162	− .197	− .230	− .261	− .291	− .319	− .346	− .371	− .396	− .419
15	− .081	− .118	− .152	− .185	− .216	− .246	− .274	− .301	− .327	− .352
16	.000	− .039	− .076	− .111	− .144	− .175	− .205	− .233	− .260	− .286
17	.081	.039	.000	− .037	− .072	− .105	− .136	− .166	− .194	− .221
18	.162	.118	.076	.037	.000	− .035	− .068	− .099	− .129	− .157
19	.245	.197	.152	.111	.072	.035	.000	− .033	− .064	− .094
20	.329	.278	.230	.185	.144	.105	.068	.033	.000	− .031

TABLE VI — MEAN ABSCISSAE

R \ N	31	32	33	34	35	36	37	38	39	40
21	.416	.360	.309	.261	.216	.175	.136	.099	.064	.031
22	.506	.445	.390	.338	.291	.246	.205	.166	.129	.094
23	.600	.534	.473	.417	.366	.319	.274	.233	.194	.157
24	.700	.626	.560	.499	.444	.393	.346	.301	.260	.221
25	.808	.725	.651	.585	.524	.469	.418	.371	.327	.286
26	.926	.831	.748	.675	.609	.549	.493	.443	.396	.352
27	1.058	.947	.853	.771	.697	.632	.572	.517	.466	.419
28	1.213	1.078	.968	.874	.792	.719	.654	.594	.539	.489
29	1.403	1.231	1.098	.988	.894	.812	.740	.675	.615	.561
30	1.668	1.420	1.249	1.116	1.007	.914	.832	.760	.695	.636
31	2.240	1.683	1.437	1.266	1.134	1.025	.932	.851	.779	.715
32		2.252	1.698	1.453	1.283	1.151	1.043	.950	.870	.798
33			2.264	1.712	1.468	1.299	1.167	1.060	.968	.887
34				2.275	1.726	1.482	1.314	1.184	1.077	.985
35					2.287	1.739	1.497	1.329	1.199	1.092
36						2.298	1.751	1.511	1.344	1.214
37							2.308	1.764	1.524	1.358
38								2.319	1.776	1.537
39									2.328	1.788
40										2.338

TABLE VI (*Continued*)

Table of x_R = Deviate of Rank R

(*Note:* R is the rank in an ordered series in which N is the total number of individuals; x_R is the abscissa of the normal curve $\phi(x)$ pertaining to the mean position of R. There is a possible error of 1 in the last place.)

R \ N	41	42	43	44	45	46	47	48	49	50
1	−2.347	−2.356	−2.365	−2.374	−2.382	−2.390	−2.398	−2.405	−2.413	−2.421
2	−1.799	−1.810	−1.820	−1.830	−1.841	−1.851	−1.860	−1.870	−1.879	−1.888
3	−1.549	−1.561	−1.573	−1.585	−1.596	−1.607	−1.617	−1.627	−1.637	−1.647
4	−1.371	−1.384	−1.397	−1.410	−1.421	−1.433	−1.445	−1.455	−1.467	−1.477
5	−1.229	−1.242	−1.257	−1.269	−1.282	−1.294	−1.307	−1.319	−1.330	−1.342
6	−1.108	−1.122	−1.137	−1.151	−1.164	−1.178	−1.191	−1.203	−1.215	−1.227
7	−1.001	−1.017	−1.032	−1.047	−1.061	−1.075	−1.089	−1.101	−1.114	−1.127
8	− .905	− .921	− .937	− .953	− .968	− .982	− .996	−1.011	−1.023	−1.037
9	− .816	− .833	− .850	− .866	− .882	− .898	− .912	− .927	− .941	− .954
10	− .733	− .752	− .769	− .786	− .803	− .819	− .834	− .849	− .864	− .878
11	− .656	− .675	− .693	− .711	− .728	− .745	− .761	− .777	− .792	− .807
12	− .581	− .601	− .620	− .639	− .657	− .675	− .691	− .708	− .724	− .739
13	− .510	− .531	− .552	− .571	− .590	− .608	− .625	− .642	− .658	− .675
14	− .442	− .464	− .485	− .505	− .524	− .543	− .562	− .579	− .596	− .613
15	− .375	− .398	− .420	− .441	− .461	− .481	− .500	− .518	− .536	− .553
16	− .311	− .334	− .357	− .379	− .400	− .421	− .440	− .459	− .478	− .496
17	− .247	− .272	− .296	− .319	− .341	− .362	− .383	− .402	− .421	− .440
18	− .184	− .210	− .235	− .259	− .282	− .304	− .326	− .346	− .366	− .385
19	− .123	− .150	− .176	− .201	− .225	− .248	− .270	− .291	− .312	− .332
20	− .061	− .090	− .117	− .143	− .168	− .192	− .215	− .237	− .259	− .279

TABLE VI — MEAN ABSCISSAE

R \ N	41	42	43	44	45	46	47	48	49	50
21	.000	− .030	− .058	− .086	− .112	− .137	− .161	− .184	− .206	− .228
22	.061	.030	.000	− .028	− .056	− .082	− .107	− .131	− .154	− .176
23	.123	.090	.058	.028	.000	− .027	− .053	− .078	− .102	− .126
24	.184	.150	.117	.086	.056	.027	.000	− .026	− .051	− .075
25	.247	.210	.176	.143	.112	.082	.053	.026	.000	− .025
26	.311	.272	.235	.201	.168	.137	.107	.078	.051	.025
27	.375	.334	.296	.259	.225	.192	.161	.131	.102	.075
28	.442	.398	.357	.319	.282	.248	.215	.184	.154	.126
29	.510	.464	.420	.379	.341	.304	.270	.237	.206	.176
30	.581	.531	.485	.441	.400	.362	.326	.291	.259	.228
31	.656	.601	.552	.505	.461	.421	.383	.346	.312	.279
32	.733	.675	.620	.571	.524	.481	.440	.402	.366	.332
33	.816	.752	.693	.639	.590	.543	.500	.459	.421	.385
34	.905	.833	.769	.711	.657	.608	.562	.518	.478	.440
35	1.001	.921	.850	.786	.728	.675	.625	.579	.536	.496
36	1.108	1.017	.937	.866	.803	.745	.691	.642	.596	.553
37	1.229	1.122	1.032	.953	.882	.819	.761	.708	.658	.613
38	1.371	1.242	1.137	1.047	.968	.898	.834	.777	.724	.675
39	1.549	1.384	1.257	1.151	1.061	.982	.912	.849	.792	.739
40	1.799	1.561	1.397	1.269	1.164	1.075	.996	.927	.864	.807
41	2.347	1.810	1.573	1.410	1.282	1.178	1.089	1.011	.941	.878
42		2.356	1.820	1.585	1.421	1.294	1.191	1.101	1.023	.954
43			2.365	1.830	1.596	1.433	1.307	1.203	1.114	1.037
44				2.374	1.841	1.607	1.445	1.319	1.215	1.127
45					2.382	1.851	1.617	1.455	1.330	1.227
46						2.390	1.860	1.627	1.467	1.342
47							2.398	1.870	1.637	1.477
48								2.405	1.879	1.647
49									2.413	1.888
50										2.421

TABLE VII

	Log* R_x										Tenths of Mean Tabular Difference					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e
0.0	0.0981	0946	0912	0877	0843	0809	0775	0742	0708	0675	3	7	10	13	17	Here e is less than 1
0.1	0.0642	0609	0576	0543	0511	0478	0446	0414	0382	0350	3	6	10	13	16	
0.2	0.0318	0286	0255	0224	0192	0161	0130	0100	0069	0038	3	6	9	12	15	
0.3	0.0008	$\bar{1}$.9978	9948	9918	9888	9858	9828	9799	9770	9740	3	6	9	12	15	
0.4	$\bar{1}$.9711	9682	9653	9625	9596	9568	9539	9511	9483	9455	3	6	9	11	14	
0.5	$\bar{1}$.9427	9399	9371	9344	9316	9289	9262	9235	9208	9181	3	5	8	11	14	
0.6	$\bar{1}$.9154	9127	9101	9074	9048	9022	8996	8970	8944	8918	3	5	8	10	13	
0.7	$\bar{1}$.8892	8867	8841	8816	8791	8765	8740	8715	8690	8666	3	5	8	10	13	
0.8	$\bar{1}$.8641	8616	8592	8568	8543	8519	8495	8471	8447	8423	2	5	7	10	12	
0.9	$\bar{1}$.8399	8376	8352	8329	8305	8282	8259	8236	8213	8190	2	5	7	9	12	
1.0	$\bar{1}$.8167	8144	8121	8099	8076	8054	8032	8009	7987	7965	2	4	7	9	11	
1.1	$\bar{1}$.7943	7921	7899	7878	7856	7834	7813	7791	7770	7749	2	4	6	9	11	
1.2	$\bar{1}$.7727	7706	7685	7664	7643	7623	7602	7581	7560	7540	2	4	6	8	10	
1.3	$\bar{1}$.7519	7499	7479	7458	7438	7418	7398	7378	7358	7339	2	4	6	8	10	
1.4	$\bar{1}$.7319	7299	7279	7260	7240	7221	7202	7182	7163	7144	2	4	6	8	10	
1.5	$\bar{1}$.7125	7106	7087	7068	7049	7031	7012	6993	6975	6956	2	4	6	7	9	
1.6	$\bar{1}$.6938	6919	6901	6883	6864	6846	6828	6810	6792	6774	2	4	5	7	9	
1.7	$\bar{1}$.6756	6739	6721	6703	6686	6668	6651	6633	6616	6598	2	3	5	7	9	
1.8	$\bar{1}$.6581	6564	6547	6530	6512	6495	6479	6462	6445	6428	2	3	5	7	8	
1.9	$\bar{1}$.6411	6395	6378	6361	6345	6328	6312	6296	6279	6263	2	3	5	7	8	
2.0	$\bar{1}$.6247	6230	6214	6198	6182	6166	6150	6134	6119	6103	2	3	5	6	8	
2.1	$\bar{1}$.6087	6071	6056	6040	6025	6009	5993	5978	5963	5947	2	3	5	6	8	
2.2	$\bar{1}$.5932	5917	5902	5886	5871	5856	5841	5826	5811	5796	2	3	5	6	8	
2.3	$\bar{1}$.5782	5767	5752	5737	5723	5708	5693	5679	5664	5650	1	3	4	6	7	
2.4	$\bar{1}$.5635	5621	5607	5592	5578	5564	5550	5536	5521	5507	1	3	4	5	7	

$$* R_x = \frac{1}{\phi(x)} \int_x^\infty \phi(x)dx, \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

TABLE VII (*Continued*)

	Log R_x										*Tenths of Mean Tabular Difference*					
x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	e
2.5	$\bar{1}$.5493	5479	5465	5451	5438	5424	5410	5396	5382	5369	1	3	4	6	7	Here e is less than 1
2.6	$\bar{1}$.5355	5341	5328	5314	5301	5287	5274	5260	5247	5234	1	3	4	5	7	
2.7	$\bar{1}$.5220	5207	5194	5181	5168	5154	5141	5128	5115	5102	1	3	4	5	7	
2.8	$\bar{1}$.5089	5076	5064	5051	5038	5025	5012	5000	4987	4974	1	3	4	5	6	
2.9	$\bar{1}$.4962	4949	4937	4924	4912	4899	4887	4874	4862	4849	1	2	4	5	6	
3.0	$\bar{1}$.4837	4825	4813	4800	4788	4776	4764	4752	4740	4728	1	2	4	5	6	
3.1	$\bar{1}$.4716	4704	4692	4680	4668	4656	4644	4633	4621	4609	1	2	4	5	6	
3.2	$\bar{1}$.4597	4585	4574	4562	4551	4539	4528	4516	4504	4493	1	2	3	5	6	
3.3	$\bar{1}$.4481	4470	4459	4447	4436	4425	4413	4402	4391	4380	1	2	3	5	6	
3.4	$\bar{1}$.4368	4357	4346	4335	4324	4313	4302	4291	4280	4269	1	2	3	4	5	
3.5	$\bar{1}$.4258	4247	4236	4225	4215	4204	4193	4182	4172	4161	1	2	3	4	5	
3.6	$\bar{1}$.4150	4140	4129	4118	4108	4097	4086	4076	4065	4055	1	2	3	4	5	
3.7	$\bar{1}$.4045	4034	4024	4013	4003	3993	3982	3972	3962	3952	1	2	3	4	5	
3.8	$\bar{1}$.3941	3931	3921	3911	3901	3890	3880	3870	3860	3850	1	2	3	4	5	
3.9	$\bar{1}$.3840	3830	3820	3810	3800	3790	3781	3770	3761	3751	1	2	3	4	5	

x	00	10	20	30	40	50	60	70	80	90	1	2	3	4	5	e
4.	$\bar{1}$.3741	3644	3549	3456	3365	3275					9	18	28	37	46	2.0
4.						3275	3187	3101	3016	2933	8	17	25	34	42	2.0
5.	$\bar{1}$.2851	2772	2692	2615	2538	2463					8	15	23	31	38	3.0
5.						2463	2389	2316	2245	2175	7	15	22	29	37	2.1
6.	$\bar{1}$.2105	2037	1970	1903	1838	1774					7	13	20	26	33	2.0
6.						1774	1710	1647	1586	1525	6	12	19	25	31	1.3
7.	$\bar{1}$.1464	1405	1347	1289	1232	1175					6	12	17	23	29	0.7
7.						1175	1120	1065	1010	0957	5	11	16	22	27	0.9
8.	$\bar{1}$.0904	0851	0799	0748	0698	0648					5	10	15	20	25	1.5
8.						0648	0598	0549	0501	0453	5	10	15	19	24	1.1
9.	$\bar{1}$.0406	0359	0312	0266	0221	0176					5	9	14	18	23	0.9
9.						0176	0131	0087	0044	0000	4	9	13	18	22	0.5

TABLE VIII

Moving the decimal point m places to the right (or left) in N is equivalent to adding m (or $-m$) to log N.

Do not interpolate in this Table. Choose the nearest number obtainable without crossing a horizontal line. (See also page 379.)

N	Log N	N	Log N	N	Log N	N	Log N	N	Log N	N	Log N
1.00	.000		.052		.103		.151		.198		.246
	.001	1.13	.053	1.27	.104	1.42	.152	1.58	.199	1.76	.246
	.002		.054		.105		.153		.200		.247
	.003		.055		.106		.154	1.59	.201	1.77	.248
1.01	.004		.056	1.28	.107	1.43	.155		.202		.249
	.005	1.14	.057		.108		.156		.203	1.78	.250
	.006		.058		.109		.157	1.60	.204		.251
	.007		.059	1.29	.110	1.44	.158		.205		.252
	.008		.060		.111		.159		.206	1.79	.253
1.02	.009	1.15	.061		.112		.160	1.61	.207		.254
	.010		.062		.113	1.45	.161		.208	1.80	.255
	.011		.063	1.30	.114		.162		.209		.256
	.012	1.16	.064		.115		.163	1.62	.210		.257
1.03	.013		.065		.116	1.46	.164		.211	1.81	.258
	.014		.066	1.31	.117		.165	1.63	.212		.259
	.015		.067		.118		.166		.213	1.82	.260
	.016	1.17	.068		.119	1.47	.167		.214		.261
1.04	.017		.069	1.32	.120		.168	1.64	.215	1.83	.262
	.018		.070		.121		.169		.216		.263
	.019		.071		.122	1.48	.170	1.65	.217		.264
	.020	1.18	.072		.123		.171		.218	1.84	.265
1.05	.021		.073	1.33	.124		.172		.219		.266
	.022		.074		.125	1.49	.173	1.66	.220	1.85	.267
	.023		.075		.126		.174		.221		.268
	.024	1.19	.076	1.34	.127		.175		.222		.269
1.06	.025		.077		.128	1.50	.176	1.67	.223	1.86	.270
	.026		.078		.129		.177		.224		.271
	.027	1.20	.079	1.35	.130		.178	1.68	.225	1.87	.272
	.028		.080		.131	1.51	.179		.226		.273
1.07	.029		.081		.132		.180		.227	1.88	.274
	.030	1.21	.082	1.36	.133		.181	1.69	.228		.275
	.031		.083		.134	1.52	.182		.229	1.89	.276
	.032		.084		.135		.183	1.70	.230		.277
1.08	.033		.085	1.37	.136		.184		.231		.278
	.034	1.22	.086		.137	1.53	.185		.232	1.90	.279
	.035		.087		.138		.186	1.71	.233		.280
	.036		.088	1.38	.139		.187		.234	1.91	.281
1.09	.037		.089		.140	1.54	.188		.235		.282
	.038	1.23	.090		.141		.189	1.72	.236	1.92	.283
	.039		.091	1.39	.142	1.55	.190		.237		.284
	.040		.092		.143		.191	1.73	.238		.285
1.10	.041	1.24	.093		.144		.192		.239	1.93	.286
	.042		.094	1.40	.145	1.56	.193		.240		.287
	.043		.095		.146		.194	1.74	.241	1.94	.288
	.044		.096		.147		.195		.242		.289
1.11	.045	1.25	.097	1.41	.148	1.57	.196	1.75	.243	1.95	.290
	.046		.098		.149		.197		.244		.291
	.047		.099		.150						
	.048	1.26	.100								
1.12	.049		.101								
	.050		.102								
	.051										

N	Log N	N	Log N	N	Log N	N	Log N	N	Log N	N	Log N
1.96	.292	2.18	.338	2.44	.387		.434	3.06	.486		.641
	.293		.339		.388	2.72	.435	3.07	.487	3.48	.542
1.97	.294	2.19	.340	2.45	.389	2.73	.436		.488	3.49	.543
	.295		.341		.390		.437	3.08	.489	3.50	.544
	.296	2.20	.342	2.46	.391	2.74	.438	3.09	.490	3.51	.545
1.98	.297		.343		.392	2.75	.439	3.10	.491		.546
	.298	2.21	.344	2.47	.393		.440		.492	3.52	.547
1.99	.299		.345	2.48	.394	2.76	.441	3.11	.493	3.53	.548
	.300	2.22	.346		.395	2.77	.442	3.12	.494	3.54	.549
2.00	.301		.347	2.49	.396		.443		.495	3.55	.550
	.302	2.23	.348		.397	2.78	.444	3.13	.496	3.56	.551
2.01	.303		.349	2.50	.398		.445	3.14	.497		.552
	.304	2.24	.350		.399	2.79	.446	3.15	.498	3.57	.553
2.02	.305		.351	2.51	.400	2.80	.447		.499	3.58	.554
	.306	2.25	.352	2.52	.401		.448	3.16	.500	3.59	.555
2.03	.307		.353		.402	2.81	.449	3.17	.501	3.60	.556
	.308	2.26	.354	2.53	.403	2.82	.450	3.18	.502		.557
	.309		.355		.404		.451		.503	3.61	.558
2.04	.310	2.27	.356	2.54	.405	2.83	.452	3.19	.504	3.62	.559
	.311		.357		.406	2.84	.453	3.20	.505	3.63	.560
2.05	.312	2.28	.358	2.55	.407		.454		.506	3.64	.561
	.313		.359	2.56	.408	2.85	.455	3.21	.507	3.65	.562
2.06	.314	2.29	.360		.409	2.86	.456	3.22	.508	3.66	.563
	.315		.361	2.57	.410		.457	3.23	.509		.564
	.316	2.30	.362		.411	2.87	.458		.510	3.67	.565
2.07	.317		.363	2.58	.412	2.88	.459	3.24	.511	3.68	.566
	.318	2.31	.364	2.59	.413		.460	3.25	.512	3.69	.567
2.08	.319	2.32	.365		.414	2.89	.461	3.26	.513	3.70	.568
	.320		.366	2.60	.415	2.90	.462		.514	3.71	.569
2.09	.321	2.33	.367		.416		.463	3.27	.515		.570
	.322		.368	2.61	.417	2.91	.464	3.28	.516	3.72	.571
2.10	.323	2.34	.369	2.62	.418	2.92	.465	3.29	.517	3.73	.572
	.324		.370		.419		.466		.518	3.74	.573
2.11	.325	2.35	.371	2.63	.420	2.93	.467	3.30	.519	3.75	.574
	.326		.372		.421	2.94	.468	3.31	.520	3.76	.575
2.12	.327	2.36	.373	2.64	.422		.469	3.32	.521	3.77	.576
	.328		.374		.423	2.95	.470	3.33	.522	3.78	.577
2.13	.329	2.37	.375	2.65	.424	2.96	.471		.523		.578
	.330		.376		.425		.472	3.34	.524	3.79	.579
2.14	.331	2.38	.377	2.66	.426	2.97	.473	3.35	.525	3.80	.580
	.332	2.39	.378		.427	2.98	.474	3.36	.526	3.81	.581
2.15	.333		.379	2.67	.428		.475	3.37	.527	3.82	.582
	.334	2.40	.380	2.68	.429	2.99	.476	3.38	.528	3.83	.583
2.16	.335		.381		.430	3.00	.477	3.39	.529	3.84	.584
	.336	2.41	.382	2.69	.431		.478	3.40	.530	3.85	.585
2.17	.337		.383	2.70	.432	3.01	.479		.531		.586
		2.42	.384		.433	3.02	.480	3.41	.532	3.86	.587
			.385			3.03	.481	3.42	.533	3.87	.588
		2.43	.386	2.71	.433		.482	3.43	.534	3.88	.589
									.535	3.89	.590
						3.04	.483	3.44	.536	3.90	.591
						3.05	.484	3.45	.537	3.91	.592
							.485	3.46	.538	3.92	.593
								3.47	.539	3.96	.594
									.540	3.94	.595
											.596
										3.95	.597
										3.96	.598
										3.97	.599

TABLE VIII (*Continued*)

Moving the decimal point m places to the right (or left) in N is equivalent to adding m (or $-m$) to $\log N$.

Do not interpolate in this Table. Choose the nearest number obtainable without crossing a horizontal line. (See also page 379.)

N	Log N	N	Log N	N	Log N	N	Log N	N	Log N	N	Log N
3.98	.600	4.56	.659	5.10		5.63		6.14	.788	6.60	
3.99	.601	4.57	.660	5.11	.708	5.64	.751	6.15	.789	6.61	.820
4.00	.602	4.58	.661	5.12	.709	5.65	.752			6.62	.821
4.01	.603	4.59	.662	5.13	.710	5.66	.753	6.16			
4.02	.604	4.60	.663	5.14	.711			6.17	.790	6.63	
4.03	.605	4.61	.664	5.15	.712	5.67		6.18	.791	6.64	.822
4.04	.606	4.62	.665	5.16	.713	5.68	.754	6.19	.792	6.65	.823
4.05	.607	4.63	.666	5.17		5.69	.755	6.20		6.66	
	.608					5.70	.756				
4.06	.609	4.64		5.18	.714	5.71	.757	6.21	.793	6.67	.824
4.07	.610	4.65	.667	5.19	.715	5.72		6.22	.794	6.68	.825
4.08	.611	4.66	.668	5.20	.716			6.23		6.69	
4.09	.612	4.67	.669	5.21	.717	5.73	.758				
4.10	.613	4.68	.670	5.22	.718	5.74	.759			6.70	.826
4.11	.614	4.69	.671			5.75	.760	6.24	.795	6.71	.827
4.12	.615	4.70	.672	5.23		5.76		6.25	.796	6.72	
4.13	.616	4.71	.673	5.24	.719						
4.14	.617	4.72	.674	5.25	.720	5.77	.761	6.26		6.73	.828
4.15	.618	4.73	.675	5.26	.721	5.78	.762	6.27	.797		
4.16	.619	4.74	.676	5.27	.722	5.79	.763	6.28	.798	6.74	
4.17	.620	4.75	.677			5.80				6.75	.829
4.18	.621	4.76	.678	5.28	.723			6.29		6.76	.830
4.19	.622			5.29		5.81	.764	6.30	.799		
4.20	.623	4.77				5.82	.765	6.31	.800	6.77	
4.21	.624	4.78	.679	5.30	.724	5.83	.766	6.32	.801	6.78	.831
4.22	.625	4.79	.680	5.31	.725	5.84		6.33		6.79	.832
4.23	.626	4.80	.681	5.32	.726						
4.24	.627	4.81	.682	5.33	.727	5.85	.767	6.34	.802	6.80	
4.25	.628	4.82	.683			5.86	.768	6.35	.803	6.81	.833
4.26	.629	4.83	.684	5.34		5.87	.769	6.36		6.82	.834
4.27	.630	4.84	.685	5.35	.728	5.88				6.83	
4.28	.631	4.85	.686	5.36	.729						
4.29	.632	4.86	.687	5.37	.730	5.89	.770	6.37	.804	6.84	.835
4.30	.633			5.38	.731	5.90	.771	6.38	.805	6.85	.836
4.31	.634	4.87								6.86	
4.32	.635	4.88	.688	5.39		5.91		6.39			
4.33	.636	4.89	.689	5.40	.732	5.92	.772	6.40	.806	6.87	.837
4.34	.637	4.90	.690	5.41	.733	5.93	.773	6.41	.807		
4.35	.638	4.91	.691	5.42	.734	5.94	.774			6.88	
4.36	.639	4.92	.692	5.43	.735					6.89	.838
4.37	.640	4.93	.693			5.95		6.42		6.90	.839
4.38	.641	4.94	.694	5.44		5.96	.775	6.43	.808	6.91	
4.39	.642	4.95	.695	5.45	.736	5.97	.776	6.44	.809		
4.40	.643	4.96		5.46	.737	5.98	.777			6.92	.840
4.41	.644	4.97	.696	5.47	.738	5.99		6.45		6.93	.841
4.42	.645	4.98	.697	5.48	.739			6.46	.810	6.94	
4.43	.646	4.99	.698			6.00	.778	6.47	.811		
4.44	.647	5.00	.699	5.49		6.01	.779			6.95	.842
4.45	.648	5.01	.700	5.50	.740			6.48			
4.46	.649	5.02	.701	5.51	.741	6.02		6.49	.812	6.96	
4.47	.650			5.52	.742	6.03	.780	6.50	.813	6.97	.843
4.48	.651	5.03		5.53	.743	6.04	.781			6.98	.844
4.49	.652	5.04	.702			6.05	.782	6.51		6.99	
4.50	.653			5.54		6.06		6.52	.814		
4.51	.654	5.05	.703	5.55	.744			6.53	.815	7.00	.845
4.52	.655	5.06	.704	5.56	.745	6.07	.783			7.01	.846
4.53	.656	5.07	.705	5.57	.746	6.08	.784	6.54		7.02	
4.54	.657	5.08	.706	5.58	.747			6.55	.816		
4.55	.658	5.09	.707	5.59		6.09		6.56	.817	7.03	.847
						6.10	.785				
				5.60	.748	6.11	.786	6.57		7.04	
				5.61	.749	6.12	.787	6.58	.818	7.05	.848
				5.62	.750	6.13		6.59	.819	7.06	.849
										7.07	

N	Log N	N	Log N	N	Log N	N	Log N	N	Log N	N	Log N
7.08	.850	7.58		8.07	.907	8.53	.931	9.01		9.50	
		7.59	.880	8.08		8.54		9.02	.955	9.51	.978
7.09		7.60	.881								
7.10	.851	7.61		8.09	.908	8.55	.932	9.03		9.52	
7.11	.852			8.10		8.56		9.04	.956	9.53	.979
7.12		7.62	.882								
				8.11	.909	8.57	.933	9.05		9.54	
7.13	.853	7.63				8.58		9.06	.957	9.55	.980
7.14	.854	7.64	.883	8.12						9.56	
7.15				8.13	.910	8.59	.934	9.07			
		7.65				8.60		9.08	.958	9.57	.981
7.16	.855	7.66	.884	8.14						9.58	
		7.67	.885	8.15	.911	8.61	.935	9.09			
7.17		7.68						9.10	.959	9.59	.982
7.18	.856			8.16		8.62				9.60	
7.19	.857	7.69	.886	8.17	.912	8.63	.936	9.11			
7.20		7.70						9.12	.960	9.61	
				8.18	.913	8.64		9.13		9.62	.983
7.21	.858	7.71	.887	8.19		8.65	.937				
								9.14	.961	9.63	
7.22		7.72		8.20	.914	8.66		9.15		9.64	.984
7.23	.859	7.73	.888	8.21		8.67	.938				
7.24	.860	7.74	.889					9.16	.962	9.65	
7.25		7.75		8.22	.915	8.68		9.17		9.66	.985
				8.23		8.69	.939			9.67	
7.26	.861	7.76	.890					9.18	.963		
		7.77		8.24	.916	8.70		9.19		9.68	.986
7.27				8.25		8.71	.940			9.69	
7.28	.862	7.78	.891					9.20	.964		
7.29	.863			8.26	.917	8.72		9.21		9.70	
7.30		7.79				8.73	.941			9.71	.987
		7.80	.892	8.27				9.22			
7.31	.864			8.28	.918	8.74		9.23	.965	9.72	
		7.81				8.75	.942			9.73	.988
7.32		7.82	.893	8.29				9.24			
7.33	.865	7.83	.894	8.30	.919	8.76		9.25	.966	9.74	
		7.84				8.77	.943			9.75	.989
7.34				8.31		8.78		9.26		9.76	
7.35	.866	7.85	.895	8.32	.920			9.27	.967		
7.36	.867	7.86				8.79	.944			9.77	.990
7.37				8.33		8.80		9.28		9.78	
		7.87	.896	8.34	.921			9.29	.968		
7.38	.868					8.81	.945	9.30		9.79	.991
		7.88		8.35		8.82				9.80	
7.39		7.89	.897	8.36	.922			9.31	.969		
7.40	.869					8.83	.946	9.32		9.81	
7.41	.870	7.90		8.37		8.84				9.82	.992
7.42		7.91	.898	8.38	.923			9.33	.970		
						8.85	.947	9.34		9.83	
7.43	.871	7.92		8.39	.924	8.86				9.84	.993
		7.93	.899	8.40				9.35	.971	9.85	
7.44		7.94	.900			8.87	.948	9.36			
7.45	.872	7.95		8.41	.925	8.88				9.86	.994
7.46	.873			8.42				9.37		9.87	
7.47		7.96	.901			8.89	.949	9.38	.972		
		7.97		8.43	.926	8.90				9.88	
7.48	.874			8.44				9.39		9.89	.995
7.49		7.98	.902			8.91	.950	9.40	.973		
				8.45	.927	8.92				9.90	
7.50	.875	7.99		8.46				9.41		9.91	.996
		8.00	.903			8.93	.951	9.42	.974		
7.51				8.47	.928	8.94				9.92	
7.52	.876	8.01		8.48				9.43		9.93	.997
7.53	.877	8.02	.904			8.95	.952	9.44	.975	9.94	
7.54				8.49	.929	8.96		9.45			
		8.03		8.50						9.95	.998
7.55	.878	8.04	.905			8.97	.953	9.46	.976	9.96	
		8.05	.906	8.51	.930	8.98		9.47			
7.56		8.06		8.52						9.97	
7.57	.879					8.99	.954	9.48	.977	9.98	.999
						9.00		9.49			
										9.99	
										10.00	1.000