INTRODUCTION

TO

MATHEMATICAL STATISTICS

BY

CARL J. WEST, Ph.D.
FORMERLY ASSISTANT PROFESSOR OF MATHEMATICS
OHIO STATE UNIVERSITY

REVISED EDITION

COLLEGE BOOK COMPANY
COLUMBUS, OHIO
1941

PREFACE

STATISTICS is among the very latest of the sciences to become mathematical as opposed to merely numerical and this almost abrupt change in methods has resulted in distinct instructional needs. By mathematical is meant the use of a small number of indexes which express the statistical information contained in the data in concise, accurate, and standardized language. In this way it is possible to replace the mass of data by a few derived measures.

Mathematical statistics has been developed from the older mathematical theory of probability. The developments of the past thirty years and especially the past ten years have resulted in a science which has grown away from its origin until it has become an almost separate branch of mathematical theory.

The development of mathematical statistics has been largely in the hands of persons of thorough mathematical training, although in a number of outstanding instances these research workers have possessed a highly developed sense of practical values.

Mathematical methods in statistics have proven so useful that literally hundreds of research workers are making wide use of them. Only a small percentage of the persons using modern statistical methods have had the mathematical training necessary to following the logical steps connecting present day methods with the basic theory of probability, or to following many of the proofs and derivations of the formulas which they frequently use. This situation is not of itself to be deplored for a comparable situation is found in other branches of science.

Though the practical worker in statistics need not of necessity be thoroughly familiar with the mathematical theory, he must understand the basic meaning of the methods and indexes which he is using. Fortunately the basic ideas which have proven of value here are not difficult to understand and do not of themselves involve higher mathematics. That is to say, the mathematics is only a tool applied to the basic concepts.

It is the primary purpose of this book to present some of the more important and more commonly used statistical indexes, and to show something of how the mathematics has been applied in the development of these indexes. Stress is laid at all points on the reader's obtaining a careful and accurate knowledge of the purpose of each index and especially of the reliance which can be placed upon each index and on each part of the method.

The student whose interest is primarily mathematical may, it is hoped, find the descriptions here presented of direct assistance toward his obtaining a full idea of the practical meaning of the formulas.

For the student who looks upon mathematics as something to be used only where necessary this book is intended to present the basic ideas of modern statistical methods in such a way that a lack of extensive mathematical training may not prevent the obtaining of a comprehensive idea of the significance of statistical indexes.

CARL J. WEST.

January, 1941

PREFACE TO THE FIRST EDITION

IT IS the aim of this book to present certain topics of elementary statistical theory which have been found useful and workable. There is no real reason why the theory of statistical methods should remain in obscurity. The necessary mathematics is largely elementary arithmetic and except in a few cases there is no need for higher mathematics. This book presupposes a reasonable familiarity with elementary mathematics only.

The idea is emphasized that a formula or method to be of practical and trustworthy value to a statistician must be so simple and direct that the final results can be interpreted in terms of the original conditions or the given data. To illustrate, if the arithmetic mean is ten percent larger in one distribution than in another what difference does this variation indicate in the forms of the distributions or in the values of the two series of measurements? If one correlation ratio is 0.54 and a second 0.59 how much more closely related are the attributes in the second than in the first? It must always be remembered that mathematics is but a tool to be used when the desired results can be more efficiently attained by its use, and that a formula is nothing more than a statement in mathematical language of a method of computation already thought out and understood. The difficulties that may arise in this subject are not primarily mathematical. They are essentially a part of the often difficult task of analyzing a statistical distribution.

* * * * *

Professor James McMahon has given most generously of his time and interest. Whatever assistance this book may afford to the practical worker in statistics is in a large measure due to the influence of Professor Walter F. Willcox, whose critical insight into the limitations and the possibilities of statistical methods together with the originality and practical initiative which permeate his research and instructional work place all his students under obligation to him.

CARL J. WEST.

CONTENTS

PAGE

CHAPTER V.

CHAPTER I

CURVE PLOTTING

Plotting the Data. Let us plot the following data of monthly average precipitation in inches at Columbus.

January	3.1	May	3.6	September	2.6
February	2.7	June	3.3	October	2.5
March	3.5	July	3.6	November	2.8
April	2.9	August	3.3	December	2.7

A horizontal straight line is first drawn and at equal distances on this line twelve points are located, one for each month. On a vertical line erected at the point corresponding to the month of January equal intervals are laid off, one for each inch of precipitation, and these intervals are subdivided into tenths. The two series of points are called the scales. It is usual to designate the horizontal and the vertical scale lines by O—X and O—Y respectively, as in Figure 1.

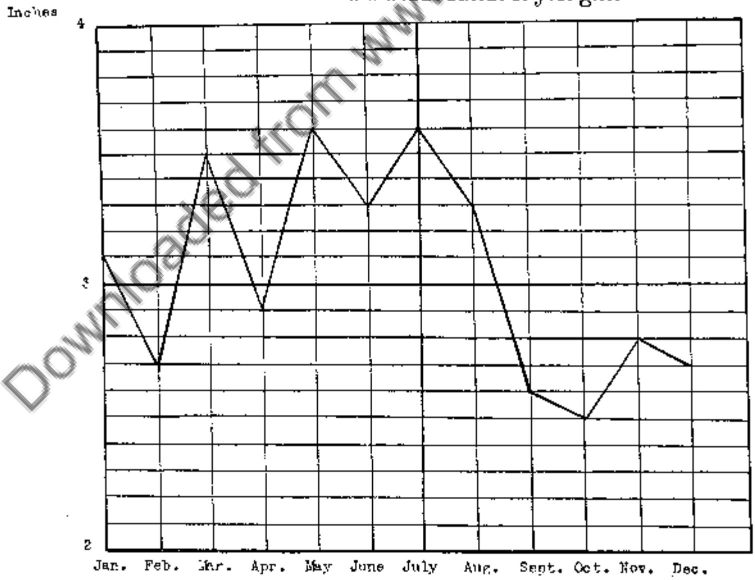


Figure 1. Monthly Average Precipitation at Columbus.

To save space the line O—X in Fig. 1 is taken at 2 inches instead of zero inches. The January precipitation is 3.1 inches. Place a dot above January, or beginning point, at a height corresponding to 3.1 inches on the vertical scale. The next point is directly above the second or February point at a distace corresponding to 2.7 inches. Continuing in this way a point is located for each month. The data is then said to be plotted or pictured point by point.

Exercises

1. Plot the following data of average monthly temperatures in degrees from 1878 to 1939 at Columbus.

Month	*Temperature*	*Month*	*Temperature*
January	29.7	July	75.2
February	30.8	August	73.0
March	40.0	September	67.0
April	51.0	October	55.0
May	62.1	November	42.2
June	70.9	December	32.6

2. Plot the life expectancy from the American Experience Table of Mortality by five year ages:

Age	*Exp.*	*Age*	*Exp.*	*Age*	*Exp.*
10	49	40	28	70	8
15	46	45	25	75	6
20	42	50	21	80	4
25	39	55	17	85	3
30	35	60	14	90	1
35	32	65	11		

3. Plot the following population data for the United States:

Year	*Population*	*Year*	*Population*	*Year*	*Population*
1790	3,929,214	1840	17,069,453	1890	62,947,714
1800	5,308,483	1850	23,191,876	1900	75,994,575
1810	7,239,881	1860	31,443,321	1910	91,972,266
1820	9,638,453	1870	38,558,371	1920	105,710,620
1830	12,866,020	1880	50,155,783	1930	122,775,046

4. Plot the following deaths per thousand from the American Experience *Table of Mortality* by five year ages:

Age	*Death*	*Age*	*Death*
10	7.5	45	11.2
15	7.6	50	13.8
20	7.8	55	18.6
25	8.1	60	26.7
30	8.4	65	40.1
35	8.9	70	62.0
40	9.8	75	94.4

Laying Off the Scales. The object of any graphic representation of statistical data is to present a vivid picture. Therefore a diagram too small or too large, or too wide or too narrow will not be as effective as will a correctly proportioned diagram. This means that the widths of the horizontal and the vertical scale intervals must be carefully chosen in order to give the complete diagram the proper proportions.

In determining the widths of the intervals account must be taken of the nature of the statistical material. If the data is of such a nature, for instance, that the measurements can be determined only to the nearest dollar it would be manifestly improper to divide the scale into intervals corresponding to cents. The wealth of the country and the value of manufactured articles are examples of statistics which do not admit of close subdivision.

It is useless to have the scale intervals finer than the smallest differences which the eye can conveniently distinguish on the diagram. This often means, even in the case of quite accurate material, that the data must be used in round numbers. In plotting population data for the United States, for instance, one million may be the smallest numerical difference that can be pictured on a diagram of ordinary size.

Horizontal and vertical lines called *coordinate lines* ordinarily are drawn to assist in carrying the divisions of the scales across the diagram.

Connecting the Points. The eye is aided in passing across a diagram if the plotted points are connected by a curve. The curve may be either a series of broken straight lines joining the points or a continuous curve passing through each point without sharp angles or abrupt changes in direction. Of the two methods the continuous curve is usually to be preferred because of the better appearance which it presents.

The general term, *curve,* it must be noted, refers to the connection of the points and not to any special geometrical shape. A straight line, a series of straight lines, are *statistical curves* just as much as a connection no part of which is straight.

The items measured, observed, or plotted may be referred to as *measurements, observed values, variates, frequencies, individual deviations, etc.,* depending on the connection or nature or source of the data.

Exercises

5. Connect the points in Exercises 1, 2, 3 and 4 by straight lines to obtain the curves.

6. Connect the points in Exercises 1, 2, 3 and 4 by curved lines passing through each point and rounded at the points to absorb the change of direction and compare with results obtained under Exercise 5.

7. Under Exercise 4 it will be noted that deaths per thousand beyond age 75 are omitted. These items are as follows:

Age	*Deaths*
80	144.5
85	235.6
90	454.5
95	1000.

Re-plot Exercise 4 using these added values. In doing this note that the scale must be changed in order to get all of the data on one curve, or else it may be better to plot the curve in two sections. If plotted in two sections it may be convenient to show only a part of the ordinates in the second section where the ordinates are so much greater than for the first section. If this latter is done some notation should be placed on the diagram to show that the scale is broken.

General Directions. Many rules can be laid down to cover the method of presenting statistical data in a diagram. Such rules are ordinarily framed with reference to the uses to be made of the diagram. Since it is the purpose of this book to describe methods of *analyzing* statistical data only such brief instructions as to form are here given as may be necessary in view of the purpose of this book.

All rules are, in fact, only details under the general rule—

a diagram must be so arranged as to present the data most effectively.

Regardless of the use to be made of it every diagram must be provided with a brief, concise, and yet accurate and comprehensive title. A careful study of the titles in any of the better known statistical reports will be especially helpful in acquiring a notion of what constitutes an adequate title. In particular, reference may be made to the titles in the year books of the various departments of the government at Washington, and the various scientific and technical journals.

All headings of columns must be clear and definite and all units of measurement of the scales must always be given, thus, "Precipitation in inches," "Temperature in degrees."

Exercises

In each of the following exercises construct a complete statistical diagram with the curve carefully drawn and an appropriate title designed for each.

8. The land area of the United States exclusive of outlying possessions for each census year from Reports of the United States Census.

9. The population of Ohio at each census year from Reports of the Census.

10. The accumulated value of $1 at 10% compound interest:

End of Year	1	2	3	4	5	6	7	8	9	10
Amount	$1.10	$1.21	$1.33	$1.46	$1.61	$1.77	$1.95	$2.14	$2.36	$2.59

11. The average yield per acre for wheat in the United States since 1900; Yearbook, Department of Agriculture.

12. Average farm price per bushel of wheat in the United States since 1900; Yearbook, Department of Agriculture.

13. Substitute the word corn for wheat in exercises 11 and 12 and construct the curves.

More than One Curve on the Same Diagram. The relationship among two or more sets of data can frequently be studied conveniently by plotting the different sets all on one diagram.

Exercises

14. Compare the rainfall curve with the temperature curve. To what extent do the two curves vary in the same directions? What conclusions can be drawn as to the tendency for the amount of rainfall to depend on the temperature?

15. Give a comparative interpretation of the curves of Exercises 11 and 12. Why should they not be expected to follow exactly the same general course?

16. Discuss, as in Exercise 15, curves of prices and yield per acre of corn.

Cumulative Curves. In the foregoing curves, values have been stated for each unit of the horizontal scale. It is frequently desirable to construct what is called a *cumulative curve* where the curve at any point is the *cumulative sum of all the preceding units.* Thus, if a curve be plotted to show the gain or loss in business month by month, it may also be desired to construct a curve showing gains or losses month by month since the beginning of the year. This latter curve would be a cumulative curve.

Exercises

17. Plot a cumulative curve of the data of monthly average precipitation.

18. Plot a cumulative curve of the data of Exercise 1.

19. Plot a cumulative curve of the data of Exercise 10.

20. What additional information is given by each of the curves in Exercises 17, 18, 19, over that given by the curves of Figure 1 and Exercises 1 and 10?

Coordinates. It is convenient to have a standardized notation for the horizontal and vertical scales. The horizontal base line is denoted by *O—X and called the axis of abscissas or simply the X-axis. The vertical line is denoted by O—Y and called the axis of ordinates or the Y-axis. The point where the two lines meet is the origin of coordinates.* Distances along the X-axis are spoken of as *x distances* or *x coordinates,* and those along the *Y-axis* as *y distances,* or *y coordinates.*

Logarithmic Curves. Where it is desired to picture relative changes or values, use is sometimes made of logarithmic curves. *A logarithmic curve is obtained by taking the logarithms of the measurements and using these logarithms as vertical distances or ordinates.* Since in multiplication logarithms are added, a constant ratio or rate will appear in the logarithmic diagram as a constant addition. Hence if there is a constant rate of change in the data the logarithmic curve will be a straight line. Whether the rate is constant or not, curves of this type are often of value for comparing different rates. However, if the rate of change is not approximately constant considerable familiarity with logarithms is necessary in interpreting the curve.

Exercises

21. Plot the curve of the following data, and then on the same diagram plot the logarithmic curve.

x	y	*log y*
1	1	.0
2	2	.30
3	4	.60
4	8	.90
5	16	1.20
6	32	1.51

22. Plot a logarithmic curve of the data of Exercise 10.

23. Plot a logarithmic curve of the data of Exercise 3.

24. Interpret the comparative significance of the curves in Exercises 10, 19, and 22.

CHAPTER II

SMOOTHING A CURVE

Interpolation. The curves of the preceding Chapter were drawn for the purpose of connecting the plotted points as an aid to the eye in following the course of the data across the diagram. However, a statistical curve can be used for other purposes than picturing the data.

The population of the United States is given by the Bureau of the Census for ten-year intervals. What has been the population from year to year? This question is essentially one of *interpolation, that is, of estimating values lying between stated or known values.*

A simple method of obtaining intermediate values from a curve consists of measuring on the vertical scale the height of the curve at the required point. Thus with the Population Curve of Exercise 3, Chapter I, which is constructed from decennial census reports, an estimate of the population for the year 1926 is given by the height of the curve above the 1926 point on the horizontal scale.

Exercises

1. Estimate the population of the United States from the curve of Exercise 3, Chapter I, for each of the years 1920-1930.

2. From Exercise 2 of Chapter I estimate the life expectancy at the ages 22, 46, 57.

3. From Exercise 4 of Chapter I estimate the deaths per thousand at the ages 22, 46, 57.

4. Appraise the accuracy of the interpolations in Exercises 2 and 3 at the younger ages as compared with the older.

5. Estimate the compound amount of $1 at 7½ years at 10% from the data of Exercise 10, Chapter I.

Under the foregoing method of interpolation an estimated value depends largely on the two consecutive given values which inclose it. But the increase in population during a decade may have occurred almost entirely during the latter years of the period so that the shape of the curve when drawn merely to connect the ten-year points may not adequately indicate this peculiarity of increase. Again the temperature for one month may have no connection with that of the preceding month and hence the curve between the points, depending as it does on the two non-related values can hardly be expected to give the actual temperature for an intermediate week or day.

It must be apparent therefore that a curve which passes through a series of more or less non-related points can not be of great value in interpolation and that the *problem of interpolation is essentially one of determining by some means or other the general course of the data and then estimating the intermediate values in conformity with this general trend.* The values obtained in this way are the most probable values. Accidental variations which bear no relation to the underlying tendencies can not be so estimated.

The Smoothing of a Curve. The curves of Chapter I, drawn as they are through each point, preserve all the variations whether they reflect an underlying trend in the data or whether they are due merely to the presence of accidental influences. The curve of Mean Monthly Temperatures, Exercise 1. of the preceding Chapter, shows distinct seasonal variations in temperature—higher temperatures in summer and lower in winter. Along with these essentially significant changes are fluctuations apparently accidental as, in one year June is warm and in another relatively cool; sometimes January is warmer than February and sometimes the reverse is true.

A curve to represent a general movement or trend must be without abrupt changes in direction and must sweep *among* the points rather than necessarily through each point. Since such a *smoothed curve,* as it is called, depends on the general or collective characteristics of the data the drawing of it must be based on collective characteristics of the measurements. One important general assumption has just been stated; namely, that the curve must ordinarily be smooth, that is, not have abrupt changes in direction. This assumption is another way of saying that the significant variations are fairly uniform from value to value and not capricious or arbitrary. A second assumption, which is presently discussed, is that areas under a curve are relatively stable and change in accordance with the underlying trend.

Smoothing by Inspection. The smoothing of a curve may be based on a study of the data and made a matter of the skill and experience of the statistician without the assistance of definitely stated methods or rules. The curve is then said to be *smoothed by inspection.*

In smoothing a curve the first step is to study the data carefully. Without such an investigation into the probable sources and extent of the irregularities and fluctuations one cannot hope to know what irregularities to smooth out and what to leave in. On the basis of the information gained by this study a preliminary curve should then be drawn freehand among the points. By successive erasures and re-drawings the finished curve can gradually be arrived at. Thus a curve showing the long time movements in the price of wheat will pass above some points and below others and how much the curve should miss any point can not well be determined without some knowledge of financial conditions, yields, etc.

The inspection method of smoothing a curve is often sufficiently accurate, especially when done by a statistician of experience and when there is a considerable element of inaccuracy in the data. Its disadvantage lies obviously in the fact that no two smoothings of the same curve will be exactly alike. The inspection method is essentially tentative and personal.

In any event, a rough preliminary draft of the curve should be made by inspection before proceeding to apply more refined methods.

Exercises

6. Smooth the curve of monthly average precipitation in Figure 1, Chapter I.

7. Smooth the temperature curve in Exercise 1, Chapter I.

8. Smooth the data in Exercise 11, Chapter I.

9. Note that Exercises 2 and 4 of Chapter I are already smoothed.

The Preservation of Areas. In the illustrative data at the beginning of Chapter I the precipitation of 3.5 inches in March is the total precipitation for the entire month. With a base of one unit, then a rectangle of height 3.5 will have an area equal to the total precipitation. Likewise the rectangle on the July unit as a base will have an area equal to 3.6 which is the July precipitation. The prices of Exercise 12, Chapter I, can in similar manner be represented by rectangles with heights equal to the respective prices and with unit bases. The population data of Exercise 3 of the same chapter may be represented by rectangles which are not adjacent, and have nine rectangles omitted between successive census years.

After the curve is smoothed each rectangle will be altered so as to have a curved top. The total area under the finished curve will then be the sum of the areas of the modified rectangles. *The First Rule of Preservation of Areas is that the curve should be so smoothed that the total area under the resulting curve is equal to the sum of the areas of the original rectangles.* Since, for instance, the monthly precipitation is made up of the sum of the daily precipitations it is reasonable to assume that the monthly sum is more stable than is the daily or weekly sum and hence we have the *Second Rule of the Preservation of Areas; namely, where possible, the areas of the individual rectangles are to remain unchanged.* This individual preservation of areas will result where, in smoothing, there is added to and subtracted from each rectangle an equal sum.

Within the requirement that the curve must be free from abrupt changes in direction the two preceding working rules furnish a fairly comprehensive basis for the smoothing of statistical data. In later chapters more detailed rules will be discussed. However, for much data the present rules are sufficient.

In the illustrative plotting, at the beginning of Chapter I, of the data of average monthly precipitation the vertical scale was laid off on a line through the January point. In constructing rectangles for smoothing, it would be convenient to have the January and other perpendiculars at the middle of the respective intervals. The zero point on the horizontal scale is then at the beginning of the first interval and the vertical distance for the first point is taken, not on the vertical scale line, but above the mid-point of the interval. Whenever the curve is to be smoothed the scale is marked off in this way, otherwise the scale is laid off as explained at the beginning of Chapter I.

Exercises

10. Applying the rectangle method of this Chapter smooth the illustrative data at the beginning of Chapter I.

11. By the rectangle method smooth the data of Exercises 11 and 12 of Chapter I.

The Adjusted Data; Interpolation. Since in general it is impossible to preserve exactly the area of each rectangle the process of smoothing will lead to values differing from the original data. The data then is said to be adjusted or graduated or smoothed by means of the curve. In accordance with the reasoning at the beginning of this chapter the adjusted values are to be taken as giving a more significant idea of the true trend of the data than does the original data. Accordingly, to obtain the best estimate of an intermediate point measure the corresponding ordinate of the smoothed curve, or measure the appropriate area under that curve. Thus the rainfall during the first week in June is obtained by measuring the area under the curve on the first one-fourth of the June base unit.

Test of a Graduation. The extent to which smoothing preserves the areas of the individual rectangles is often taken as a test of the appropriateness of the smoothing or graduation. The smoothed curve is said to fit the data and the term "goodness of fit" is used to denote the appropriateness of the methods used in the process of constructing the smooth curve. One measure of the goodness of fit is the extent to which the areas of the individual rectangles are preserved. In applying this test two columns of numbers are set down, in one the original values and in the other the adjusted values. The differences are then taken and studied. Other conditions being equal the smoothing with the smallest differences is the best.

Exercises

12. Discuss the goodness of fit of each of the curves smoothed in the preceding exercises.

Determining the General Trend of the Data. The characteristics of a movement of prices over a number of years can be determined from the smoothed curve. Thus a general upward trend of prices may be shown by a rise in the curve.

A general movement may be pictured by drawing a straight line, unless there are significant indications of a curved trend, or more than one straight line where there seems to be more than one distinct movement. This amounts to a straight line smoothing of the data. With data not conforming closely to a straight line there is likely to be some uncertainty in the exact location of the straight line or lines but since the lines are but the pictures of the ideas of general increases or decreases the uncertainty is neither greater nor less than is the uncertainty in the ideas of the general movements themselves.

Periodic Data. In smoothing and determining the general trend of data care must be taken that the data is not smoothed to conform to a straight line when there is an inherent periodicity in the material. The data of Exercise 12 of Chapter I exhibits significant tendencies for the values to be high for a few years and then consistently lower for a few years and then higher, and so on, through more or less regular and uniform cycles. In smoothing such data the idea should be to determine a uniform cycle and then smooth the data into the curve made up of the determined cycles. The problem of smoothing such data is complicated by the fact that the curve in addition to being composed of a series of similar loops or arches also has a tendency to rise or fall. Thus, imports of the U. S. have increased on the whole during the past 50 years though there have been increases and decreases following each other in fairly regular periods.

Occasionally it is possible to draw a periodic curve by an inspection method but it is usually necessary to rely upon some form of *sine curve* which requires considerable mathematical knowledge to construct.

Significance of Informal Methods. The informal inspection methods of plotting, smoothing, interpolating and graduating data referred to in this and the preceding chapter are, it may be repeated, adequate in many cases. Even if more elaborate methods are to be applied it is always well to make a preliminary use of inspection methods.

The preliminary inspections may give the statistician the information which his purposes require. And they may be of the greatest value in determining what more elaborate methods, if any, should be applied.

Makeham's Law of Mortality. It should be of interest to read at least the explanatory part of Appendix III on Makeham's Law of Mortality for an illustration of how a trend may be discovered from simple considerations.

Exercises

13. Plot the following data, smooth, and then examine it for periodicity and other general trends so as to give an answer to the question whether the month of October is now warmer or cooler that at other times of the past 62 years.

Mean Temperature for October, 1878-1939—Columbus, Ohio.

Year	*Temp.*	*Year*	*Temp.*	*Year*	*Temp.*	*Year*	*Temp.*
1878	54	1894	55	1910	58	1926	54
1879	62	1895	48	1911	54	1927	59
1880	52	1896	50	1912	56	1928	58
1881	60	1897	60	1913	54	1929	53
1882	59	1898	55	1914	58	1930	54
1883	56	1899	59	1915	57	1931	58
1884	59	1900	62	1916	55	1932	55
1885	51	1901	56	1917	48	1933	54
1886	54	1902	56	1918	58	1934	56
1887	51	1903	55	1919	60	1935	55
1888	49	1904	54	1920	60	1936	56
1889	49	1905	54	1921	54	1937	53
1890	54	1906	53	1922	57	1938	57
1891	53	1907	50	1923	52	1939	57
1892	54	1908	56	1924	59		
1893	55	1909	50	1925	47		

14. Plot the following data of maximum wind velocity in miles per hour for each day in October, 1939, at Columbus, Ohio, and smooth this data, interpreting the smoothed curve for trends.

	Maximum Velocity		*Maximum Velocity*		*Maximum Velocity*
Date	*Miles per Hour*	*Date*	*Miles per Hour*	*Date*	*Miles per Hour*
1	14	11	13	21	25
2	15	12	25	22	31
3	12	13	28	23	9
4	17	14	19	24	15
5	28	15	13	25	28
6	20	16	14	26	21
7	12	17	20	27	29
8	24	18	13	28	25
9	24	19	25	29	18
10	31	20	19	30	13
				31	16

CHAPTER III

FREQUENCY CURVES

Frequency. The following data of the heights of 750 freshmen students may be taken for purposes of illustrating the meaning of the term *frequency.*

The measurements are classified to show the number of individuals for each inch of height.

Measurements of Heights of Students

Height	Number	Height	Number
61	2	68	126
62	10	69	109
63	11	70	87
64	38	71	75
65	57	72	23
66	93	73	9
67	106	74	4
			750

Height, the attribute or characteristic here under consideration, is in this table measured to the nearest inch, giving a group or class interval of one inch. A class interval or class is ordinarily designated by the value of its middle measurement, and the class limits are located on either side at a half unit's distance from this mid-value. All individuals, for instance, with height between 67.5 and 68.5 belong to class 68; here the limits are 67.5 and 68.5 and the class is designated by the number 68. For purposes of computation instead of using the mid-values 61, 62, 63, etc., the classes may be numbered 1, 2, 3, etc, and these numbers used as class numbers. Again, the classes may be numbered in both ways from some point, as 68, within the range. This latter numbering would give classes as follows: —7, —6, —5, —4, —3, —2, —1, 0, +1, +2, etc.

The objects measured or enumerated are referred to as *variates* or simply as *individuals*.

The size or *frequency* of a class is the number of individuals within that class, and the *total frequency* is the sum of all the class frequencies. The table as a whole constitutes a *frequency distribution* and shows the number of times each class occurs.

To illustrate the method of constructing a frequency distribution let us take the following data:

Chicago Monthly Top Beef Cattle Prices

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1916	$9.85	$9.75	$10.05	$10.00	$10.90	$11.50	$11.30	$11.50	$11.50	$11.60	$12.40	$13.00
1915	9.70	9.50	9.15	8.90	9.65	9.95	10.40	10.50	10.50	10.60	10.55	11.60
1914	9.50	9.75	9.75	9.55	9.60	9.45	10.00	10.90	11.05	11.00	11.00	11.40
1913	9.50	9.25	9.30	9.25	9.10	9.20	9.20	9.25	9.50	9.75	9.85	10.25
1912	8.75	9.00	8.85	9.00	9.40	9.60	9.85	10.65	11.00	11.05	11.00	11.25
1911	7.10	7.05	7.35	7.10	6.50	6.75	7.35	8.20	8.25	9.00	9.25	9.35
1910	8.40	8.10	8.85	8.65	8.75	8.85	8.60	8.50	8.50	8.00	7.75	7.55
1909	7.50	7.15	7.40	7.15	7.30	7.50	7.65	8.00	8.50	9.10	9.25	9.50
1908	6.40	6.25	7.50	7.40	7.40	8.40	8.25	7.90	7.85	7.65	8.00	8.00
1907	7.30	7.25	6.90	6.75	6.50	7.10	7.50	7.60	7.35	7.45	7.25	6.55
1906	6.50	6.40	6.35	6.35	6.20	6.10	6.50	6.85	6.95	7.30	7.40	7.90
1905	6.35	6.45	6.35	7.00	6.85	6.35	6.25	6.50	6.50	6.40	6.75	7.00
1904	5.90	6.00	5.80	5.80	5.90	6.70	6.65	6.40	6.55	7.00	7.30	7.65
1903	6.85	6.15	5.75	5.80	5.65	5.15	5.65	6.10	6.15	6.00	5.85	6.00
1902	7.75	7.35	7.40	7.50	7.70	8.50	8.85	9.00	8.85	8.75	7.40	7.75
1901	6.15	6.00	6.25	6.00	6.10	6.55	6.40	6.40	6.60	6.90	7.25	8.00
1900	6.60	6.10	6.05	6.00	5.85	5.90	5.85	6.20	6.15	6.00	6.00	7.50
1899	6.30	6.25	5.90	5.85	5.75	5.75	6.00	6.65	6.90	7.00	7.15	8.25
1898	5.50	5.85	5.80	5.50	5.50	5.35	5.65	5.75	5.85	5.90	6.25	6.25
1897	5.50	5.40	5.65	5.50	5.45	5.30	5.25	5.50	6.00	5.40	6.00	5.65
1896	5.00	4.75	4.75	4.75	4.55	4.65	4.60	5.00	5.30	5.30	5.45	6.50
1895	5.80	5.80	6.60	6.60	6.40	6.00	6.00	6.00	6.00	5.60	5.00	5.50

The width of the classes must first be determined. It would be possible to have a class for each quotation but it would be found highly inconvenient. The error introduced by the grouping of the measurements is ordinarily not of great practical significance. A general rule in determining the width of the classes, and hence of the number of classes, is to make as wide classes as is practically feasible in view of the purposes of the analysis. The number of classes is perhaps most often from ten to twenty. In this case the width is taken as fifty cents and the limiting quotations of each class are included in the class.

On a ruled sheet the classes are written to the left with space to the right for scoring. The data is examined and a score made for each occurrence of the class. Thus Class I with the range $4.50-$4.99 appears February, March, April, May, June, July, 1896. As an occurrence is observed a mark or score is made.

After the scoring is completed the frequency of each class, that is, the number of tallies or scores for each class, is noted and written in a column.

The above operation results in the following frequency distribution:

Class	Frequency Count	Class	Frequency Count
$4.50—4.99	6	$9.00— 9.49	13
5.00—5.49	14	9.59— 9.99	18
5.50—5.99	34	10.00—10.49	5
6.00—6.49	47	10.50—10.99	7
6.50—6.99	25	11.00—11.49	9
7.00—7.49	24	11.50—11.99	5
7.50—7.99	18	12.00—12.49	1
8.00—8.49	12	12.50—12.99	0
8.50—8.99	15	13.00—13.49	1
			264

Exercises

1. Construct a frequency table from the Chicago Monthly Top Beef Cattle Prices using class intervals of one dollar and compare with the distribution obtained with a class interval of fifty cents.

2. Study the frequency distributions of population with respect to age from a Report of the United States Census with special reference to the size of the various class intervals and note two general forms of stating the frequencies of the classes.

3. Examine the different forms of frequency distributions appearing in the report of the Medico-Actuarial Society's Investigations, Vols. I, II, III, IV; also in Biometrika, Agricultural Experiment Station Bulletins and in other accessible sources.

4. In which of the exercises of Chapter I is the data in the frequency distribution form?

Plotting a Frequency Distribution. The illustrative data at the beginning of this Chapter is plotted by locating 14 equidistant points on a horizontal line, one for each height class from 61 to 74 inches inclusive. Then at the middle of each interval a vertical line is erected with a height proportional to the corresponding class frequency. In this way a point is obtained for each class.

As in Chapter II, a rectangle is constructed on each interval. It must be apparent that a rectangle in the case of the frequency distribution has in every case a significant statistical meaning—it is the frequency of the class. Hence the sum of the areas of all the rectangles is the total frequency of the distribution.

Smoothing the Frequency Curve. With the rectangles drawn, the smoothing of a frequency distribution is in no wise different from the smoothing of the data discussed in the preceding chapter. However, for the frequency curve the two rules of the permanence of areas have a stronger justification because of the more definite significance of the areas under the curve.

With practice in the construction of statistical diagrams and curves the rectangles may be dispensed with and the curve drawn by inspection. Also the broken line obtained on joining the ends of the ordinates, called the frequency polygon, may be smoothed by inspection into the required curve. Smoothing by inspection frequently gives as accurate results as the data will justify.

Exercises

5. Smooth the illustrative data at the beginning of this Chapter.

6. Smooth the frequency distribution of Chicago Monthly Top Beef Cattle Prices for 50 cent intervals.

7. From data obtained from a financial paper construct the frequency distribution and smooth curve of the prices of preferred stocks for any one market day.

8. Draw the smoothed curve of the following weight distribution of students:

Weight Class	102	107	112	117	122	127	132	137	142	147	152	157	162	167	172	177	182	187
Frequency	8	13	20	48	76	93	93	110	93	49	56	31	22	13	11	3	2	9

9. Construct the smooth curve of the distribution of ages of a class of high school graduates:

Ages	11	12	13	14	15	16	17	18
Numbers	0	7	45	186	114	61	8	0

Use of the Frequency Curve. The frequency curve does not give a chronological picture of the variations in the data. Instead it shows the number of times that each value occurs.

The frequency curve of precipitation for a dryer climate is located to the left of that for a more moist climate because months with small precipitation occur more frequently in the dryer region. A frequency curve of higher prices lies further to the right than does that of lower prices, so that by constructing the frequency curves of comparative price data it can be readily discovered which series of prices tends to be higher.

Exercises

10. Prepare from the following data the frequency distribution of mean monthly temperatures in degrees at Columbus for January and for July. Plot both distributions on the same diagram and compare.

Years	*Jan.*	*July*	*Years*	*Jan.*	*July*	*Years*	*Jan.*	*July*
1878	..	79	1899	29	76	1920	22	71
1879	26	78	1900	33	76	1921	34	79
1880	44	75	1901	30	80	1922	27	74
1881	24	79	1902	29	75	1923	34	75
1882	33	72	1903	28	74	1924	26	71
1883	27	74	1904	23	73	1925	29	73
1884	20	74	1905	24	75	1926	28	74
1885	23	77	1906	37	73	1927	30	74
1886	24	72	1907	34	74	1928	30	75
1887	27	80	1908	30	76	1929	28	74
1888	27	73	1909	33	72	1930	29	77
1889	34	74	1910	28	75	1931	33	79
1890	39	74	1911	34	76	1932	40	75
1891	33	70	1912	19	75	1933	39	76
1892	24	74	1913	37	76	1934	34	80
1893	19	76	1914	34	76	1935	31	78
1894	35	76	1915	28	73	1936	24	80
1895	24	74	1916	36	79	1937	38	75
1896	31	74	1917	30	74	1938	31	76
1897	26	77	1918	16	72	1939	35	75
1898	33	78	1919	33	76			

Typical or Representative Data. In discussing an increase in prices it is impossible to quote all past prices and recourse must be had to a typical list of prices. The condition of trade in certain industries, for instance, is taken as indicative of the condition of all business. In comparing the prices of beef and the prices of corn the real aim of the investigation may be to find an underlying connection between the two series of values—a connection which will be indicative in any particular year. In such a study the historical statistics of the two price variations are in reality used as representative, as typical, of the manner in which the two prices are related. The frequency form of distribution is peculiarly well adapted to typical data.

Random Sampling. Typical or representative data is usually spoken of as data of *random sampling.* The term *population* is frequently used in a general sense to mean the body of data from which a sample is taken. *By random sampling is meant a taking of a part of the data and making the selection in such a way as not to presuppose any indications.* In other words, data selected by random sampling is assumed to be free from prejudice or bias.

Bias or prejudice would be introduced in the selection of 750 students, to refer to the data of this Chapter as an illustration, if only students of weight over 150 pounds were measured. It is possible that bias might be introduced through the selection of students from one section of the country only. Another possibility for bias would be introduced by taking only students of a certain age. Bias or prejudice, however, is in a sense a relative term and has a meaning only in connection with the purposes for which the data is to be used. If the Student Height Data at the beginning of this Chapter is to be taken as showing the distribution of the heights of freshmen students in college then samples from students of a specific age or specific weight would, in all probability, not give reliable distribution for a general freshmen student population. On the other hand, it might be desired to study the total distribution of students of a special weight class or of certain ages in which cases the samples would not be biased.

In random sampling data there will be variations not arising from bias or prejudicial selection and which have no significance in reference to trends. Thus when 100 coins are tossed again and again there are in very few throws exactly 50 heads and 50 tails.

The errors of random sampling or accidental variations follow the laws of probability and are amenable to mathematical computation. In dealing with representative data it is always necessary to know the limits within which there is a reasonable probability that differences in data are the results of accidental errors of random sampling, and hence are not significant indices of basic differences in the trends.

Random sampling, it must be evident, will yield the most significant and reliable results when the data from which it is taken is *homogeneous* so that it has as few variable characteristics as possible. Thus, if the students referred to were all of one nationality, age and weight the resulting frequency distribution of heights would be expected to give a better fit when applied to another similar set of students than if the group from which the selection was made was comprised of various ages and weights. This matter of homogeneity will be discussed later and some mathematical indices derived for measuring it.

A smooth curve properly drawn is presumed to depict the *underlying trend* in the data, the connection between two series of variates as ages and heights of students, for illustration. The usefulness of such a curve to the statistician, aside from the picture which it presents of the data in hand, lies in the possibilities that the same underlying trend will hold for other data.

One condition, to repeat, under which it may be expected that the underlying trend will apply is that the data be similar to the data from which the trend was derived. This idea is commonly expressed as "conditions being equal."

It must be evident that the more nearly the data is homogeneous, that is, with the individual variates alike in everything except the characteristic being measured,—the more nearly homogeneous the data is—the more reliable and trustworthy will be the derived trends.

Practical Value of the Theory of Random Sampling. The mathematical terminology used in describing and measuring the accidental variations due to random samply may tend to conceal the practical importance of the theory of random sampling. Again, the use of such illustrations, as coin tossing, may tend to give the idea that the applications of this theory might be comparatively unimportant.

Such material as coin tossing is used for illustrations because the application of the laws of probability and hence of random sampling are laid bare, as it were, in such phenomena. These phenomena are not complicated by other conditions. In other words, data derived from coin tossing, say, is highly homogenous. The price of wheat is the result of the interaction of almost innumerable causes, whereas the percentages to be obtained in coin tossing are the results of causes which are comparatively simple.

The degree of dependency which can be placed in a sample taken from a population is, speaking generally, one of the main concerns of the statistician. He must accordingly understand the laws which define the possibilities for errors. In beginning his study of these laws he naturally wishes to work with simple illustrative data.

Classes of Statistical Errors. A clear distinction must be had as to classes of statistical errors.

Data may be in error because of known and understood causes and such errors may be removed by the application of the proper corrections. Wind pressure may cause a more or less constant deflection; instruments of measurement may have a known error; a chronometer may have a standardized correction to be added to or subtracted from each time observation.

A second clases of errors may be labeled mistakes such as copying a number incorrectly. Comparison with other data or differencing are two useful methods of removing this type of error.

Accidental errors constitute a third class. For illustration, the repeated tossing of a coin only infrequently gives results evenly balanced between heads and tails. This type of error has been referred to as variations due to random sampling.

CHAPTER IV

AVERAGES

The Arithmetic Mean. Let us add the January cattle prices in the data of page 27, Chapter III, and divide the sum by the number of items. The result is $7.19. In this way a number, *the arithmetic mean,* is obtained. The characteristic property of this number is that each of the given values may be replaced by it without altering the total sum of the values.

It is usual to speak of the arithmetic mean simply as the mean unless, in order to distinguish the arithmetic mean from some other mean, there is special need for the defining word arithmetic.

Exercises

1. Determine from the data of Exercise 1, Chapter I, the arithmetic mean of the monthly precipitation at Columbus.

2. Find from the data of page 27 the arithmetic mean of the 1895 Top Beef Cattle prices and compare with the 1915 mean.

3. On the assumption that the population of the United States increased uniformly from 1920 to 1930 find the value of the annual increase and then the estimated population for 1906.

4. Compute the arithmetic mean of the Chicago Monthly Top Beef Cattle prices for the years 1895 to 1916.

5. By first assigning each monthly price of Exercise 4 to the appropriate 50-cent class and computing the arithmetic mean of the prices when so altered determine the effect on the value of the arithmetic mean by substituting the class prices for the exact values. Use the class numbers in computation and translate the result in terms of the proper interval.

6. In Exercise 4 there are 264 entries in the sum to be added. Show that much of the labor of the addition can be avoided by selecting the equal prices, then multiplying each by the number of times it occurs, and adding the resulting products to obtain the total sum of prices.

The results of Exercises 4 and 6 suggest the computing of the mean from a frequency table in accordance with the following rule: multiply each deviation by its frequency, add the resulting products, and divide this total sum by the total frequency. The quotient is the value of the mean. Thus, from the frequency distribution of Top Beef Cattle Prices of Chapter III, obtained on page 27, 6, 14, 34, 47, 25, 29, 18, 12, 15, 18, 18, 5, 7, 9, 5, 1, 0, 1, the mean price is given by the expression—

$$d = \frac{1\times6+2\times14+3\times34+4\times47+5\times25+6\times29+7\times18+8\times12}{264}$$
$$\frac{+9\times15+10\times18+11\times18+12\times5+13\times7+14\times9+15\times5+}{264}$$
$$\frac{+16\times1+17\times0+18\times1}{264},$$

$= 6.61$, where d is the distance of the computed mean from the origin.

The mean class is thus 6.61; that is, 6.61 of the 50-cent intervals or \$3.30 from the origin which is the mid-point of the class preceding the first. The mid-point of this class is 4.24 hence the mean is 7.54.

Whenever the frequency table is available, the method just described is usually the shortest method of computing the value of the mean. However, if the frequency distribution is not needed for any other purpose and especially if an adding machine is at hand the saving of time in the computation of the mean does not ordinarily justify the compilation of a frequency table merely for the one purpose of finding the mean.

The following is the computation for mean height from the data at the beginning of Chapter III.

Let us take the origin at height 60. Then the computation scheme will be as follows:

Computation of the Mean
Student Height Data

Class	*Deviation*	*Frequency*	*Dev. Times Freq.*
61	1	2	2
62	2	10	20
63	3	11	33
64	4	38	152
65	5	57	285
66	6	93	558
67	7	106	742
68	8	126	1,008
69	9	109	981
70	10	87	870
71	11	75	825
72	12	23	276
73	13	9	117
74	14	4	56
		750	5,925

$$d=\frac{5,925}{750}=7.90$$

Hence the mean height is 7.9 classes, that is, 7.9 inches from the origin, 60, and is therefore equal to 67.9 inches.

Statistical Properties of the Arithmetic Mean. What is the statistical significance and interpretation of the arithmetic mean? If a higher price were substituted for one of the January beef cattles prices the resulting arithmetic mean would be larger, but not as much larger as the individual price because in the process of obtaining the mean the price increase is divided by the total number of prices. Hence a larger mean denotes that, as a whole the values of the distribution are greater, and a smaller arithmetic mean is to be interpreted as indicating a relatively lower series of values. Furthermore, the arithmetic mean is relatively more stable than is an individual measurement.

That is, since all increases and decreases are divided by the number of variates the changes in the value of the arithmetic mean are relatively smaller than are those of the individual values. Thus a decrease of 50 cents must occur in each of the months in order to decrease the arithmetic mean by the same amount. A decrease of 50 cents in one-half the variates decreases the arithmetic mean by only 25 cents, and so on.

The relative stability of the arithmetic mean when applied to the student height data means that if several groups of 750 students were measured for height and the frequency distributions tabulated and the means computed for each group it would be found that the means would vary but little while the frequency of any one class, 67 inches for instance, would vary considerably from distribution to distribution.

It is to be noted that a single increase of 50 cents in the price of one month has exactly the same effect on the value of the arithmetic mean as does a 10 cent increase in the prices of each of five months. But is this true statistically? Should the exceptionally high price be given so much weight? Should the persons of exceptional height be emphasized so strongly in the group of persons whose height is measured?

This emphasizing of extreme values raises a question of whether the value of the mean may always be significant. Whether an item is unduly large can be determined only from a study of the data itself for the mean conveys no information whatever as to the distribution of the variates; it tells only of their general size. *That is, the statistical function of the arithmetic mean is essentially to measure the size or magnitude of the data as a whole.*

Theorem. *In any disribution the sum of the deviations from the mean is zero.* That is, the sum of the positive deviations or the measurements to the right of the mean is equal to the sum of the negative deviations or the measurements to the left of the mean. The distance of the mean from any origin is obtained by taking the sum of the deviations from that origin and dividing by the total frequency, hence when this distance is zero, that is, when the origin is at the mean, the total sum of the deviations must be zero.

Weighted Arithmetic Mean. An apparent modification of the arithmetic mean is illustrated by the following. It is desired

Computation of the Mean
Student Height Data

Class	*Deviation*	*Frequency*	*Dev. Times Freq.*
61	1	2	2
62	2	10	20
63	3	11	33
64	4	38	152
65	5	57	285
66	6	93	558
67	7	106	742
68	8	126	1,008
69	9	109	981
70	10	87	870
71	11	75	825
72	12	23	276
73	13	9	117
74	14	4	56
		750	5,925

$$d=\frac{5,925}{750}=7.90$$

Hence the mean height is 7.9 classes, that is, 7.9 inches from the origin, 60, and is therefore equal to 67.9 inches.

Statistical Properties of the Arithmetic Mean. What is the statistical significance and interpretation of the arithmetic mean? If a higher price were substituted for one of the January beef cattles prices the resulting arithmetic mean would be larger, but not as much larger as the individual price because in the process of obtaining the mean the price increase is divided by the total number of prices. Hence a larger mean denotes that, as a whole the values of the distribution are greater, and a smaller arithmetic mean is to be interpreted as indicating a relatively lower series of values. Furthermore, the arithmetic mean is relatively more stable than is an individual measurement.

That is, since all increases and decreases are divided by the number of variates the changes in the value of the arithmetic mean are relatively smaller than are those of the individual values. Thus a decrease of 50 cents must occur in each of the months in order to decrease the arithmetic mean by the same amount. A decrease of 50 cents in one-half the variates decreases the arithmetic mean by only 25 cents, and so on.

The relative stability of the arithmetic mean when applied to the student height data means that if several groups of 750 students were measured for height and the frequency distributions tabulated and the means computed for each group it would be found that the means would vary but little while the frequency of any one class, 67 inches for instance, would vary considerably from distribution to distribution.

It is to be noted that a single increase of 50 cents in the price of one month has exactly the same effect on the value of the arithmetic mean as does a 10 cent increase in the prices of each of five months. But is this true statistically? Should the exceptionally high price be given so much weight? Should the persons of exceptional height be emphasized so strongly in the group of persons whose height is measured?

This emphasizing of extreme values raises a question of whether the value of the mean may always be significant. Whether an item is unduly large can be determined only from a study of the data itself for the mean conveys no information whatever as to the distribution of the variates; it tells only of their general size. *That is, the statistical function of the arithmetic mean is essentially to measure the size or magnitude of the data as a whole.*

Theorem. *In any disribution the sum of the deviations from the mean is zero.* That is, the sum of the positive deviations or the measurements to the right of the mean is equal to the sum of the negative deviations or the measurements to the left of the mean. The distance of the mean from any origin is obtained by taking the sum of the deviations from that origin and dividing by the total frequency, hence when this distance is zero, that is, when the origin is at the mean, the total sum of the deviations must be zero.

Weighted Arithmetic Mean. An apparent modification of the arithmetic mean is illustrated by the following. It is desired to obtain an index of food prices by taking the mean of the price quotations of 15 articles of food. It is decided, however, that one of the quotations should be given twice the weight of the other articles. This is done by multiplying this quotation by two and taking the double quotation in the total sum. The article is said to have a *weight* of two. The idea of weight introduces no new principles into the computation of the arithmetic mean. Stated in another way, there is no change in the properties of the arithmetic mean if some of the variates are identical in value. Again, when the arithmetic mean is computed from a frequency distribution the frequencies may be looked upon as weights.

Adjustment or Graduation Formulas. An adjustment or graduation formula of wide and convenient adaptability to the smoothing of data is based on the arithmetic mean. It is called the moving average.

A *moving average* based on the average of five consecutive terms is the result of adding the first five terms and dividing by five and using this average as the third term, and then using the average of the second to the sixth terms for the fourth term and so on. When the average has been computed for each such set of five terms a new series will be at hand for all except the original two beginning terms and the final two. These four terms may be smoothed out by inspection, if necessary.

The resulting series after the substitution of the first series of moving averages will give a smoother curve than a curve from the original data. The process of the moving average can be repeated as applied to the successive averages. The moving average method of graduating or smoothing data is especially applicable to data where there are unaccountable but similar fluctuations in the data from class to class. This method works best where there are a fairly large number of classes.

An extensive application of this method has been made in the graduating of mortality tables. It is often used in smoothing data in which the general trend is obscured by the presence of more or less regular fluctuations. In this latter case the number of classes grouped together should be determined by the lengths of the cycles of the fluctuations. If the cycles are irregular in length the method of the moving average is not likely to yield satisfactory results.

Exercises

7. Smooth the data of Student Heights at the beginning of Chapter III by taking the means of each successive three terms, then of five terms.

The Geometric Mean. The arithmetic mean, it has been seen, may be substituted for each item of the data and leave the sum of all the items unchanged. Where the data consists of percentages of increase or decrease it may be convenient to obtain an average percentage which when used in place of each of the variable percentages will give the same final figure. Such an average is called the *geometric mean.*

Let the price of a certain article for each year from 1934 to 1939, to illustrate the geometric mean, be expressed as a percentage of the preceding year as follows (assuming 100 for the 1934 price), 105, 118, 109, 102, 115. The percentage change from 1934 to 1939 is obtained by multiplying together the five percentages and is approximately 158. What uniform annual percentage of increase will give the same percentage of increase of 1939 over 1934? Let $(1 + r)$ be the constant multiplier or percent. Then we have

$$\begin{aligned}(1 + r)^5 &= 105 \times 118 \times 109 \times 102 \times 115, \\ &= 1.58415.\end{aligned}$$

$$\begin{aligned}\text{and } (1 + r) &= \sqrt[5]{1.58415}, \\ &= 1.0964 \text{ (by logarithms).}\end{aligned}$$

Each of the unequal increases in the series may therefore be replaced by the factor 1.0964 and still give the same product.

The population of continental United States in 1930 was 122,497,000 and in 1920, 105,711,000. On the assumption of a uniform rate of increase during the decade what should be the value of this uniform rate in percent? As above, we have

$$(1+r)^{10} = 122{,}497/105{,}711 = 1.158791.$$

$$\text{Hence } (1+r) = \sqrt[10]{1.158791}$$
$$= 1.0148$$

It may be noted that according to this method the population in 1923 was equal to $105{,}711{,}000 \times (1.0148)^3 = 110{,}474{,}338$.

For any but the simplest problems the computation of the geometric mean cannot be accomplished without the use of logarithms. The following computation of the geometric mean of student heights from the data of Chapter III illustrates the process.

The geometric mean height $= (61^2 \times 62^{10} \times 63^{11} \times 64^{38} \times 65^{57} \times 66^{93} \times 67^{106} \times 68^{126} \times 69^{109} \times 70^{87} \times 71^{75} \times 72^{23} \times 73^{9} \times 74^{4})^{\frac{1}{750}}$

Hence $750 \times \log$ geometric mean $=$

$$\begin{aligned} &2 \log 61 + 10 \log 62 + 11 \log 63 + \\ &38 \log 64 + 57 \log 65 + 93 \log 66 + \\ &106 \log 67 + 126 \log 68 + 109 \log 69 + \\ &87 \log 70 + 75 \log 71 + 23 \log 72 + \\ &9 \log 73 + 4 \log 74 = 1373.\ 70315. \end{aligned}$$

On dividing by 750, we have log geometric mean $= 1.83160$. The number of which this is the logarithm is 67.858 or 67.86, to two places of decimals.

Hence the geometric mean height is 67.86. It is interesting to note that this geometric mean for this data, 67.86, is very close to the arithmetic mean which was found to be 67.90.

Properties of the Geometric Mean. The geometric mean, unlike the arithmetic mean, is most affected by the smaller deviations because a small factor in a product has a proportionately greater influence on the result of a multiplication than does a large factor.

Each property of the arithmetic mean has a corresponding property for the geometric mean because the logarithm of the geometric mean is the arithmetic mean of the logarithms of the deviations. From this logarithmic correspondence all the properties of the geometric mean can be derived from those of the arithmetic mean. It is apparent, for instance, that the geometric mean is the result of a series of deviations multiplied together in a way exactly parallel to that of the arithmetic mean and a series of terms added together. Illustrations of this parallel are, a chain of relative prices and a series of price increases; interpolation on the assumption of a uniform rate and of a uniform increase; compound interest and of simple interest.

Makeham's Law of Mortality. Reference is made to the descriptive part of Appendix III for an application of the general idea of rates of change. The derivation of Makeham's Law of Mortality as given in this Appendix should be studied by readers with some mathematical training.

The Median. Let the years 1901 to 1939 inclusive be arranged in order of the March precipitation at Columbus beginning with the lowest. We then have, with the data measured to hundredths of an inch:

1910	0.28	1923	3.04
1915	1.19	1936	3.10
1931	1.34	1920	3.32
1937	1.56	1917	3.59
1929	1.76	1927	3.97
1901	1.82	1903	4.13
1918	1.85	1924	4.28
1905	1.87	1938	4.32
1930	2.13	1922	4.54
1926	2.16	1912	4.56
1932	2.20	1919	4.58
1934	2.20	1906	4.59
1925	2.25	1916	4.88
1911	2.36	1904	4.93
1939	2.41	1907	5.21
1914	2.46	1933	5.44
1902	2.63	1908	6.03
1909	2.68	1921	7.66
1928	2.79	1913	8.09
1935	2.81		

The middle year, 1935, in this ordered arrangement is called the *median year* with respect to March precipitation, the median precipitation of 2.81 inches being that of the median year, with nineteen years having a smaller precipitation and nineteen years having a larger precipitation.

In general the *median individual is defined as the individual so located that there are as many individuals with a greater value of the characteristic as with a lesser value,* and the middle value of the measured characteristic is spoken of as the *median value* of the *characteristic.*

If the number of variates or measurements is even, the median is assumed to lie between the two middlemost variates.

It is obvious that the above median precipitation year might have been obtained by a simple process of counting and inspection of the data without setting down the variates in order.

Exercises

8. From the data of Exercise 10, Chapter III determine the median temperatures for January and July at Columbus.

9. From the Chicago Monthly Top Beef Cattle price data of Chapter III determine the median monthly price.

When the data is in the form of a frequency distribution the computation of the position of the median is much facilitated. All that is necessary then is to start from one extremity of the distribution and include successive classes until half the total frequency is obtained. The only point of difficulty in this case is when the median is located within a class. Then it is necessary to interpolate within the median class for the more exact position of the median. To illustrate the method of interpolation let us find the medium student height from the data at the beginning of Chapter III. Half of the number of variates is 375. Counting from the lower extremity we find, up to and including class 67, a frequency of 317, so that it is necessary to take 58 individuals from class 68. Hence it may be assumed that the position of the median will be 58/126 of a unit or 0.46 inches from the left boundary of class 68. Since this boundary is at 67.5 the median is located at 67.96 inches.

Geometrically, the median deviation locates the ordinate *which divides the area under the frequency curve into two equal parts.*

The median can be found directly from a *cumulative curve* by drawing a horizontal line through the point on the vertical scale corresponding to half the total frequency. The abscissa of the point of crossing of this horizontal line and the curve is the median deviation.

Exercises

10. By drawing the cumulative curve locate the median student height.

11. From the frequency distribution of Chicago Monthly Top Beef Cattle prices of Chapter III determine the median price by using the cumulative curve.

12. What is the median point of population as determined by the Bureau of Census?

13. Distinguish the median point of population from the center of population.

Quartiles. Each half of the distribution, one on either side of the median, may be divided into two equal parts. These two points of division are the *First and Third Quartiles.*

The two quarters and the median thus divide the variates into four classes of equal frequencies.

In data having predominately large frequencies near the median the quartiles are relatively close to the median, and in widely scattered data the quartiles are relatively far from the median. This property of the quartiles is developed and applied in the following chapter.

Deciles. The decile variates are the variates which separate the frequency into ten equal classes. The median is, of course, the fifth decile but the quartiles are not deciles. The chief use of the deciles, like that of the quartiles, is in determining the shape of the distribution.

Exercises

14. Determine the quartile and decile prices from the data of Chicago Monthly Top Beef Cattle prices of Chapter III.

Statistical Properties of the Median. The position of the median ordinate depends solely on the relative values and not on the actual values of the variates. The data need be given with only enough exactness to permit the arrangement of the variates in order with respect to the attribute considered. Moreover, it is only the arrangement near the median value that must be carefully attended to, consequently the median can not give detailed information of the variates at the extremities of the ranges.

There is apparently no apriori reason why the value of the median should not show considerable variation from sample to sample taken from the same material, but in practice it is found that the median shows as high, if not higher, degree of stability than does the arithmetic mean. Thus if a second group of 750 students were measured as to height and the median computed it would most likely be found to differ only slightly from that of the group already discussed. This slowness of change in the median means that the median is not greatly affected by the presence of accidental and irrelevant influences. That is, differences in the value of the median are not likely to be merely accidental and hence the median significantly measures properties of the material. For instance, a distribution of wages showing a higher median wage is most likely to be significantly a group of higher wages.

The Probable Deviation. The median variate divides the data into two classes of equal frequencies. Hence it is an even chance that an individual selected at random will fall into a designated one of the two classes. If the median height of freshmen students is 68 inches it is an even bet, for instance, that a student concerning whose height nothing is known has a height less than 68 inches.

Likewise it is an even bet that a student selected at random will have a height between the first and third quartiles. The range from the median to the third or first quartile, one-half of the range within which the chances are even for an individual measurement to lie, is called the *probable deviation.*

Exercises

15. Determine the probable deviation for Chicago Monthly Top Beef Cattle prices from the data of Chapter III.

The Mode. It is to be noted that in the frequency distribution of student heights, class 68 has the greatest frequency or number of students and that the high point on the frequency curve is within the same class. The class of greatest frequency is called the *modal class* and under the curve the deviation with the highest ordinate, the *modal deviation. A mode is thus defined as a class or deviation of greatest frequency, more accurately, it is the class or deviation of greater frequency than that of either the class immediately greater or immediately less.* This second definition allows for distributions having more than one mode.

Exercises

16. Determine the modal deviation of the student weight classes from the data of Exercise 8, Chapter III.

17. Determine the location of one or more modal prices of Chicago Monthly Top Beef Cattle prices from the data of Chapter III.

It is possible to locate the mode within a class by a process of interpolation similar to that described in the determination of the median but by far the easiest method is to construct the smooth frequency curve and determine the abscissa or deviation of the greatest ordinate.

When the data seems to have more than one mode care must be exercised in deciding whether to smooth out the apparent modes. In a frequency distribution of monthly temperatures it is evident that there are summer and winter modal temperatures. The telephone calls data of Exercise 18, this Chapter, shows more than one mode. On the other hand, the data of age distribution reported by the United States Census Bureau shows a tendency for the frequencies at the even ages to be larger than at the odd ages. This latter tendency is partly due to the fact that persons who are uncertain as to their exact age seem to show a preference for an even number. These apparent modes should be smoothed out. Data with essentially one mode is said to be *unimodal;* with more than one mode, *multimodal.*

Exercises

18. Smooth the following data of the telephone calls for one day at a business exchange and locate the modes.

Time	6-7	7-8	8-9	9-10	10-11	11-12	Noon 12-1	1-2	2-3
Calls	1595	3430	6389	6904	7282	7358	6361	5659	6186
Time	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12
Calls	6597	6510	6093	4508	4210	2289	1197	916	314

19. Do the same for the following residence calls.

Time	6-7	7-8	8-9	9-10	10-11	11-12	Noon 12-1	1-2	2-3
Calls	1256	3796	6604	4098	4240	3816	5852	4421	3136
Time	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12
Calls	4344	3268	4541	4778	4039	2088	1176	665	187

Statistical Properties of the Mode. Since the modal class or deviation is that of greatest frequency, that is, since more variates belong to that class than to any other within its immediate vicinity, the mode is the most typical of the variates of a distribution. If any one variate is to be selected as descriptive of the data the modal variate should be that variate. *The mode is accordingly said to define the type of the distribution.* The significance of the mode as a type depends, of course, on the relative preponderance of its frequency. Thus the frequency of height 68 in the student height distribution of Chapter III is 126 and the total frequency of the classes near the modal class is a large percentage of the total of all the frequencies. In the Chicago Monthly Top Beef Cattle prices of Chapter III the modal price class of $6.00-$6.49 has a frequency of 47 and the decline in frequencies on either side of this class is not as rapid as is shown in the height data. Data showing a strong tendency to concentrate about the mode is said to be *highly stable* or *true to type.* Measures of *trueness to type* are discussed in the following chapter.

The position of the mode may depend only on the values of a few variates so that the mode, like the median, gives little information of the extremes of the range.

The mode cannot be accurately determined by an elementary process of arithmetic as can the median and the mean.

The mode being the predominating value, the type, the fashion, is what is ordinarily in the popular mind when an average is spoken of. The statement that the average person spends one-third of his income in rent is most likely to mean that more persons spend about that percentage than any other percentage.

Consistency of Averages. The arithmetic mean, the geometric mean, the median, and the mode for the student height data of page 26 are all close together in value—67.90, 67.86, 67.06, and 68. respectively. Thus, close agreement is an indication that the data is not erratic and hence is stable, reliable and truly representative or typical of the population from which it is a sample.

It is easy to set down figures of a distribution which may show a wide divergence in the averages but such divergences do not ordinarily arise in actual data unless the data has complex tendencies. It accordingly appears that any marked differences among the averages always call for a critical examination of the data for the causes of such differences.

It will usually be found that where the data shows a distinct *mode* that the median and arithmetic mean will be in close agreement with the mode. Where the frequencies are comparatively high near the ends of the distribution there may be distinct differences among these three averages. Later methods will give exact mathematical indices of the forms of distributions but the fact must not be overlooked that the study of these averages gives very important information on the form of a distribution.

CHAPTER V

THE FORM OF A DISTRIBUTION

Dispersion. It is stated in the preceding Chapter that the significance of the mode as a representative of the data depends on the extent to which the data conforms to the mode as a type. That is, if the sum of the frequencies near the mode is a relatively large percentage of the total frequency the modal deviation is highly typical and the data is not subject to great variations. The word variation is used because if a certain type does not predominate in the data different samples will have a tendency to show widely differing, that is, varying distributions. To illustrate, if the modal frequency in another sample of 750 student heights is only 95 instead of the 126 for the data already studied with a similar reduction in the other larger frequencies and with consequent larger frequencies farther from the mode, then this second distribution will not be so true to the type expressed by the mode as in the first distribution.

To repeat, a distribution with small frequencies at the ends of the ranges and with the frequencies concentrated about a point is said to be *true to type,* to be *highly stable.* Some of the methods of measuring the extent to which the data is scattered or spread or dispersed about the class of concentration are now to be considered.

Measures of Dispersion. Because the *breadth of the range* depends on the usually uncertain data at the extremes it does not furnish a reliable measure of the extent to which the data is dispersed. As given in Chapter III the range of student heights is 14 inches, from 61 to 74 inches, so that the inclusion of a single student of height 58 inches would increase this range by more than twenty per cent.

We have seen that in theory the dispersion should be measured from the mode but in practical statistical work the mean, median, and mode usually differ so little in position that it is ordinarily permissible to measure the dispersion from the mean. As was shown in the preceding chapter the arithmetic mean is most convenient from the mathematical standpoint.

The *sum of the deviations about the mean* is useless as a measure of dispersion because, as was proved in Chapter IV, this sum is zero regardless of the spread or dispersion of the distribution.

Mean Deviation. Since the object in measuring dispersion is to determine the deviation of the variates from an average it is the numerical amount of a deviation that counts and not its direction. Hence a logical measure of dispersion is obtained by adding the deviations, all counted positive, and then dividing this sum by the total frequency. This gives the *mean deviation.*

The form for the computation of the mean deviation is the same as for the arithmetic mean except that all negative signs are disregarded.

The following is the computation of the mean deviation from the arithmetic mean of the Student Height Data.

The arithmetic mean has already been computed at 67.9 inches.

Computation of the Mean
Student Height Data

Class No.	*Diff.*	*Freq.*	*Prod.*
1	6.9	2	13.8
2	5.9	10	59.0
3	4.9	11	53.9
4	3.9	38	148.2
5	2.9	57	165.3
6	1.9	93	176.7
7	0.9	106	95.4
8	0.1	126	12.6
9	1.1	109	119.9
10	2.1	87	182.7
11	3.1	75	232.5
12	4.1	23	94.3
13	5.1	9	45.9
14	6.1	4	24.4
		750	1,424.6

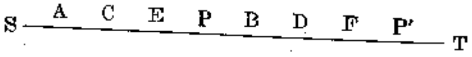
On dividing 1424.6 by the total frequency we have 1.9.

The mean deviation is accordingly 1.9 classes. Since each class interval is one inch the mean deviation from the arithmetic mean is 1.9 inches. This result shows that the average length of the variations from the arithmetic mean is 1.9 inches in this illustrative data.

For purposes of comparing the stability of different distributions it is desirable to divide the mean deviation by the mean or median, whichever is used as a base. When this is done the mean deviation is expressed as a fraction of the base average. For instance, it seems reasonable to say that a mean deviation of 0.3 with an arithmetic mean of 20 has the same significance as a mean deviation of 0.9 based on an arithmetic mean of 60.

Because, as is presently proved, the mean deviation is least when taken about the median there is a theoretical advantage in computing the mean deviation about the median. When so done there is a certain degree of standardization which is not attained with any other average as a base, but the point is not of great practical importance unless the median and the arithmetic mean differ markedly.

Proof that the Mean Deviation Is Smallest When Taken About the Median. Let *P* be a point on the line S—T between the points A and B. The sum of the deviations of P from A and B is, without regard to the sign of the deviations PB + PA, and this sum is equal to AB. If *P* should lie without the segment AB, as *P'*, the sum of the two deviations would be greater than AB. Likewise the sum of the distances of P from any other two points C and D is least when *P* lies between them. Hence the total sum of deviations of P from any number of points is least when there are as many points on one side of P as on the other; that is, when *P* is the median of the points.



Exercises

1. According to the measure supplied by the mean deviation which is the more variable, the monthly mean temperature or the monthly mean precipitation at Columbus, based on data given in a previous Chapter?

2. From the data of student heights and student weights of Chapter III determine which is the more variable as measured by the mean deviation.

Statistical Properties of the Mean Deviation. The mean deviation as a measure of dispersion has much to be said for it—it takes all the variates into account; it takes each variate according to its size and hence, as the arithmetic mean, may be unduly influenced by extreme variates.

The mean deviation is an index of dispersion of practical importance and frequently gives a sufficient measure of dispersion, though for many, if not most, purposes the mean squared deviation to be presently discussed is more convenient.

The Mean Squared Deviation. The mathematically simplest device for eliminating negative signs is that of squaring the terms. If the difference between each deviation and the mean be squared, the sum of the squares added, and the resulting sum divided by the total frequency the *mean squared deviation,* thus obtained, is a measure of dispersion which is arithmetically more convenient than is the mean deviation.

The computation of the mean squared deviation differs from the computation of the mean deviation only in that the deviations from the mean are squared before multiplication by the frequencies. It is, of course, possible to compute directly from the data without using the frequency table but ordinarily only a slight error is introduced by combining the actual values into reasonably narrow classes and much labor may be saved because only one multiplication is then required for each class instead of one for each individual variate.

Computation of Mean Squared Deviation from the Mean Student Height Data

Class	*Diff.*	*Squares*	*Frequency*	*Products*
1	6.9	47.61	2	95.22
2	5.9	34.81	10	348.10
3	4.9	24.01	11	264.11
4	3.9	15.21	38	577.98
5	2.9	8.41	57	479.37
6	1.9	3.61	93	335.73
7	0.9	.81	106	85.86
8	0.1	.01	126	1.26
9	1.1	1.21	109	131.89
10	2.1	4.41	87	383.67
11	3.1	9.61	75	720.75
12	4.1	16.81	23	386.63
13	5.1	26.01	9	234.09
14	6.1	37.21	4	148.84
			750	4,193.50

On dividing the sum of products by the total frequency we have a mean squared deviation of 5.59 classes.

Exercises

3. Determine the mean squared deviation about the arithmetic mean of student weights from the data of Exercise 8, Chapter III.

4. Determine the mean squared deviation about the arithmetic mean of Chicago Monthly Top Beef Cattle prices of Chapter III.

5. Determine the mean squared deviation about the arithmetic mean of monthly precipitation at Columbus from the data of Chapter I.

6. Determine the mean squared deviation of monthly temperatures at Columbus from the data of Exercise 1, Chapter I.

The above method of computing the mean squared deviation involves fractional deviations. By the following short rule fractions can be avoided in the computation.

Short Rule for the Mean Squared Deviation. Under the short rule for computing the mean squared deviation an integral deviation near the actual arithmetic mean is selected and the differences between each deviation and this selected point are computed. Then each of the differences are squared and multiplied by the corresponding frequencies. The sum of these products is divided by the total frequency which result gives the mean squared deviation from the selected point. The mean squared deviation is obtained from the arithmetic mean and from the value just computed by subtracting from this computed value the square of the difference between the true arithmetic mean and the selected integral point. The proof of this latter relationship is as follows: Let the mean squared deviation about the actual arithmetic mean be denoted by the square of the Greek letter σ (sigma), and the mean squared deviation about any other point by the square of the same symbol written with a prime, σ'. We then have, on recalling that the letter d is used to denote the deviation of the arithmetic mean from the origin, the following formula:

$$\sigma^2 = \sigma'^2 - d^2.$$

To prove this formula let the deviations from the selected point be denoted by X with a subscript for each class and the deviations from the arithmetic mean by x and let the distance of the mean from the selected origin be denoted by d. Then $X = x + d$ for each individual or class in the distribution.

The standard deviation is obtained by squaring the X for each class, adding, and dividing by the total frequency. Performing these operations we have

$$X_1^2 = x_1^2 + 2dx_1 + d^2$$

$$X_2^2 = x_2^2 + 2dx_2 + d^2$$

$$\begin{array}{ccccc} . & = . & . & . \\ . & = . & . & . \\ . & = . & . & . \end{array}$$

and on adding we have

$$\Sigma X^2 = \Sigma x^2 + 2d\ \Sigma x + \Sigma d^2.$$

But $\Sigma x = o$, by the theorem of Chapter IV which states that the sum of the variations from the arithmetic mean is zero.

Also $\Sigma X^2 = N\sigma'^2$,
and $\Sigma x^2 = N\sigma^2$,
and $\Sigma d^2 = Nd^2$.
Hence, we have,

$$N\sigma'^2 = N\sigma^2 + Nd^2 ,$$

or, $\sigma'^2 = \sigma^2 + d^2$.
By transposition,

$$\sigma^2 = \sigma'^2 - d^2 .$$

To apply the foregoing short method for computing the Standard Deviation let us take in the Student Height Data a selected origin at class 68. The mean squared deviation is then obtained by the following computation.

Computation of Mean Squared Deviation by Shortened Method Student Height Data

Class	*Dev.*	*Dev. Squared*	*Freq.*	*Prod.*
1	7	49	2	98
2	6	36	10	360
3	5	25	11	275
4	4	16	38	608
5	3	9	57	513
6	2	4	93	372
7	1	1	106	106
8	0	0	126	0
9	1	1	109	109
10	2	4	87	348
11	3	9	75	675
12	4	16	23	368
13	5	25	9	225
14	6	36	4	144
			750	4,201

The mean squared deviation from class 8, that is, 68, is

$$\frac{4201}{750} = 5.60.$$

Since the arithmetic mean is 67.9, $d = 68 - 67.9 = 0.1$ and $d^2 = 0.01$.

Therefore $\sigma^2 = 5.60 - 0.01 = 5.59$

and $\sigma = 2.36$.

The mean squared deviation here computed, 5.59, agrees with the same value computed on a preceding page.

Exercise

7. Recompute the mean squared deviation about the arithmetic mean of Exercises 3, 4, 5, 6 using the shortened method.

Mean Squared Deviation Least About the mean. The mean squared deviation is least when taken about the arithmetic mean. This fact of the minimum value follows at once from the formula

$$\sigma'^2 = \sigma^2 + d^2 .$$

Thus taking the standard deviation about a point numerically distant from the arithmetic mean by d increases the mean squared deviation by d^2.

Variance and Standard Deviation. It has just been seen that the mean squared deviation about any point other than the mean is equal to the mean squared deviation about the mean plus the square of the distance of such point from the mean and hence the mean squared deviation is least when taken from the mean. There is accordingly a peculiar fitness in the mean squared deviation about the mean as a measure of dispersion of the data. *The term variance is used to denote the mean squared deviation about the mean. In this terminology the standard deviation is the square root of the variance.*

This minimum characteristic gives a practical and theoretical preference to the standard deviation over that of any other mean squared deviation. For this reason, and because certain other computations are rendered simpler by so doing, the mean squared deviation about any other value than the arithmetic mean is seldom computed even though the idea of trueness of type centers about the mode. Since the mean and the mode rarely differ by more than a small amount the square of this difference will be relatively still smaller and as a result, the difference between the square of the standard deviation and the mean squared deviation about the mode is ordinarily negligible.

Properties of the Standard Deviation. Since a small value for the standard deviation can arise only when the variates are closely concentrated about the mean or mode and since a large value must be due to a relatively high frequency of the variates near the extremes of the distribution, the standard deviation is a measure of the dispersion of the data. Because the effect of squaring is to diminish the importance of the smaller values and to exaggerate the importance of the larger values a small value for the standard deviation shows conclusively that the data is highly true to type and stable, while, on the other hand, a large value may to some extent be due to the presence of the larger frequencies of the extreme variates and hence not altogether significant. But even with this qualification in regard to large frequencies near the limits of the range the *standard deviation is a most practicable and reliable index of the dispersion of data.*

Exercises

8. Discuss the comparative variabilities of the distributions for which the standard deviations have been computed in the preceding exercises of this Chapter.

9. Does a standard deviation of 2.4 for height of students denote a smaller variation than a standard deviation of 15 pounds of weight?

The Coefficient of Variability. As in the case of the mean deviation the significance of a value for the standard deviation depends on the size of the variates. A variation of 10 feet in a measurement of five miles is of the same degree of accuracy as a variation of 2 feet in one mile.

It is, therefore, reasonable to divide the standard deviation by the mean in order to express it as a fraction of the size of the variates. This quotient is ordinarily quite small, so that it is usual to multiply it by 100. The resulting coefficient—*100 times the standard deviation divided by the mean*—is called the *coefficient of variability.*

For the Student Height Data the coefficient of variability is accordingly:

$$\frac{2.36}{67.9} \times 100 = 3.48$$

Exercises

10. Compare the value of the coefficient of variability for height with that for weight for students as shown by the data of a preceding Chapter.

11. Discuss the comparative variability of monthly precipitation and monthly temperatures at Columbus from the data of Chapter I.

The Quartiles as Measures of Dispersion. The distance from the median to the third quartile is the interval that includes half the frequencies to the right of the median. Now if these distances are relatively large it must mean that the frequencies at the center are not large in comparison with the total frequency. That is, if the first and the third quartiles are close together the distribution must be closely concentrated about the median; must be highly typical; must show a low degree of variability; because in every case one-half the total frequency is included between these two quartiles. If the interval is narrow the ordinates near the mean must be tall, that is, the frequencies in the center must be predominantly large, in order to include half the total frequency. If the data has a flat frequency curve so that the degree of variability is large and the trueness to type small the two quartiles will be comparatively far apart.

Ordinarily the distance between the first quartile and the median is approximately equal to the distance from the median to the third quartile so that the distance from the median to the third quartile is taken as the index of dispersion of the distribution. This distance is called the *probable deviation.*

Since half the total number of frequencies are included between the two quartiles the chances are even that an individual of the group, selected at random, will have a deviation lying between the quartile deviations. In other words, the chances are even that an individual selected at random from the group will have a deviation numerically less than the probable deviation. If in one group of 750 students, for instance, it is an even bet that a student selected at random has a height between 64 and 72 inches and in a second group the range for even chances is from 67 to 69, the second group with the narrow range between the two quartiles is said to be the more *true to type.*

Formula for the Probable Deviation. The probable deviation can always be found by the simple process of locating the quartiles. It is proved in the following chapter that for a certain special, though very frequently occurring, form of distribution the probable deviation is equal to the standard deviation multiplied by a constant, 0.6745.

In symbols, we have P.E. $= 0.6745\ \sigma$, where the symbol P.E., inherited from the theory of errors developed by Gauss, denotes the probable deviation. From the Student Height Data the probable deviation is 0.6745×2.36 which equals 1.59. This value for the probable error means that the chances are even that a student selected from this group at random will have a height between $67.9 - 1.59$ and $67.9 + 1.59$ or 66.31 and 69.49 inches. This result is usually written 67.9 ± 1.59.

If the distribution is markedly unsymmetrical the above formula may not hold accurately and there are symmetrical distributions for which it does not hold exactly. But extreme accuracy in the matter of an index of dispersion is not necessary or desirable. The formula is generally used regardless of the form of the distribution.

Variance vs. Probable Deviation. The term variance, which has been defined as the mean squared deviation about the mean, is more convenient in many forms of statistical work than the probable deviation. In using the variance theory it is not necessary either to extract the square root of the mean squared deviation or to multiply the square root by the fraction 0.6745. It must be evident that comparative variances tell essentially the same story in regard to dispersion as do the comparative probable errors. About the only added information obtained from the probable error is the defining of a range within which half the frequencies lie. It is apparent that this idea is closely related to the median idea.

Standard Deviation of the Arithmetic Mean. The arithmetic mean in the Student Height data has been computed previously at 67.9 inches. The mean height of a second group of 750 students from the same student population would most likely not differ greatly from 67.9 but it is not at all likely that it would be exactly the same as that of the first group. Let group after group be taken and the value of the mean computed for each group. The values of these means would themselves form a frequency distribution from which a mean and standard deviation could be obtained.

Now if the student data is highly typical and stable the variation in the successive means will be within a small range and hence the standard deviation of the means will be relatively small. Let us assume the value which we have obtained by actual observation, namely 67.9, is the best estimate of the true mean of the height of all such students; that is, that the deviation of greatest frequency in the frequency distribution of means is 67.9. Then the standard deviation in this distribution will be the standard deviation of the mean. It can be proved that the *standard deviation of the mean is obtained from the standard deviation of the variates by dividing that standard deviation by the square root of the number of individuals or frequencies.*

In a formula, the standard deviation of the mean is $\frac{\sigma}{\sqrt{N}}$ and the variance of the mean is $\frac{1}{N}\sigma^2$

Likewise the P.E. of the mean is $0.6745 \frac{\sigma}{\sqrt{N}}$

The standard deviation of the arithmetic mean gives a measure of the reliability or significance of the arithmetic mean. It shows that the larger the number of frequencies the more we can rely upon the computed arithmetic mean because the variability of the means as measured by the standard deviation is decreased as N becomes larger. It shows also that the *reliability of the arithmetic mean is increased as the square root of the number of individuals,* frequencies, or observations increases. That is, the arithmetic mean for a distribution of 75,000 students from the same student population, as has been previously used, which is 100 times the frequency of the data already stated would be only ten times as accurate. Likewise, a frequency ten times as large would increase the accuracy only by the square root of ten which is somewhat more than three times. Multiplying the total number of variates by 1,000 increases the accuracy only slightly more than thirty times.

This fact, which has just been stated; namely, that increasing the number of observations increases the accuracy only by the square root of that number, has great practical significance.

It tends to show that after a reasonable number of observations have been made, or frequencies have been obtained, additional observations become of decreasing importance. In other words, it is logical to be content with data of reasonable volume.

Standard Deviation of the Standard Deviation. The standard deviation of the standard deviation may be explained by a process of reasoning similar to that for the standard deviation of the mean. The formula for this standard deviation is:

$$\text{Standard deviation of the standard deviation is } \frac{1}{\sqrt{2N}}\sigma$$

and

$$\text{The P.E. of standard deviation is } 0.6745\frac{\sigma}{\sqrt{2N}}$$

It appears from the preceding formula that the standard deviation of the standard deviation is equal to the standard deviation of the arithmetic mean divided by the square root of 2, which makes the standard deviation about 0.7 as variable as the arithmetic mean. Since the standard deviation has small variability, a difference in standard deviations is more likely to be significant than the same difference for the arithmetic means. *It should be apparent from these general comments on the standard deviation that the standard deviation deserves the important place which it occupies in statistical methods.*

Exercises

12. Compute the standard deviation of the standard deviation of the student heights of a preceding Chapter.

13. Compute the standard deviation of the standard deviation of the student weights of a preceding Chapter.

The Deciles as Measures of Dispersion. The position of the deciles shows the spread of the variates in the distribution. If the deciles near the middle of the distribution are close together and the deciles near the ends of the range are far apart the distribution is highly stable and true to type. Because there are nine decile positions to observe in a distribution the decile is not as simple a measure of dispersion as is the quartile or standard deviation, though this very fact of greater detail may in some cases be of advantage.

Exercise

14. Using the results of Exercise 14 of Chapter 4 examine the variability of the Chicago Monthly Top Beef Cattle prices as shown by deciles.

Symmetrical and Asymmetrical Distributions. The curve of students heights is essentially of the same shape to the right of the highest point as it is to the left. It is a symmetrical curve. Statistically the fact of symmetry means in this case that there is no tendency for students to be either tall or short; that there is no selection between the tall and short; that the chances for a tall person to belong to the student group are equally as good as those of a short person; that there is absolutely no connection between being a member of this student group and being tall or being short.

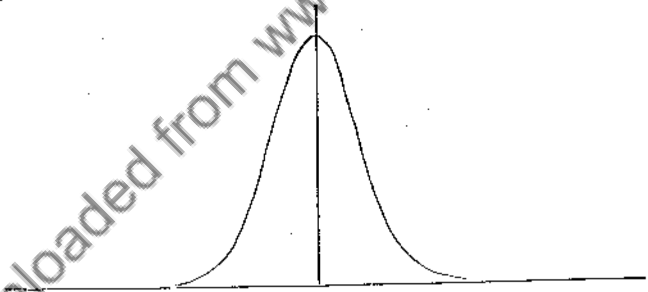


FIG. II. A Symmetrical Curve.

On the other hand, the curve of height of the members of a police force would have a longer range to the right than to the left because extremely short persons are excluded. The curve in this case is said to be asymmetrical. Asymmetry in a curve denotes the presence of selection in the data; of a dependence; of an expressed preference for certain values of the attribute.

The Position of the Averages and Asymmetry. In a symmetrical curve the mean, median and mode coincide. In an asymmetrical curve the mean, median and mode do not coincide.

The cutting off of the range to the left tends to move the mean to the right because the longer deviations are to the right, and it has been seen that the mean is most affected by the longer or extreme deviation. This places the median to the left of the mean. The mode will tend to be moved to the left of the median because both the effect of the moving of the mean to the right and of the shortening of the left range with a consequent heaping up of the frequencies within the left half. The result is that the three averages are then in the order—mode, median, mean. It has been verified experimentally that for moderately asymmetrical distributions the distance of the median from the mode is about one-third the distance of the mean from the mode.

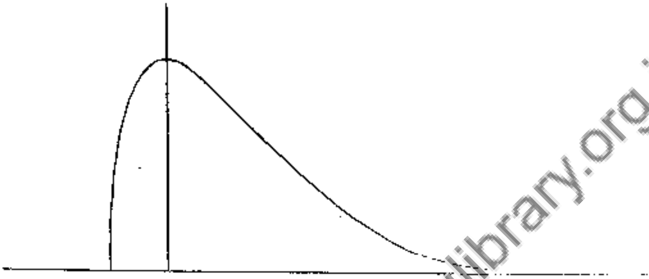


FIG. III. An Asymmetrical or Skew Curve.

Skewness. *An asymmetrical curve is said to be skew. Skewness is positive when the longer range is to the right and negative when the longer range is to the left.*

Measures of Skewness. Since the mode and mean are separated to an extent depending on the degree of skewness present, a logical measure of skewness is the difference between the mean and the mode. Because a large difference between the positions of the mean and the mode in widely spread-out data may not be as significant as a smaller difference in highly concentrated data it is advisable to divide this difference by the standard deviation. Hence we have,

$$Skewness = \frac{Mean - Mode}{\sigma}.$$

A Second Measure of Skewness is obtained as follows. Any measure of skewness must take into account the distinction between positive and negative deviations. The total sum of deviations from the mean is zero regardless of the form of the distribution. The standard deviation involves the deviations as squares and hence obliterates the distinction between positive and negative deviations. The mean cubed deviation, however, will serve as a measure of skewness. The longer deviations to the right, if the skewness is positive, will be more powerfully affected by the operation of cubing than will the shorter deviations to the left and hence the total sum of cubed deviations will be positive. It is well to extract the cube root of the mean cubed deviation and then in order to express the skewness as a fraction of the spread of the distribution to divide the result by the standard deviation. Further discussion of the subject of skewness is deferred to the Chapter on Moments.

Exercises

15. Using the second measure of skewness compute the skewness of the student height data already given.

16. Using the second measure of skewness compute the skewness of the student weight data already given.

CHAPTER VI

THE NORMAL PROBABILITY CURVE

The Equation of a Frequency Curve. As discussed in Chapter II, a smoothed curve is a graphic estimate of what would be the course of the data if it could be freed from accidental variations. *The smoothed curve is therefore the geometric representation of a law of connection or variation.* It shows, for instance, the variation of temperature with the seasons; the tendency for precipitation to depend on the month of the year; the most likely percentage of students at each height.

The presence of an underlying law of connection in the data implies the presence of an algebraic law connecting the x and the y coordinates. *The algebraic statement of the law giving the most probable value of y in terms of x is called the equation of the curve.*

If the equation is given, the ordinate can be computed for any abscissa and hence the curve can be located by plotting a sufficient number of computed points.

In some distributions it is possible to discover a law of connection directly from the data, and then without an extended computation to translate this law into the proper algebraic form. We shall discuss in this chapter the equation of one type of curve —the normal probability curve. This form of curve is suited to the representation of a large class of distributions. And the theory of the normal probability curve can be made use of in the determination of the standard and probable deviations and in the discussion of certain other properties even for a distribution to which it does not apply with sufficient accuracy to be adopted as the form of the smoothed curve.

Statistical Significance of the Normal Curve. The frequencies in data designated by a normal probability curve are merely the frequencies which result from the accidental variations in random sampling. They are the accidental variations which come purely from chance. For illustration of a normal probability curve assume that ten coins are placed in a cup and thoroughly shaken and then thrown so as to show heads or tails and a record made of the results. Out of 100 such throws, 5 heads and 5 tails will probably occur the greatest number of times. Six heads, four tails will occur not quite as many times. The probability will be still less for the occurrence of 3 heads and 7 tails and so on for the possible combinations. If a curve were plotted with x representing each number of heads and y the number of times each number of heads may occur it would be found that with $x = 5$, y would probably be the highest. For $x = 4$ and $x = 6$ the y's would be closely equal but somewhat shorter than the greatest value for y and so on for each value of x.

In this tossing of coins it is evident that in general there is no causal connection between the number of times a coin is thrown and the proportion of heads. In fact, a distribution of the type just considered results when the frequencies are merely the result of chance.

The foregoing may be restated as follows: *In a frequency distribution where there is no causal connection each frequency is the algebraic sum of an indefinitely great number of small elemental accidental influences which are all equal and each of which is as likely to be positive as negative.* This statement is not only logical but it is possible to derive the equation of the normal probability curve from this defining statement. One form of derivation is given in Appendix I. This equation is of the widest use throughout analytical statistical methods. Even where the data is not distributed normally many of the methods and formulas from the normal probability equation can be applied with sufficient accuracy for practical purposes.

The Equation of the Normal Curve. The equation of the normal probability curve is

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

where N is the total frequency of the distribution; σ, the standard deviation; π, the well known constant 3.14159; and e is a constant which is taken numerically equal to 2.71828. In this form of the equation x is measured from the arithmetic mean as origin. The derivation of this equation, as is shown in the Appendix, is based entirely on the foregoing statement that the ordinates of the normal probability curve are the resultants of a large number of elemental influences which are in themselves very small and which are equally likely to be positive or negative. It may be noted that in order to have a normal distribution it is not at all necessary that it be possible to compute actually the values of the elemental factors; it is only their existence under the above assumptions that is predicated.

The Curve of the Normal Equation. The mathematics of the normal probability curve, especially the derivation of the equation is fairly complex, but with the help of tables which have been prepared the normal probability curve is actually quite simple in its applications. The form of the equation may be somewhat intricate but the practical uses of the curve can be readily understood without going deeply into the higher mathematics of the matter.

The shape of the normal probability curve follows from general considerations. Since the equation of this curve is derived from the assumption that all the ordinates are the resultants of a large number of elements which are equally likely to be positive or negative it is logical to expect this curve to have the same shape on the positive side as on the negative side. Again, with an absence of selection, it is reasonable to conclude that the resulting distribution will be symmetrical with as large a distribution to the right as to the left. And it is also apparent, that the frequencies at the center will be high and those at the ends of the range small.

In the normal equation,

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

Since x appears in the equation only as a square, $-x$ gives the same value for y as does $+x$. This is the algebraic way of saying that the curve is symmetrical about the y axis.

It is to be noted that x appears in a negative exponent. It is a well known algebraic fact that a quantity with a negative exponent is really a fraction, so that a negative exponent in the numerator appears as a positive exponent when transferred to the denominator. It must accordingly be apparent that as x becomes larger the denominator of this fraction becomes larger and hence the fraction itself becomes smaller. That is, y becomes smaller as x becomes numerically larger. On the other hand, no matter how large x may be there is always, even though it be microscopic, some value for y. All of this means that this symmetrical curve sinks both to the right and to the left and continually approaches the axis but never actually reaches it. It will be seen in a moment that the points of the curve beyond some very definite limits are negligible.

To re-state in general language, there is always a possibility that the elements might be so grouped as to produce any value—there is a remote possibility that a coin tossed 100 times might yield heads each time. Since this possibility exists there must always be some ordinate, however small, for each value of x and hence the curve continually approaches the x axis as a limit but never becomes actually equal to zero.

The study of the normal probability curve is facilitated by writing the equation in the following form:

$$y = \frac{N}{\sigma} \cdot Z,$$

$$\text{Where } Z = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \frac{x^2}{\sigma^2}}$$

It is to be noted that the x in the expression for Z is divided by σ so that the independent variable is made up of the various values of x standardized by dividing by the standard deviation. In other words, Z is a general form which can be fitted to any normal distribution data. The value of the standard deviation, σ, determines the shape of the curve for any distribution. Where σ is large the result of dividing each x by σ will be relatively smaller than where σ is small. Consequently a curve with a large σ will be more spread out, that is, more flat than where σ is small

To summarize these comments on the normal probability curve it may be said that all normal probability curves are of the same general shape and have the same geometric and algebraic characteristics but differ from distribution to distribution by the degree of flatness which is measured by the standard deviation.

Where the data is highly typical the standard deviation will be small and the ordinates of the curve will decrease more rapidly so that the curve will be steep or relatively very high in the center. The logic of a flat or steep normal curve can also be supported by mathematical reasoning.

It must be apparent that if a table giving the values of Z for each ordinary value of $\frac{x}{\sigma}$ was at hand it would make possible the ready computation of values for y. In the following table, which is an illustrative excerpt from Sheppard's "Table for Statisticians and Biometricians", values of Z are presented.

Table of Ordinates and Areas of the Normal Curve

x/σ	Z	*Areas*	x/σ	Z	*Areas*
0.0	0.399	0.000	1.2	0.194	0.385
0.1	0.397	0.040	1.4	0.150	0.419
0.2	0.391	0.079	1.6	0.111	0.445
0.3	0.381	0.118	1.8	0.079	0.464
0.4	0.368	0.155	2.0	0.054	0.477
0.5	0.352	0.191	2.2	0.035	0.486
0.6	0.333	0.226	2.4	0.022	0.492
0.7	0.312	0.258	2.6	0.014	0.495
0.8	0.290	0.288	2.8	0.008	0.497
0.9	0.266	0.316	3.0	0.004	0.499
1.0	0.242	0.341	3.2	0.002	0.499

In computing the ordinates, each x is measured from the mean and divided by the standard deviation. Then the Table is entered with $\frac{x}{\sigma}$. Finally multiplication of the values so found by the ratio, N/σ, gives the successive values for y.

Computation of Areas Under the Curve. In the following illustrative computation of the normal curve of the distribution of student heights the ordinates are computed for the boundaries (as the fractional deviations in the first column denote) instead of the midpoints of the class intervals. This is done, as will be presently explained, for convenience in finding the areas under the curve. In the computation scheme, the first column is for the deviations; the second for the deviations from the mean; the third, the deviations from the mean divided by the standard deviation; the fourth, the values of Z obtained from the table; and the fifth column shows the desired values of the ordinates which are obtained by multiplying Z by N/σ. The sixth, seventh and eighth columns are explained on a following page.

Table of Z's and Corresponding Areas for Student Height

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
						Student Ht.	*Area*
Deviations	x	$\frac{x}{\sigma}$	Z	y	*Areas*	*Areas*	*Frequencies*
0.5	—7.4	—3.20	0.002	1.	.001	0.7	1.6
1.5	—6.4	—2.77	0.01	3.2	.003	2.3	5.2
2.5	—5.4	—2.34	0.03	9.7	.010	7.5	13.5
3.5	—4.4	—1.91	0.06	19.5	.028	21.0	32.3
4.5	—3.4	—1.47	0.14	45.5	.071	53.3	58.5
5.5	—2.4	—1.04	0.23	74.7	.149	111.8	91.5
6.5	—1.4	—0.61	0.33	107.1	.271	203.3	121.5
7.5	—0.4	—0.17	0.39	126.6	.433	324.8	126.7
7.9	—0.0	—0.00	0.40	127.2	.500	(375.0)	—
8.5	+0.6	+0.26	0.39	126.6	.602	451.5	107.3
9.5	+1.6	+0.69	0.31	100.7	.745	558.8	94.5
10.5	+2.6	+1.13	0.21	68.2	.871	653.3	52.5
11.5	+3.6	+1.56	0.12	39.0	.941	705.8	27.0
12.5	+4.5	+1.99	0.06	19.5	.977	732.8	12.0
13.5	+5.6	+2.43	0.02	6.5	.993	744.8	3.7
14.5	+6.6	+2.86	0.01	3.2	.998	748.5	0

$N = 750$

$\sigma = 2.36$

$N/\sigma = 3.18$

$\text{Mean} = 67.9$

Exercises

1. Plot a normal curve for the distribution of student weights according to the data of Chapter III.

2. Compare the curve obtained in Exercise 1 with the smooth curve of Chapter III. How closely do they agree?

Area Under the Normal Curve. It has been seen that the total area under a normal curve must equal the total frequency. In Sheppard's table the value of this area for limits of x are given. In the Table of page 68, a few of these values are given. It is to be noted that this Table is so arranged that the area starts from the mean and extends to the right or the left.

In the computation form for the ordinates of the student heights curve the areas from Sheppard's Table are set down in column (6) for the values of $\frac{x}{\sigma}$ of column (3). The ordinates through the boundaries of the classes are taken to facilitate the computation of the class areas. It is to be noted that these areas are so arranged that the area through the mean is one-half, and it is to be noted that the last area of the Table is not quite one. In column (7) the cumulative student height areas are given as obtained by multiplying each item of column (6) by 750. In the final column, (8), the areas are subtracted or added so as to show the frequencies in each class.

It should be noted that all the values for x in the table of student heights of page 69 are at the boundaries of the student height classes. The mean has been computed as 67.9 hence the middle class is 67.5 to 68.5 with 68 as the middle ordinate. This brings the computation in line with the frequency distribution at the beginning of Chapter III. Since areas are measured from the left the class one area is 1.6 or to the nearest whole number 2; the next is 5.2 or 5, and so on down one by one.

Goodness of Fit. The area frequencies of column (8) of the preceding Table are re-stated in terms of whole numbers in the following table in order to compare the computed frequencies with the original frequencies.

The Adjusted Distribution of Student Heights

Class	*Computed Frequencies*	*Original Frequencies*	*Positive Difference*	*Negative Difference*
1	2	2	..	..
2	5	10	..	5
3	14	11	3	..
4	32	38	..	6
5	59	57	2	..
6	91	93	..	2
7	122	106	16	..
8	127	126	1	..
9	107	109	..	2
10	95	87	8	..
11	53	75	..	22
12	27	23	4	..
13	12	9	3	..
14	4	4	..	..
	750	750	37	37

The goodness of fit of this normal curve is indicated by the differences of the fourth and fifth columns. The differences are taken positive when the adjusted values exceed the original frequencies. The sum of the positive and of the negative differences shows a fairly close fit, though the size of the individual differences must also be taken into account in estimating the closeness of fit.

Further Comments on Areas. It may be readily shown by the usual methods of determining areas under a curve that the mathematics of the normal probability curve calls for an area of unity between the curve and the X-axis.

The areas are from the central ordinate to the ordinate corresponding to the x/σ. It will be noted from the Z table that when x/σ is 3.2 practically the entire area of half of the curve has been included, the entire half area being 0.5 and the area here being shown as 0.499. It should be noted also that the x giving half of the area to one side is somewhat less than 0.7, 0.7 being 0.258 instead of 0.25,—it is 0.6745.

Exercises

3. Test the closeness of fit of the normal curve of student weights plotted in Exercise 1.

4. Compare the closeness of fit of the normal curves of weight and height.

5. The table of values of Z are multiplied by N/σ to give the actual ordinates while the areas are multiplied by N. Explain this difference (an elementary knowledge of calculus is required).

Preliminary Determination of Normality. Before attempting to fit a normal curve to a given distribution the data should be analyzed to determine whether the fundamental condition of normality is present, that is, whether the data is apparently subject only to accidental variations. The data should be plotted and the smooth curve drawn by the methods of Chapter II. Then if a normal distribution is indicated a normal curve should be fitted.

A mathematical measurement of normality will be derived in a later chapter.

Probable Deviation in a Normal Distribution. The quartiles divide the two halves of the area into equal parts. Hence, in the Z table the value of x/σ which corresponds to an area of 0.25, gives the value of the probable deviation. This value of x/σ is there found by interpolation to be equal to 0.6745. Therefore, the deviation of the quartile is 0.6745 times the standard deviation. This demonstrates the rule for obtaining the probable deviation, namely, multiplying the standard deviation by 0.6745.

The formulas for the probable deviation of the arithmetic mean and of the standard deviation referred to in the preceding Chapter are derived on the assumption that the two are each normally distributed.

It can be shown mathematically that even when the form of distribution is distinctly non-normal the ordinary rules for finding the probable deviations hold with an approximation close enough for practical purposes, and experimentation with different forms of distributions bears out the mathematical conclusions.

Exercises

6. What is the deviation corresponding to the ordinate which marks off three-fourths of the area to the right of the mean?

7. What part of the area under the normal curve is included between the mean and the ordinate with a deviation of two times the standard deviation? Three times the standard deviation? Four times the standard deviation?

The results of Exercise 7 show that the occurrence of a deviation of three times the standard deviation is highly improbable. That is, a deviation greater than about three times the standard deviation must significantly indicate that the measurement is not that of an individual taken from the same material; that it does not belong to the same distribution but to another distribution which has some conditions different from the first. To illustrate, the standard deviation of student heights is 2.36 inches and the mean height is 67.9 inches. One would, according to this theory, be justified in concluding that a person with a height of 76 inches $(67.9 + 3 \times 2.36 = 74.98)$ does not belong to the same student group.

While it is not advisable to place implicit confidence in the tests furnished by the theory of probable deviations to the extent that the results which it indicates are accepted without some independent verification, or at least justification, yet when used with judgment they are extremely valuable aids in practical statistical work. In every case it establishes cautionary limits, as, for instance, one would not ordinarily be justified in concluding that a variate with a deviation much greater than two or three times the standard deviation belonged to the same distribution. On the other hand, if a number of measurements of height should each consistently exceed those of the student distribution it might then be concluded with much certainty that the individuals measured were taken from a population distinctly different from the student population. And the conclusion would be justified even though the deviations were considerably less than two or three times the standard deviation.

Least squares. The equation of the normal probability curve furnishes a very convenient and easy way of seeing the reasonableness of the principle of *least squares*.

According to the principle of least squares the best fitting curve, or the best graduation of data, is one in which the sum of the squares of the differences between the original data and the graduated or fitted data, is a minimum. It must be noted that according to this principle the "best" is not based on having the sum of the differences between the observed and the adjusted values the least possible. This principle says that the *squares* of such differences must be the least possible.

It is obvious that one advantage of the least squares test of "goodness of fit" is that since each difference is squared no distinction is made between positive and negative values because the square of a negative value has a positive sign. It has been seen that the mean deviation, where all deviations are taken as positive, is a measure of dispersion. It might accordingly be possible to add all the mean dispersions, but one difficulty is that, though such a test is simple in words, mathematically the differences themselves regardless of sign are not readily amenable to analysis. On the other hand, squares of differences which naturally point toward the standard deviation are much easier to handle in computations.

To give an idea in outline of how the principle of least squares is true where the data follows the normal probability curve let us take a number of points $(x_1\ y_1)$ $(x_2\ y_2)$ $(x_3\ y_3)$ . . . . Then if the points are on the probability curve each set of values of x and y must satisfy the equation. Hence on substituting we have,

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

for each point. That is, we will have this expression with (x_1, y_1), (x_2, y_2), ... , respectively, substituted.

The foregoing values for y_1 . y_3 . y_3 . . . etc. give the respective probabilities that these points lie on the normal probability curve. Hence, the probability that all these points line on the curve at the same time is the probability of the occurrence of all these phenomena and this is the product of the individual probabilities. The product of the y's is an expression with the sum of the squares of the x's in the numerator of the negative exponent of e.

In passing over the various questions in regard to the constants let it be noted that the exponent of e involves the expression $(x_1^2 + x_2^2 + x_3^2 + \ldots\ldots)$. Since this latter expression occurs in a negative exponent, which makes the expression a fraction, the values of the *fraction* will be largest when the denominator is smallest, which, in this instance, will occur when the just stated sum of the squares of the x's is smallest. That is, the probability that all points are on the curve is greatest when the sum of the squares of the differences is a minimum.

CHAPTER VII

THE CORRELATION TABLE

The Correlation Table. From the records of physical measurements of students, of which the data at the beginning of Chapter III is a part, a tabulation was made of the heights of students whose weight was from 130 to 134 pounds—a weight class which may be denoted by the middle weight, 132 pounds—and the following distribution obtained:

Height	62	63	64	65	66	67	68	69	70	71	72	73	74
Number	2	0	4	9	18	18	17	8	8	4	3	1	1

The distributions were likewise obtained for each other five-pound intervals from 102 to 187 pounds. Instead of writing each of these distributions separately it is more convenient to write them together in one table called, for reasons explained later, a *correlation table.* In this way we have the following table:

Correlation Table of Height and Weight

Height in Inches

		61	62	63	64	65	66	67	68	69	70	71	72	73	74	To'ls
Weight in Pounds	187	.	.	.	..	..	..	..	1	3	2	2	1	.	.	9
	182	.	.	.	..	..	..	1	..	..	..	..	.	1	.	2
	177	.	.	.	..	..	..	..	..	..	1	1	..	.	1	3
	172	.	.	.	..	..	..	..	1	1	1	6	2	.	.	11
	167	.	.	.	..	..	..	..	1	2	6	1	2	1	.	13
	162	1	.	.	..	..	..	2	2	3	8	2	2	2	.	22
	157	.	.	.	..	..	4	1	6	7	5	7	1	.	.	31
	152	.	.	.	..	2	2	3	14	10	12	11	.	1	1	56
	147	.	.	.	..	2	3	7	5	12	9	8	3	.	.	49
	142	.	.	.	..	7	12	10	17	17	8	15	5	2	.	93
	137	.	1	.	3	4	14	20	24	21	11	9	2	.	1	110
	132	.	2	.	4	9	18	18	17	8	8	4	3	1	1	93
	127	1	1	1	7	7	11	15	16	18	9	5	2	.	.	93
	122	.	1	4	2	12	17	16	14	4	5	..	.	1	.	76
	117	.	2	2	10	9	6	6	7	2	2	2	.	.	.	48
	112	.	.	2	7	3	3	3	..	..	..	2	.	.	.	20
	107	.	3	1	5	2	1	1	..	..	..	..	.	.	.	13
	102	.	.	1	..	..	2	3	1	1	..	..	.	.	.	8
	Totals	2	10	11	38	57	93	106	126	109	87	75	23	9	4	750

The writing of the distribution in this compact tabular form greatly facilitates the study and comparison of the two characteristics or attributes.

It is to be noted that there is a decided increase in weight with an increase in height; that there are no extremely tall persons in the group who are at the same time extremely light in weight; that there are practically no persons who are both short and extremely heavy. It also appears that there is a closer connection between height and weight for the shorter and lighter individuals than for persons with medium values of the two characteristics.

Definitions and Symbols. The properties, as height and weight, are called the *attributes* or *characteristics.*

The horizontal deviations are called *the x classes or deviations,* and the vertical, *the y classes or deviations.* Each sub-class or sub-group thus has a value of x and of y associated with it. It is convenient to number the x and y classes from left to right and from bottom to top, respectively, and use these numbers for class numbers instead of the actual class values. Thus there are 17 persons with height 66 inches and weight 122 pounds. In terms of x and y, the sub-class $(x = 6, y = 5)$ has a frequency of 17; the sub-class $(x = 5, y = 6)$ has a frequency of 7.

The columns and rows are spoken of as *arrays;* the *columns as y arrays of type x* and the *rows as x-arrays of type y.* Or the specific names of the data may be given to the arrays—the weight array of height 67 inches; the height array of weight 132 pounds. It should be noted that the weight array of height type 67 inches is the distribution with respect to weight of the students having a height of 67 inches.

A y array of type x and an x array of type y are said to be *arrays of opposite sense.* Two y arrays or two x arrays are *arrays of the same sense.*

The frequency of a y array is denoted by the symbol n_x where x is the type of the array. The frequency of an x array is denoted by the symbol n_y, where y is the type. The frequency of a subclass is denoted by the symbol n_{xy}, where x and y are the deviations of the subclass, that is, the types of its two arrays. Thus, $n_{61:} = 2$; $n_{:132} = 93$; $n_{66:142} = 12$, or if the simpler class numbers are used, $n_{1:} = 2$; $n_{:7} = 93$; $n_{6:9} = 12$. When the latter form of class numbers is employed it is necessary to distinguish between x and y class numbers by means of a colon. Sometimes the distinction between x and y deviations or class numbers is made by the use of subscripts as $n_{x_1y_2}$.

Exercises

1. Write the values of $n_{3:4}$ $n_{9:2}$ for the height-weight data.

2. Practice stating the frequencies of the various arrays and subgroups; e. g. the frequency of the weight array of type 8 (68) is 126.

3. Note that $n_{1:7} + n_{2:7} + \ . \ . \ . \ n_{11:7} = n_{:7} = 93$, for the height-weight data.

4. Write other statements in the form of that of Exercise 3.

The mean of the vertical column of totals is called the *mean of all the weights,* and in general, *the mean of the y's,* and is denoted by the symbol $\bar{y}$. It is the mean weight for all heights.

Likewise the mean of all the x's is denoted by the symbol $\bar{x}$.

The means of the weight arrays are denoted by the symbols, $\bar{y}_{61}$, $\bar{y}_{62}$, $\bar{y}_{63}$.

In general the *mean of the y array of type x is denoted by the symbol* $\bar{y}_x$. The standard deviation of all the y's is denoted by σ_y and of all the x's by σ_x.

Exercise

5. From the following data construct the correlation table of Monthly Top Hog and Top Beef Cattle prices at Chicago.

Chicago Monthly Top Hog Prices

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1916	$8.10	$8.90	$10.10	$10.05	$10.35	$10.15	$10.25	$11.55	$11.60	$10.35	$10.35	$10.80
1915	7.40	7.25	7.05	7.90	7.95	7.95	8.12	8.05	8.50	8.95	7.75	7.10
1914	8.00	8.90	9.00	8.95	8.67	8.50	9.30	10.20	9.75	9.00	8.25	7.75
1913	7.80	8.70	9.62	9.70	8.85	9.00	9.62	9.40	9.65	9.10	8.30	8.15
1912	6.70	6.57	7.95	8.20	8.05	7.30	8.50	9.00	9.27	9.42	8.30	7.85
1911	8.30	7.90	7.35	6.90	6.50	6.72	7.55	7.95	7.80	6.90	6.72	6.60
1910	9.05	10.00	11.20	11.00	9.35	9.80	9.60	9.70	10.10	9.65	8.70	8.10
1909	6.70	6.95	7.15	7.60	7.55	8.20	8.45	8.32	8.60	8.40	8.45	8.75
1908	4.72	4.70	6.35	6.45	5.90	6.67	7.10	7.10	7.60	7.20	6.40	6.15
1907	7.05	7.25	7.10	6.90	6.65	6.42	6.65	6.72	7.00	7.00	6.32	5.30
1906	5.72	6.42	6.55	6.82	6.67	6.85	7.00	6.75	6.82	6.85	6.50	6.55
1905	5.00	5.12	5.55	5.72	5.65	5.70	6.17	6.45	6.20	5.80	5.25	5.35
1904	5.20	5.30	5.82	5.50	4.95	5.40	5.90	5.80	6.37	6.30	5.25	4.87
1903	7.10	7.65	7.87	7.65	7.15	6.45	6.10	6.20	6.45	6.50	5.50	4.90
1902	6.85	6.60	6.95	7.50	7.50	7.95	8.25	7.95	8.20	7.92	6.95	6.80
1901	5.47	5.65	6.20	6.25	6.05	6.30	6.40	6.75	7.37	7.10	6.30	6.90
1900	4.92	5.10	5.55	5.85	5.57	5.42	5.55	5.57	5.70	5.55	5.12	5.10
1899	4.05	4.05	4.00	4.15	4.05	4.09	4.70	5.00	4.90	4.90	4.35	4.45
1898	4.00	4.27	4.17	4.13	4.80	4.50	4.17	4.20	4.15	4.00	3.85	3.75
1897	3.60	3.75	4.25	4.25	4.05	3.65	4.00	4.55	4.65	4.40	3.80	3.60
1896	4.45	4.35	4.35	4.15	3.75	3.60	3.70	3.75	3.50	3.60	3.67	3.65
1895	4.80	4.65	5.30	5.42	4.97	5.10	5.70	5.40	4.65	4.50	3.85	3.75

Chicago Monthly Top Beef Cattle Prices

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1916	$9.85	$9.75	$10.05	$10.00	$10.90	$11.50	$11.30	$11.50	$11.50	$11.60	$12.40	$13.00
1915	9.70	9.50	9.15	8.90	9.65	9.95	10.40	10.50	10.50	10.60	10.55	11.60
1914	9.50	9.75	9.75	9.55	9.60	9.45	10.00	10.90	11.05	11.00	11.00	11.40
1913	9.50	9.25	9.30	9.25	9.10	9.20	9.20	9.25	9.50	9.75	9.85	10.25
1912	8.75	9.00	8.85	9.00	9.40	9.60	9.85	10.65	11.00	11.05	11.00	11.25
1911	7.10	7.05	7.35	7.10	6.50	6.75	7.35	8.20	8.25	9.00	9.25	9.35
1910	8.40	8.10	8.85	8.65	8.75	8.85	8.60	8.50	8.50	8.00	7.75	7.55
1909	7.50	7.15	7.40	7.15	7.30	7.50	7.65	8.00	8.50	9.10	9.25	9.50
1908	6.40	6.25	7.50	7.40	7.40	8.40	8.25	7.90	7.85	7.65	8.00	8.00
1907	7.30	7.25	6.90	6.75	6.50	7.10	7.50	7.60	7.35	7.45	7.25	6.35
1906	6.50	6.40	6.35	6.35	6.20	6.10	6.50	6.85	6.95	7.30	7.40	7.90
1905	6.35	6.45	6.35	7.00	6.85	6.35	6.25	6.50	6.50	6.40	6.75	7.00
1904	5.90	6.00	5.80	5.80	5.90	6.70	6.65	6.40	6.55	7.00	7.30	7.65
1903	6.85	6.15	5.75	5.80	5.65	5.15	5.65	6.10	6.15	6.00	5.85	6.00
1902	7.75	7.35	7.40	7.50	7.70	8.50	8.85	9.00	8.85	8.75	7.40	7.75
1901	6.15	6.00	6.25	6.00	6.10	6.55	6.40	6.40	6.60	6.90	7.25	8.00
1900	6.60	6.10	6.05	6.00	5.85	5.90	5.85	6.20	6.15	6.00	6.00	7.50
1899	6.30	6.25	5.90	5.85	5.75	5.75	6.00	6.65	6.90	7.00	7.15	8.25
1898	5.50	5.85	5.80	5.50	5.50	5.35	5.65	5.75	5.85	5.90	6.25	6.25
1897	5.50	5.40	5.65	5.50	5.45	5.30	5.25	5.50	6.00	5.40	6.00	5.65
1896	5.00	4.75	4.75	4.75	4.55	4.65	4.60	5.00	5.30	5.30	5.45	6.50
1895	5.80	5.80	6.60	6.60	6.40	6.00	6.00	6.00	6.00	5.60	5.00	5.50

Exercises

6. From the data of Exercise 5, construct the correlation table of **hog prices and months of the year,** and comment on its significance.

7. From data obtained from a financial journal **construct a correlation table of the** prices of common and preferred stocks.

8. In the correlation table of Exercise 5 does there appear **to be a sharp tendency** for the beef cattle arrays to vary with the changing **live hog prices?** **Is the** tendency more pronounced at some parts of the **table than at others?**

Center. The point of intersection of a horizontal line through the mean of all the *y's;* that is, $\bar{y}$, and the vertical line through the mean of all the *x's;* that is $\overline{x}$, is the *center* of the correlation table.

Frequency Surface. The frequency curve has been defined as a curve whose ordinates are proportional to the frequencies of the respective classes. The same idea may be applied to the correlation diagram. Let an ordinate be erected on each square proportional to, or equal to, the frequency of that sub-class. On the sub-class with height 68 and weight 137 the ordinate would be 24; on height 71 and weight 152, the ordinate would be 11, and so on. A smoothed surface through the ends of these ordinates would be the frequency surface for the correlation distribution. Over any point the height of the surface or the length of the ordinate would give the corresponding number of occurrences for the pair of measurements of the sub-class.

The distance of a point above the X-Y plane is denoted by the Z coordinate. On the surface a point accordingly has three coordinates X, Y, and Z.

It must be apparent from the definition of correlation that in correlated data the frequency surface has some definite shape which in most cases may be fairly simple. For uncorrelated data the surface is folded and crinkled without any order whatever or if it is smoothed the resulting surface will be a plane parallel to the X-Y plane.

Correlation. In the table of student heights and weights there is a decided tendency for heaviness and tallness to be associated and for lightness and shortness to be associated. There is likewise a pronounced tendency for the prices of live hogs and beef cattle to vary together. It is to be noted that the two series of measurements do not vary together in every case, that is, there are months in which the price of hogs is low but the price of beef cattle is high. But when all the months of an array are taken together there is evident a general tendency for an increase in beef cattle prices to be accompanied by an increase in hog prices. *Two characteristics are said to be correlated when there is a tendency for the changes in the value of one to depend on the changes in the value of the other.* The two characteristics may increase together or one may increase while the other decreases and even in a part of the table the movement of the changes may be together and in another part the two series of changes may move in opposition.

In *uncorrelated data* there is no tendency for the distributions of the arrays to change from type to type.

In perfectly *correlated data* there is an exact connection between the values of the two characteristics. If height and weight were perfectly correlated, for instance, all persons of a given height, say 68 inches, would be the same weight and hence all the frequencies of the weight array of type 68 would lie within a single sub-group. Between the two extremes of perfect and of no correlation there are all degrees of correlation.

Exercises

9. Study the degrees of correlation shown by the tables constructed in working the exercises of this Chapter.

CHAPTER VIII

THE CORRELATION RATIO

The Mean as Representative of the Array. In Chapter IV it was stated that the modal deviation is the most frequent deviation; that is, the most typical deviation of a distribution. Because the mode cannot be computed by a simple and uniform process of arithmetic the mean is a more practicable representative of the array. And this substitution of the mean for the mode will rarely produce a serious error.

Since the mean of the frequencies of an array is taken as the representative of the deviations of the array it is apparent, from the definition of correlation of Chapter VII, that the amount or degree of correlation in the data will be indicated by the variation in the means from array to array.

Regression Curves. The variation in the means of the arrays is shown graphically by the curve of means, which is called a *regression curve.*

Since there are two sets of arrays there are two regression curves.

Exercises

1. From the correlation diagram of Exercise 7 of the preceding Chapter compute and compare the mean of all the top beef cattle prices and the mean of all the hog prices and find the center of the table.

Correlation and the Regression Curves. In uncorrelated data the mean of an array does not depend on the type of the array; that is, does not change from array to array, and hence the unchanging value of the respective means of the arrays must be the same as the mean of all the *y's,* or at least this must be true when the regressive curve is smoothed.

The *regression curve for uncorrelated data* therefore approximates a straight line coinciding with the horizontal axis through the center. For correlated data the regression curve diverges or deviates from this position of coincidence with the axis. It must be noted that the shape of the regression curve is theoretically without effect on the degree of correlation present in the data. It is the variations in the distances of the means of the arrays from the axis that count in determining the degree of correlation present. Hence any numerical measure of the extent of correlation in the data must depend on the deviation of the means from the horizontal axis through the center.

Since there are two regression curves and two axes there are two correlations in each correlation table and their numerical measures involve the deviations of the respective regression curves from the corresponding straight lines through the center. Thus the dependence of height on weight and of weight on height are two distinct correlations.

Mean Squared Deviation of the Means of Arrays. The mean squared deviation is the most convenient measure of the deviations of the means of the arrays. In computing this measure the means of the arrays expressed in classes are first computed and written in a vertical column and then the difference between each mean and the mean of all the variates is set down in a column. Because the differences are used only in the squared form it is not necessary to retain a negative sign.

The next column in the computations on page 84, contains the squares of the differences. Since the means of the array are used as the representatives of the individuals of the respective arrays each of these individuals is possessed of the squared deviations. Hence each square must be multiplied by the respective frequencies of the corresponding arrays. The resultant products form the final column. The sum of this last column is the total sum of squared deviations and this sum divided by the total frequency is the mean squared deviation of the means of the arrays.

The Correlation Ratio. The mean squared deviation just obtained would be a significant measure of correlation were it not for the fact that it does not take into account the dispersion of the data as a whole. Without changing the mean and the total frequency of even one *y*-array, it would be possible to spread out each array to twice its length. This alteration would leave unchanged the mean squared deviation of the means of the arrays from the horizontal axis. It is evident that the value of the mean squared deviation of the means of the arrays is of less significance in the more spread-out data. Hence the dispersion of the data as a whole must be considered in interpreting the value of the mean squared deviation.

The dispersion of the data as a whole is given by the standard deviation of the frequencies of the totals in the vertical sum column. The smaller this mean squared deviation the more significant is the deviation of the means, and the larger this standard deviation the less significant is the deviation of the means. It is therefore reasonable to divide the square root of the mean squared deviation of the means of the arrays by the standard deviation obtaine dfrom the marginal column. The quotient is called the *correlation ratio,* and is denoted by the Greek letter η.

The computation of the correlation ratio for the dependence of student weight on height follows from the computation of the mean squared deviation of the means of the array.

The means and the one standard deviation were computed in the usual manner. We have, for the data as a whole, $\bar{y} = 7.9$, and $\sigma^2_y = 9.79$.

Computation of the Correlation Ratio
Student Heights and Weights

(1)	(2)	(3)	(4)	(5)
n_x	$\bar{y}_x$	$\bar{y} - \bar{y}_x$	$(\bar{y} - \bar{y}_x)^2$	$n_x(\bar{y} - \bar{y}_x)^2$
2	9.5	1.6	2.56	5.12
10	4.7	3.2	10.24	102.40
11	3.9	4.0	16.00	176.00
38	4.6	3.3	10.89	413.82
57	6.1	1.8	3.24	184.68
93	6.8	1.1	1.21	112.53
106	7.0	0.9	0.81	85.86
126	8.0	0.1	0.01	1.26
109	8.8	0.9	0.81	88.29
87	9.7	1.8	3.24	281.88
75	9.9	2.0	4.00	300.00
23	10.3	2.4	5.76	132.48
9	10.8	2.9	8.41	75.69
4	10.5	2.6	6.76	27.04
750				1987.05

Hence we have from the definition of the correlation ratio,

$$\text{Mean squared deviation} = \frac{1987.05}{750} = 2.6494,$$

$$\eta^2 = \frac{1987.05}{750 \times 9.79} = 0.27062,$$

$$\eta = 0.520.$$

Exercises

2. Compute the value of η for the correlation of Chicago Monthly Top Hog prices with Chicago Monthly Top Beef Cattle prices as shown in the table of Exercise 5 of the preceding chapter.

Two Values for η in Each Table. From the method of computation it is clear that there are two values for the correlation ratio, η, in each correlation table, one for each regression curve. The correlation ratio of weight with height, for instance, may differ considerably from the correlation ratio of height with weight; the dependence of precipitation on temperature may be of a decidedly different degree from that of temperature on precipitation. The two values of η do not ordinarily differ markedly but there can be no apriori assurance that they will be essentially of equal value and hence it is necessary to compute the two values separately in case both are desired. To distinguish between the two measures the symbol η_y is used for the dependence of y on x, and the symbol η_x refers to the dependence of x on y.

Exercises

3. Compute the correlation ratio of height with weight and compare with the value of η in Exercise 2.

4. Compute the value of η from the live stock price table of Exercise 5, Chapter VII, for beef cattle prices with hog prices, and compare this value of η with that of Exercise 2.

Limiting Values of the Correlation Ratio. In theory, the means of the arrays lie exactly in the axis for data of zero correlation. Each separate item, therefore, in the mean squared deviation of the means is zero and hence η is zero for absolutely uncorrelated data.

Because each term of the mean squared deviation of the means is squared and hence necessarily positive any accidental fluctuations of the means of the arrays in data of essentially zero correlation increase the value of η. Since there are no compensating fluctuations, the result is that small values of η are likely to be too large and hence the statistical significance of η for data of a small degree of correlation is open to question. The degree of correlation in such cases cannot be greater than the value of η would indicate and it may be less. It must be evident from the nature of the error that for material showing a considerable degree of correlation the error from this source is negligible.

Again, an inspection of the method of computing the correlation arrays will show that for perfectly correlated data this computation is precisely the same as for the computation of the standard deviation. Hence for perfect correlation

$$\eta^2 = \frac{N\sigma^2}{N\sigma^2} = 1 .$$

Exercises

5. Compare the values of η that have been computed with the general appearance of correlation in the tables.

6. Can a tendency be detected for the two values of η to be closer together in value for highly correlated data than for data of smaller correlation?

Probable Deviation of the Correlation Ratio. It can be proved that the probable deviation of a correlation ratio is given by the formula

$$P.E._{\eta} = 0.6745 \frac{(1 - \eta^2)}{\sqrt{N}}$$

This probable error formula supports the previous statement that η is increasingly reliable as its value becomes closer to unity.

Exercises

7. Compute the probable deviations of the correlations ratios of this Chapter.

In working with correlations, especially where the total frequencies are not large, it is always well to obtain a considerable number of distributions. Then if there proves to be a consistency in the value of η greater confidence can be placed in those values than if there was only one distribution. Thus if ten groups of 750 students were each measured for height and weight and the computed values for η should show a decided tendency to agree in value, increased significance could be given to the values of η.

Spurious Correlation. In interpreting the computed value of any measure of correlation care must be taken that the correlation is not merely apparent and a result of the nature of the data. The *illusion* of correlation may be a result of some general change which may affect the attributes alike and to an extent which tends to obscure the more detailed inter-effects among the attributes. The mathematical computations will develop evidence of correlation but the point is that in such cases the significance of the correlation is in question.

Hog and cattle prices are affected alike by the general levels of prices and hence any computation is likely to show some degree of correlation. The question in such cases is how much dependence to place in a computed measure of correlation as an indication of what relationship to expect when price levels remain fairly constant.

Exercises

8. In which of the correlations of this Chapter is there a possibility of spurious correlation?

9. Show that in correlating index numbers especial care is necessary in interpreting values of η.

10. Show that where there is an element of spurious correlation present the correlation is real in so far as the measurements themselves are concerned.

CHAPTER IX

THE COEFFICIENT OF CORRELATION

A Much Used Measure of Correlation. In practical statistical work a much used measure of correlation is the correlation coefficient. This index is nothing more nor less than the correlation ratio where the means of the arrays lie on a straight line. The assumption of a straight line for the means can safely be made in much correlation analysis even though the means do not lie strictly on a straight line.

Here is another instance where some theoretical accuracy can safely be sacrificed to obtain the advantages of better mathematical adaptability.

The correlation ratio has a simple and logical basis as a measure of correlation and should be well understood before proceeding to the study of the correlation coefficient. The idea of regression curve was developed in the preceding chapter. The regression lines are the basic ideas for the correlation coefficient.

Linear Regression. A straight line fitted to the means of the arrays is called a *line of regression.* A line of regression smooths the curve of regression. Whenever a curve of means approximates a straight line the regression is said to be *sensibly linear.* If the regression curve, within the limits of accuracy of the data, is exactly a straight line the regression is said to be *truly linear.*

The slope of a regression line shows the broad general tendencies of the connection between the attributes. Does weight tend to increase as height increases? Does the monthly precipitation increase with an increase of temperature? If so, at about what rate? These are questions which can be answered by observing the slopes of the regression curves. It may happen that in some correlation tables the regression curves deviate so widely from straight lines that the regression lines have but little significance.

Exercises

1. Draw by inspection the regression lines on the correlation table of student heights and weights.

2. In Exercise 1 estimate the comparative degrees of correlation shown by the two regression lines.

The Equations of the Lines of Regression. Let the coordinate axes be the two lines through the center determined by the means of all the variates as described on page 78. Then x and $\bar{y}_x$ are the coordinates of a point on the one regression line and $\bar{x}_y$ and y on the other. It must be understood, however, that the values of $\bar{y}_x$ and $\bar{x}_y$ here referred to are the adjusted or fitted means of the arrays so that unless the regressions are truly linear these values will differ from the values obtained by actual computation.

It is demonstrated in Chapter XII that the equation of the regression line of the means of the y arrays is

$$\bar{y}_x = r\frac{\sigma_y}{\sigma_x}x.$$

And of the means of the x arrays,

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

where σ_y is the right hand marginal standard deviation and σ_x the bottom marginal standard deviation. The constant r is defined by the equation, $r = \dfrac{\Sigma\Sigma n_{xy}.\, x.y}{N\sigma_x \,.\, \sigma_y}$. The expression $\Sigma\Sigma n_{xy}\,.\,xy$ is a symbolic way of saying: the sum obtained by multiplying the frequency of each subclass by its deviation from the horizontal axis and then by its deviation from the vertical axis and then obtaining the sum of all such products.

According to the first of the two regression equations the mean weight for height 71 is obtained by substituting the value for x measured from the mean and multiplying and dividing as the formula directs. We found that for this data of student measurements $\sigma_y = 3.13$ and $\sigma_x = 2.36$. The value of r is found presently to be 0.50. The array is distant from the mean 3.1. Hence, $\bar{y}_{71} = 0.50. \frac{3.13}{2.36} . 3.1 = 2.06$ weight classes from the mean weight.

The Coefficient of Correlation. Let us now compute the correlation ratio using, however, in case the regression is not truly linear, not the actual means of the array but the means given by the regression line. The deviation of a mean from the horizontal axis has just been found to be $r. \frac{\sigma_y}{\sigma_x} . x$. The square of this quantity multiplied by the frequency of the array is $n_x . x^2 . \left\{\frac{r^2 \sigma_y^2}{\sigma_x^2}\right\}$. The last factor is the same for each array and the sum of the other factors leads to the standard deviation of all the x's. Hence we have, on carrying out the multiplications for each array and adding, $r^2 \frac{\sigma_y^2}{\sigma_x^2} . \Sigma n_x . x^2 = \frac{r^2\sigma_y^2}{\sigma_x^2} . N\sigma_x^2 = Nr^2\sigma_y^2$.

Therefore, the mean squared deviation of the regression means of the arrays is $\frac{Nr^2\sigma_y^2}{N} = r^2\sigma_y^2$. On dividing this mean squared deviation by the square of the standard deviation of all the y's we have $\frac{r^2\sigma_y^2}{\sigma_y^2} = r^2$. That is, the constant r reveals itself as the correlation ratio when the regression means are used instead of the true means. It is called the *coefficient of correlation.*

Other Definitions. The foregoing may be summarized into certain definitions. A *regression line* is a line passing near each of the points on a scatter diagram. The $\bar{y}_x$ of such a line is the average value for the given x.

The *regression coefficient* measures the average increase of y per unit increase of x. It is the b in the equation $y = bx$ and is equal to $r \,.\, \dfrac{\sigma_y}{\sigma_x}$.

It has been seen that the variance of y is σ_y^2 and the variance of x is σ_x^2. The *cross moment*, $1/N \,.\, \Sigma xy$, which is $r \,.\, \sigma_x\sigma_y$, is called *covariance*.

Computation of r. For computation purposes the summation $\Sigma_x\Sigma_y n_{xy} yx$ can be arranged in the following manner. Let the subgroup frequencies of a given y array be each multiplied by the respective deviations, all deviations being measured from the axis through the center, and the products summed. Divided by the frequency of the array this sum gives the mean $\bar{y}_x$. Hence the summation for the array is equal to the product of the mean $\bar{y}_x$ and the frequency n_x. On making this substitution the original summation formula becomes $\Sigma n_x \,.\, \bar{y}_x \,.\, x$, or $\Sigma n_x(\bar{y}_x - \bar{y})\ (x - \overline{x})$ from the original axis.

In the course of the computation of the correlation ratio the means $\bar{y}_x$ are obtained and hence to the computation schedule of page 84 only the additional column for the x deviation of each array is needed. Then the multiplication of the corresponding values from the n_x, $(\bar{y}_x - \bar{y})$, and $(x - \overline{x})$ columns gives the column which sums into the quantity $\Sigma n_x\ (\bar{y}_x - \bar{y})\ (x - \overline{x})$. This sum divided by the product of the three factors N, σ_x and σ_y gives the required value for r.

The following table shows the computation of the coefficient of correlation for student heights and weights.

The Coefficient of Correlation for Student Heights and Weights

Computation of r

(1) x	(2) n_x	(3) $\bar{y}_x$	(4) $\bar{y}_x - \bar{y}$	(5) $x - \bar{x}$	(6) $n_x(x - \bar{x})$	(7) $n_x(x - \bar{x})(\bar{y}_x - \bar{y})$
1	2	9.5	+1.6	—6.9	— 13.8	— 22.08
2	10	4.7	—3.2	—5.9	— 59.0	+188.80
3	11	3.9	—4.0	—4.9	— 53.9	+215.60
4	38	4.6	—3.3	—3.9	—148.2	+489.06
5	57	6.1	—1.8	—2.9	—165.3	+297.54
6	93	6.8	—1.1	—1.9	—176.7	+194.37
7	106	7.0	—0.9	—0.9	— 95.4	+ 85.86
8	126	8.0	+0.1	+0.1	+ 12.6	+ 1.26
9	109	8.8	+0.9	+1.1	+119.9	+107.91
10	87	9.7	+1.8	+2.1	+182.7	+328.86
11	75	9.9	+2.0	+3.1	+232.5	+465.00
12	23	10.3	+2.4	+4.1	+ 94.3	+226.32
13	9	10.8	+2.9	+5.1	+ 45.9	+133.11
14	4	10.5	+2.6	+6.1	+ 24.4	+ 63.44

$$\Sigma n_x\ (x - \bar{x})\ \ (\bar{y}_x - \bar{y}) = 2775.05$$

$$\bar{x} = 7.9;\ \bar{y} = 7.9,$$
$$\sigma_x = 2.36;\ \sigma_y = 3.13,$$
$$N = 750,$$

$$r = \frac{\Sigma n_x\ (\bar{y}_x - \bar{y})\ \ (x - \bar{x})}{N\sigma_x\ \sigma_y} = \frac{2775.05}{750 \times 2.36 \times 313}$$

$$= 0.50$$

Exercises

3. Compute r for the Chicago Monthly Top Hog and Top Beef Cattle data.

4. Compare the values of r in Exercise 3 and in the height-weight data with the corresponding values for η.

5. Does there seem to be a tendency for η and r to agree more closely for highly correlated data than for material of small correlation?

6. Compare the amount of labor involved in the computation of η with that involved in the computation of r.

Statistical Properties of the Coefficient of Correlation. Unlike the correlation ratio the coefficient of correlation expresses a property of the correlation table as a whole and not merely of one or the other of the two correlations of the table.

Again, unlike the correlation ratio, a negative sign for r has a significance. It indicates that the regression line has a negative slope and hence that the connection between the attributes is inverse; that is, one attribute increases while the other decreases.

Because both positive and negative values of r can occur there is no tendency, as there is in the case of η, for small values of r to be larger than the actual degree of correlation would warrant.

Because the coefficient of correlation is based on the regression lines some data may have regression curves which deviate so much from a straight line that computed values for r have little significance. In periodic data exhibiting a sine curve form for the regression curve the correlation may be high but the departure of the regression from linearity is so wide the value of r understates the correlation and hence its applicability in such data is not of significant importance.

A characteristic importance of the coefficient r is in determining the slopes of the regression lines. It furnishes the most convenient method for defining the general tendencies in the data. The rise of prices, for instance, during the last fifteen years can be readily measured by the rate of rise of the regression line.

It is to be noted that $r = \dfrac{\Sigma xy}{N\sigma_x\sigma_y}$ has two factors, an x and a y, in each term of both the numerator and the denominator and hence is unchanged by multiplying each x or y by the same factor. *That is, r is independent of the unit of measurement of the class intervals.*

Since both x and y are measured from the respective means it is likely that both x and y will be consistently positive or negative at the same time for correlated data clustering along a regression line. That is, the products $x\,y$ will be consistently positive or negative for correlated data. Where the data is not closely correlated some of the products will be positive and some negative. In other words, the sum of the products Σxy, will be numerically large for highly correlated data and small for uncorrelated data.

Limiting Values for r. It may be shown mathematically that for perfectly correlated data $r = 1$. In perfectly correlated data, $y = \frac{r\sigma_y}{\sigma_x} x$, gives an exact value for y for each x, say $y = Kx$. Then $\Sigma xy = \Sigma x \cdot Kx = K\Sigma x^2 = KN\sigma_x^2$. Also, for perfect correlation $\sigma_y^2 = \frac{1}{N} \Sigma K^2 \cdot x^2 = K^2 \cdot \sigma_x^2$. Hence $r^2 = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{KN\sigma_x^3}{KN\sigma_x^2} = 1$. That is, r reduces to unity for perfectly correlated data.

The two coefficients of regression from the regression lines $y = b_{yx} \cdot x$ and $x = b_{xy} \cdot y$ bear an interesting relation to each other through r. For $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$ and hence $b_{yx} \cdot b_{xy} = \frac{r\sigma_y}{\sigma_x} \cdot \frac{r\sigma_x}{\sigma_y} = r^2$. That is, r is the geometric mean of the two regression coefficients.

Reasoning from the relation of r to η, we see that for truly linear regression perfect correlation leads to a value of r equal to unity. The unity value for r will be positive or negative according as the correlation is direct or inverse. According to the underlying theory of the coefficient of correlation for data in which a regression is not linear the value of r cannot be unity even though there is perfect correlation and hence for non-linear regression r is necessarily smaller in value than the degree of correlation would require.

In data of zero correlation it is clear that the regression line coincides with the axis and hence the value of r must be zero.

Test for Linearity of Regression. It would be suspected from the preceding theory and discussion that the difference between η and r should be an indicator of the departure of the regression from linearity. A somewhat more convenient measure of this departure than the simple difference is the difference of the squares of η and r.

Probable Deviations. The following probable deviations can be derived:

$$P.E. \text{ of } r = 0.6745 \frac{(1-r^2)}{\sqrt{N}}$$

$$P.E. \text{ of } (\eta^2 - r^2) = \frac{1.35}{\sqrt{N}} \sqrt{\eta^2 - r^2}$$

Exercises

7. Compute the regression equations for each of the correlation tables of Chapter VII.

8. How can the value of r be obtained graphically from the regression lines? Is this a practicable method of finding the value of r?

9. Compute the measure of departure from linearity, $(\eta^2 - r^2)$, for the correlation tables of Chapter VII.

10. A correlation table has two measures of departure from linearity. Show that one regression may be linear and the other non-linear.

11. Show that if the value of r is high the regressions must both be approximately linear.

CHAPTER X

CORRELATION FROM RANKS

Rank in a Series. Where the data consists of order or rank in a series in respect to the characteristics there is a method of determining correlation from such ranks. Let us define *rank* as position in a series so that an individual of rank *one* would have no individuals above or before it; an individual of rank *two* would have one individual before it, etc.

To pass from rank to variate correlation, that is, the types of correlation already described it is necessary to know the form of distribution of the values of the characteristics. Only for normal distribution has the requisite theory been developed. It is consequently necessary to employ the same formulas for other forms of distributions, although this may open the way to inaccuracies.

Let the ranks of the same individual in regard to the respective characteristics be ν_x and ν_y. Let there be N individuals and let $\overline{\nu_x}$ and $\overline{\nu_y}$ denote the respective means of the two series and σ_{ν_x} and σ_{ν_y} the standard deviations.

Also let all the measurements of each characteristic be distinct in value; that is, let there be no equal measurements.

Theorem I. The mean ranks $\overline{\nu_x}$ *and* $\overline{\nu_y}$ *are equal to* $(N+1)/2$.

Since there are as many ranks as individual measurements and since the ranks proceed uniformly from 1 to N the mean is $(N+1)/2$.

Theorem II. The standard deviation of the ranks are each equal to $\frac{1}{12}(N^2-1)$.

$$\text{For } N\sigma_{\nu_x}{}^2 = \Sigma(\nu_x - \overline{\nu_x})^2$$

$$= \Sigma\nu_x{}^2 - 2\overline{\nu_x} \,.\, \Sigma\nu_x + \Sigma\overline{\nu_x}^2,$$

$$= \frac{1}{6}N(N+1)(2N+1) - 2\overline{\nu_x} \,.\, \frac{N(N+1)}{2} + N\overline{\nu}_x{}^2,$$

from applying the formulas $\Sigma N^2 = 1/6N\,(N+1)\,(2N+1)$ and $\Sigma N = 1/2N\,(N+1)$. Substituting further we have

$$N\sigma^2_{\nu_x} = 1/6N\ \ (N+1)\,(2N+1) - \frac{N(N+1)^2}{2} + \frac{N\,(N+1)^2}{4},$$

and on reducing,
$$= \frac{1}{12}\,(N^3 - N).$$

Therefore
$$\sigma^2_{\nu_x} = \frac{1}{12}\,(N^2 - 1).$$

The following theorem is necessary for the computation of rank correlation.

Theorem III. If $\sigma_x = \sigma_y$, $r = 1 - \dfrac{\sigma^2{}_{(x-y)}}{2\sigma_x{}^2}$.

For, $(x-y)^2 = x^2 + y^2 - 2xy$, and $\Sigma(x-y)^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$, or $N\sigma^2{}_{(x-y)} = N\sigma_x{}^2 + N\sigma_y{}^2 - 2\Sigma xy$.

But $\Sigma xy = r\ .\ N\ .\ \sigma_x\ .\ \sigma_y$.

Therefore, $N\sigma^2{}_{(x-y)} = N\sigma_x^2 + N\sigma_y^2 - 2\ \ Nr\sigma_x\sigma_y$,

$$\text{and } r = \frac{\sigma_x{}^2 + \sigma_y{}^2 - \sigma^2{}_{(x-y)}}{2\sigma_x\sigma_y}$$

$$\text{If } \sigma_x = \sigma_y,\ r = 1 - \frac{\sigma^2{}_{(x-y)}}{2\sigma_x{}^2}.$$

where $\sigma^2{}_{(x-y)}$ is the mean squared deviation of the difference between x and y.

Exercises

1. Show how to compute the value of r from the data of student height and weight by the formula $r = \dfrac{\sigma_x{}^2 + \sigma_y{}^2 - \sigma^2{}_{(x-y)}}{2\sigma_x\sigma_y}$.

Theorem IV. The correlation coefficient of the ranks ν_x and ν_y is given by the formula,

$$r\,\nu_x\nu_y = 1 - \frac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2-1).}$$

On making use of Theorem III, we have,

$$r\,\nu_x\nu_y = 1 - \frac{\sigma^2\,(\nu_x - \nu_y)}{2\sigma^2_{\nu_x}}$$

$$= 1 - \frac{\Sigma(\nu_x - \nu_y)^2}{2N\sigma_{\nu_x}{}^2}$$

$$= 1 - \frac{\Sigma(\nu_x - \nu_y)^2}{2\,.\,\frac{1}{12}\,.\,N(N^2-1)}$$

$$= 1 - \frac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2-1)}$$

To illustrate the method let us compute the rank correlation between yearly mean temperature and yearly mean rainfall for Ohio from the data arranged in ranks.

The order of the twenty-four years in respect to temperature is written in the first column and in respect to rainfall in the second. The ties are disposed of by assigning the ranks in the inverse order of the time, thus with 1903 and 1902 each at 50.5 in the full data, 1903 is given rank 15 and 1902, 16. The third column contains for each year the differences in rank with respect to the two attributes, temperature and rainfall, and the fourth the squared differences. On adding the fourth column and applying the formula $r = 1 - \dfrac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2-1)}$ we find $r = 0.07$.

Computation of Correlation from Ranks

Year	(1) *Temp.*	(2) *Rainfall*	(3) *Difference*	(4) *Sq. Diff.*
1911	1	4	3	9
1910	17	6	11	121
1909	13	5	8	64
1908	5	19	14	196
1907	22	3	19	361
1906	9	14	5	25
1905	20	10	10	100
1904	24	17	7	49
1903	15	16	1	1
1902	16	13	3	9
1901	18	24	6	36
1900	2	21	19	361
1899	11	18	7	49
1898	6	2	4	16
1897	14	11	3	9
1896	7	9	2	4
1895	21	23	2	4
1894	3	22	19	361
1893	10	7	3	9
1892	19	14	5	25
1891	8	12	4	16
1890	4	1	3	9
1889	11	20	9	81
1888	23	8	15	225

$N = 24$ $\qquad\qquad \Sigma(\nu_x - \nu_y)^2 = 2{,}140$

$N(N^2 - 1) = 13{,}800$ $\qquad 6\Sigma(\nu_x - \nu_y)^2 = 12{,}840$

$$r = 1 - \frac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2 - 1)} = 1 - 0.93 = 0.07$$

Ties in Rank. The application of the formula

$$r_{\nu_x \nu_y} = 1 - \frac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2 - 1)}$$ is straightforward and direct. The only uncertainty arises from ties in the measurements. Thus in the preceding illustrative example it was found from the data that the temperature for each of the two years 1907 and 1894 was 52.3. What ranks are to be assigned to each of the measurements? In order to avoid complicating details in an illustrative

problem, in the preceding computation we gave the latter year the numerically smaller rank, but ordinarily it is better to base the ranks on either of the two plans:

(1) *The Bracket Rank Method,* under which the ties are assigned the same rank and that equal rank is taken as the rank next greater than that of the individual preceding the ties. The next individual after the ties takes the same rank as if preceding ties had each been given ranks differing by unity. Thus under this method the ranks of the illustrative example are as given in the table below.

(2) *The Mid-Rank Method,* under which all ties are given the same rank but that rank is the rank of the mid-individual. In the column below the two methods may be compared.

Under either method the total number of ranks must be the same and equal to N.

Rank of Ties

	Temperature	*Bracket Method*	*Mid-Rank Method*
1911	52.6	1	1
1900	52.3	2	3
1894	52.3	2	3
1890	52.3	2	3
1908	52.1	5	5.5
1898	52.1	5	5.5
1892	51.7	7	7.5
1891	51.7	7	7.5
1906	51.6	9	9.5
1893	51.6	9	9.5
1899	51.5	11	11
1889	51.1	12	12
1909	50.9	13	13
1897	50.6	14	14
1903	50.5	15	15.5
1902	50.5	15	15.5
1910	50.4	17	17
1901	50.2	18	18
1892	50.1	19	19
1905	50.0	20	20
1895	49.9	21	21
1907	49.6	22	22
1888	49.5	23	23
1904	48.6	24	24

Exercises

2. Compute $r\,\nu_x\nu_y$ from the above "bracket method" ranks.
3. Compute $r\,\nu_x\nu_y$ from the above "mid-rank-method" ranks.

Standard Deviation of the Rank Coefficient. The Standard Deviation of r when computed from ranks is in accordance with the formula $\sigma^2{}_{\nu_x\nu_y} = \dfrac{1}{\sqrt{N}}\,(1 - r^2{}_{\nu_x\nu_y})$.

Perfect Rank Correlation. Ranks are perfectly correlated, according to the formula, when $\Sigma(\nu_x - \nu_y)^2 = o$; that is, when each individual has the same rank in both series. Also there is perfect negative correlation when temperature and rainfall are inversely related so that the year with the highest temperature is the year with the lowest rainfall and so on up to the year with the lowest temperature which is associated with the highest rainfall.

Uncorrelated Data. According to the formula, the sum of the squares of the differences of the ranks is equal to the sum of the squares of the ranks when $r = o$. Thus when $r = o$ subtracting the ranks has lost its significance—and this is exactly the idea of zero correlation.

Hence the rank coefficient, r, is accurately significant for both perfect and zero correlation.

A Correction Formula for the Rank Coefficient. There is no assurance however that in general the rank r will exactly express the true variate correlation. For instance, note the two following series of deviations.

100, 80, 70, 65, 62, 60, 55, 50, 40, 20; and 100, 99, 98, 97, 96, 95, 10, 9, 8, 7.

The ranks are the same in each series, namely,

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

The coefficient $r\,\nu_x\nu_y$ which depends solely on the ranks, has the same value for a series of which the first is typical as it does for a series of which the second is typical. And yet the two distributions are fundamentally distinct in form.

Therefore, except for the two extreme cases of data of very high and of very low correlation, the value of a correlation coefficient computed from ranks must be interpreted with caution.

For a distribution which is approximately normal in form the following correction formula for r has been derived by Pearson: $r_{xy} = 2 \sin \frac{\pi}{6} \cdot r_{\nu_x \nu_y}$.

From the Table below the values of r_{xy} can be obtained directly from the value of $r_{\nu_x \nu_y}$ for each 0.05 of $r_{\nu_x \nu_y}$.

Corresponding Values of r_{xy} and $r_{\nu_x \nu_y}$.

$r_{\nu_x \nu_y}$	r_{xy}	$r_{\nu_x \nu_y}$	r_{xy}
0.00	0.00	0.55	0.57
0.05	0.06	0.60	0.62
0.10	0.10	0.65	0.67
0.15	0.16	0.70	0.72
0.20	0.20	0.75	0.77
0.25	0.26	0.80	0.87
0.30	0.31	0.85	0.86
0.35	0.36	0.90	0.91
0.40	0.42	0.95	0.96
0.45	0.47	1.00	1.00
0.50	0.52	....	....

Probable Deviation of r_{xy} Computed from Ranks. As given by Pearson: $P.\,E.$ of r_{xy} from ranks $= \frac{0.7063}{\sqrt{N}}\,(1 - r^2)$.

Exercises

4. Determine r_{xy} from the value of $r_{\nu_x \nu_y}$ computed on page 99.

5. Compute the value of the rank r from the data of other exercises and compare with the computed values of the variate r.

The Accuracy of the Coefficient r_{xy} when computed from Ranks. When the measurements are arranged in ranks and the coefficient is computed from the ranks alone, the computation is based on the relatively limited information which the ranks can convey. Hence the resulting coefficient cannot be as trustworthy and reliable as the moment coefficient. However, when a detailed correlation table cannot be constructed owing to a paucity of information, it may still be possible to determine the rank of the individual. If proper allowance is made for the necessarily wide inaccuracy of the computed result, the rank coefficient is better than no coefficient at all for such inaccurate or indeterminate data.

CHAPTER XI

THE MOMENTS OF A DISTRIBUTION

Definitions and Notation. The first moment, obtained by multiplying each deviation by the corresponding frequency, adding the resulting products and dividing by the total frequency of the distribution, was discussed in Chapter IV in connection with the arithmetic mean. The second moment, in which the deviations are squared before multiplication by the frequencies, was discussed in Chapter V. The third and fourth moments, with the deviations cubed and raised to the fourth power respectively, were referred to in Chapter V.

Obviously the moments may be computed about any point by obtaining the deviations from that point and raising to the appropriate power, etc. For most purposes, however, the second and higher moments are computed about the mean which thus serves as a standard origin for the moments.

The moments about the mean are denoted by the symbols μ_1, μ_2, μ_3, μ_4, etc., where the subscripts refer to the order of the moments; that is, the index of the power to which the deviations are raised. Under the same system of notation, the moments about any other point are denoted by μ_1', μ_2', μ_3', μ_4', etc.

The moments about the mean may be computed directly by first computing the mean and then subtracting the value of the mean from each deviation and using the resulting differences in the computations for the moments. This method of computing the moments has the advantages of simplicity and directness but it usually leads to troublesome fractions and it ordinarily involves more labor than the indirect methods which are described in this chapter.

Transformation Formulas for the Moments about the Mean. The formulas for the moments about the mean in terms of the moments about a fixed point will now be derived. Let d be the mean deviation, that is, the distance of the mean from the fixed point of reference, and let the x's be measured from the mean. Then corresponding to a given value of x there will be the deviation x' about the fixed point, so that $x' = x + d$.

From the definition of a moment we have,

$$\mu'_1 = \frac{1}{N}\Sigma(x+d)y = \frac{1}{N}\Sigma xy + \frac{Nd}{N},$$

$$= \mu_1 + d = d, \textit{ since } \Sigma xy \text{ is zero, (Theorem, Chapter IV)};$$

$$\mu'_2 = \frac{1}{N}\Sigma(x+d)^2y = \frac{1}{N}\Sigma x^2y + 2\frac{d}{N}\Sigma xy + \frac{Nd^2}{N},$$

$$= \mu_2 + d^2, \text{ since } \Sigma xy = 0;$$

$$\mu'_3 = \frac{1}{N}\Sigma(x+d)^3y = \frac{1}{N}\Sigma x^3y + \frac{3d}{N}\Sigma x^2y + \frac{3d^2}{N}\Sigma xy + \frac{Nd^3}{N}$$

$$= \mu_3 + 3d\mu_2 + d^3;$$

$$\mu'_4 = \frac{1}{N}\Sigma(x+d)^4y$$

$$= \frac{1}{N}\Sigma x^4y + \frac{4d}{N}\Sigma x^3y + \frac{6d^2}{N}\Sigma x^2y + \frac{4d^3}{N}\Sigma xy + \frac{Nd^4}{N}$$

$$= \mu_4 + 4d\mu_3 + 6d^2\mu_2 + d^4.$$

Transposing a part of the terms in the four preceding equations and changing the signs, we have the following equations which express each moment about the mean in terms of the corresponding moment about the fixed point and the moments of lower order about the mean:

$$\mu_1 = \mu_1' - d = 0, \text{ since } \mu_1' = d;$$
$$\mu_2 = \mu_2' - d^2;$$
$$\mu_3 = \mu'_3 - 3d\mu_2 - d^3;$$
$$\mu_4 = \mu'_4 - 4d\mu_3 - 6d^2\mu_2 - d^4.$$

These formulas for transferring the moments from a fixed point to the mean are arranged in what is called the *continuous form;* that is, they begin with the moment of lowest order and proceed step by step to the fourth moment.

Exercises

1. Compute the third and fourth moments for the student height data at the beginning of Chapter III.

2. By taking the fixed point of reference at various points show that for the data of Student Heights the third and fourth moments are least when computed about the arithmetic mean.

3. Find the first, second, third and fourth moments about the mean of a distribution with frequencies proportional to the successive terms in the expansion of the binominal $(p + p)^n$.

Ans. $\mu_2 = npq$; $\mu_3 = npq\,(p - q)$; $\mu_4 = npq\,3\,(n - 2)\,(pq + 1)$.

The computation of the moments about the mean either directly or by first computing about a convenient origin and then transforming to the mean is open to the serious practical objection that there are no convenient methods of checking the results. The arithmetic of the following summation method is comparatively brief and admits of satisfactory checks on the correctness of the results.

Summation Method of Computing the Moments. The derivation of the formulas of the summation method is somewhat detailed but entirely elementary throughout.

Let us take a distribution with the five frequencies, y_1, y_2, y_3, y_4, y_5, corresponding to values of x equal to 1, 2, 3, 4, 5. By the ordinary direct method, the first moment about the point $x = o$ is $y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5$ divided by N. Now let us arrange the y's in vertical order and add in the manner indicated in the second column following.

(1)

$$\begin{array}{l} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \hline \Sigma y \end{array}$$

(2)

$$\begin{array}{r} y_1 + y_2 + y_3 + y_4 + y_5 \\ y_2 + y_3 + y_4 + y_5 \\ y_3 + y_4 + y_5 \\ y_4 + y_5 \\ y_5 \\ \hline y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5 \end{array}$$

(3)

$$\begin{array}{r} y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5 \\ y_2 + 2y_3 + 3y_4 + 4y_5 \\ y_3 + 2y_4 + 3y_5 \\ y_4 + 2y_5 \\ y_5 \\ \hline y_1 + 3y_2 + 6y_3 + 10y_4 + 15y_5 \end{array}$$

(4)

$$\begin{array}{r} y_1 + 3y_2 + 6y_3 + 10y_4 + 15y_5 \\ y_2 + 3y_3 + 6y_4 + 10y_5 \\ y_3 + 3y_4 + 6y_5 \\ y_4 + 3y_5 \\ y_5 \\ \hline y_1 + 4y_2 + 10y_3 + 20y_4 + 35y_5 \end{array}$$

(5)

$$\begin{array}{r} y_1 + 4y_2 + 10y_3 + 20y_4 + 35y_5 \\ y_2 + 4y_3 + 10y_4 + 20y_5 \\ y_3 + 4y_4 + 10y_5 \\ y_4 + 4y_5 \\ y_5 \\ \hline y_1 + 5y_2 + 15y_3 + 35y_4 + 70y_5 \end{array}$$

The sum of the second column is thus the same as for the first moment. By the direct method the second moment about the same point is $y_1 + 4y_2 + 9y_3 + 16y_4 + 25y_5$ divided by N. Let us designate the sum of column (2), when divided by N, by S_1; the second divided by N, by S_2; the third when divided by N, by S_3, etc. That is,

$$S_2 = \frac{y_1+2y_2+3y_3+ \ . \ . \ .}{N}, \ S_3 = \frac{y_1+3y_2+6y_3+ \ . \ . \ .}{N},$$

$$S_4 = \frac{y_1+4y_2+10y_5+ \ . \ . \ .}{N}, \ S_5 = \frac{y_1+5y_2+15y_3+ \ . \ . \ .}{N},$$

It is apparent on inspection that $2S_3 - S_2$ is the second moment. In symbols,

$$\frac{2}{N}(y_1+3y_2+6y_3+10y_4+15y_5) - \frac{1}{N}(y_1+2y_2+3y_3+4y_4+5y_5)$$

$$= \frac{1}{N}(y_1+4y_2+9y_3+16y_4+25y_5).$$

That is, $\mu'_2 = 2S_3 - S_2$.

The third moment about the same point of reference is

$$\frac{1}{N}(y_1+8y_2+27y_3+64y_4+125y_5).$$

For this moment the following relation is readily verified:

$$\mu'_3 = 6S_4 - 6S_3 + S_2.$$

Extending the reasoning to the case of the fourth moment, we have

$$\mu'_4 = 24S_5 - 36S_4 + 14S_3 - S_2.$$

We thus have four relations connecting the moments with the S's:

$$\mu'_1 = S_2,$$
$$\mu'_2 = 2S_3 - S_2,$$
$$\mu'_3 = 6S_4 - 6S_3 + S_2,$$
$$\mu'_4 = 24S_5 - 36S_4 + 14S_3 - S_2.$$

Transferred to the mean as origin by the formulas of page 104 these moments become

$$
\begin{aligned}
&S_2 = d\,;\\
&\mu_2 = \mu'_2 - d^2 = 2S_3 - S_2 - d^2 = 2S_3 - d(1+d)\,;\\
&\mu_3 = \mu'_3 - 3d\mu_2 - d^3 = 6S_4 - 6S_3 + S_2 - 3d\mu_2 - d^2,\\
&\quad = 6S_4 - 3\mu_2 - 3d\,(1+d) + d - 3d\mu_2 - d^3,\\
&\quad = 6S_4 - 3\mu_2\,(1+d) - d\,(1+d)(2+d)\,;
\end{aligned}
$$

and similarly, $\mu_4 = 24S_5 - 2\mu_3\{2(1+d)+1\}$
$$- \mu_2\{6\,(1+d)\,(2+d) - 1\} - d\,(1+d)\,(2+d)\,(3+d).$$

It is evident that the same relations hold for a larger number of classes than the five which we have assumed for the purpose of illustrating the method.

These relations connecting the moments about the mean with the sums obtained by this process of summation are materially shorter and more convenient than the direct formulas. It will be noticed that the sum of any column is the largest number in the next column, so that a satisfactory check on the summation is afforded.

The following computations for the data of student heights illustrates the summation method.

Computations of this length should never be attempted without first arranging a complete form with a place for each number and that place so chosen that the number is in its most convenient location. The entire computation should be planned before the arithmetic is begun. In this computation the frequencies are accumulated from the bottom. Thus 4, (4 + 9), (13 + 23), . . . are the sums of column (2), and similarly for columns (3), (4), (5). Then each column is added and the sums each divided by the total frequencies. The quotients so obtained are the "S's".

Computation of Moments by Summation—Student Heights

Class	(1) *Freq.*	(2)	(3)	(4)	(5)
1	2	750	5925	28463	105421
2	10	748	5175	22538	76958
3	11	738	4427	17363	54420
4	38	727	3689	12936	37057
5	57	689	2962	9247	24121
6	93	632	2273	6285	14874
7	106	539	1641	4012	8589
8	126	433	1102	2371	4577
9	109	307	669	1269	2206
10	87	198	362	600	937
11	75	111	164	238	337
12	23	36	53	74	99
13	9	13	17	21	25
14	4	4	4	4	4
Totals	750	5925	28463	105421	329625

(1). $d = S_2 = 7.9 \quad S_3 = 37.95 \quad S_4 = 140.56 \quad S_5 = 439.5.$

(2). $d(1+d) = 70.31$

(3). $d(1+d)\ (2+d) = 696.069$

(4). $3(1+d) = 26.7 \qquad \mu_2 = 2S_3 - d(1+d) = 5.592$

(5). $4(1+d) + 2 = 37.6 \qquad \sigma = \sqrt{\mu_2} = 2.36$

(6). $6(1+d)\ (2+d) - 1 = 527.66$

$$\mu_3 = 6S_4 - \mu_2\ (4). - (3). = -2.015$$

$$\mu_4 = 24S_5 - \mu_2\,(5) - \mu_2.\ (6). - (3+d).\ (3) = 85.937$$

Correction Formulas for the Moments. All the methods that have been proposed for finding the moments assume that the frequencies are concentrated at the center of each class. Actually the deviations are continuously distributed from one end of the range to the other so that there is nothing in the nature of the data to correspond to the classes, mid-ordinates, etc. A certain degree of error is therefore introduced by these methods. We are not really working with the actual deviations but with the artificial classes built up from the actual deviations. In how far then are facts, which hold for the classes, of significance for the actual variates? It may well be that in ordinary statistical work the closeness of the measurements may not warrant taking these errors into account but the corrections are easily applied and frequently make a significant difference in the results. However, the corrections should not be applied to data not accurate enough to warrant such care no matter if the corrections are easily applied. The methods adopted in computation must never be such as to presuppose more accuracy than is actually present in the data.

When the distinction is made between the moments as calculated from the class frequencies and deviations and the moments calculated under the assumption of continuous variation, it is customary to denote the values as computed by ν_1, ν_2, ν_3, ν_4, and ν'_1, ν'_2, ν'_3, ν'_4, and reserve the corresponding μ's for the values under the assumption of continuity. When no account is taken of the distinction between the discrete and continuous series of frequencies, the μ's alone are used. The ν's are often spoken of as the raw or unadjusted moments and the μ's as the adjusted moments.

The adjustment or correction formulas are:

$$\mu_1 = \nu_1 = 0$$
$$\mu_2 = \nu_2 - {}^1/_{12}$$
$$\mu_3 = \nu_3$$
$$\mu_4 = \nu_4 - \tfrac{1}{2}\nu_3 + {}^7/_{240}$$

The theory of these corrections is due to Dr. Sheppard and Professor Pearson.

According to the underlying mathematical theory these correction formulas hold in strictness only for a frequency curve with high contact at each end. When these conditions are not satisfied it is probably best not to apply the corrections.

Theorem I. Changing the unit of measurement of the deviations; that is, multiplying each deviation by a constant, multiplies a moment by the constant raised to a power equal to the order of the moment. For,

$$\mu_n = \frac{1}{N}\Sigma x^n y \qquad \text{and} \qquad \Sigma(rx)^n y = r^n \Sigma x^n y.$$

Theorem II. Multiplying or dividing each frequency by a constant does not change the moments. For,

$$\frac{\Sigma x^n ry}{\Sigma ry} = \frac{r\Sigma x^n y}{r\Sigma y} = \frac{\Sigma x^n y}{\Sigma y}$$

Because the values of the third and fourth moments depend on the unit of measure of the deviations it is usual to employ these two moments in the forms β_1 and β_2 respectively, where $\beta_1 = \mu^2_3/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$. The denominators are powers of the standard deviation which measure the dispersion. The powers are arranged so as to show that β_1 and β_2 are independent of the unit of measure of x. Let us write

$$\beta_1 = \frac{N(\Sigma x^3 y)^2}{(\Sigma x^2 y)^3} \text{ and } \beta_2 = \frac{N(\Sigma x^4 y)}{(\Sigma x^2 y)^2}.$$

Then let x be changed into rx where r is any constant. This gives

$$\beta_1 = \frac{N(\Sigma x^3 y)^2 \cdot r^6}{(\Sigma x^2 y)^3 \cdot r^6} = \frac{N(\Sigma x^3 y)^2}{(\Sigma x^2 y)^3} = \frac{N \cdot N^2 \cdot \mu_3^2}{N^3 \mu_2^3} = \frac{\mu_3^2}{\mu_2^3}.$$

It appears that β_1 is a measure of the symmetry of a curve because μ_3 is zero for a symmetrical curve. A negative sign for $\sqrt{\beta_1}$ comes from a negative value for μ_3. When the curve extends farther to the right μ_3 is positive.

The significance of β_2 arises from the value of μ_4 as compared with σ. It may be shown that for the normal curve β_2 equals 3. For a curve having μ_4 more than three times σ^4 there must be a comparative spread of the variates away from the center. Mathematically speaking μ_4 is made up of the fourth powers of the deviations and σ^4 is the square of the average mean square deviation. Hence large deviations enlarge μ_4 more than σ^4.

The term *Kurtosis* has been applied to $\beta_2 - 3$. *Kurtosis* is accordingly a measure of the *flatness* of a curve in comparison with the normal curve.

Exercises

4. Show that adding a constant to each deviation changes the moments.

5 Show that adding a constant to each frequency changes the moments.

6. Show that the square of Σxy *is* $\Sigma x^2y^2 + \Sigma\Sigma x_s y_s x_t y_t$ where the subscripts are attached in the second summation to indicate the product of unequal deviations, and all deviations are measured from the mean and by actually computing the separate value of each summation verify the relation for the distribution 1, 2, 5, 2, 1.

The Moments and the Equation of the Smoothed Curve. It is shown in Chapter II that a smooth curve is fitted on the basis of principles which are assumed true for the data as a whole. One such principle is that of equality of area which assumes that the area under the curve is equal in numerical value to the total frequency of the distribution.

The Principle of equality of moments assumes in addition to the equalities of area and of total frequency that the first, second, third and fourth moments computed from the adjusted frequencies, are respectively equal to the same moments computed from the data.

We may look upon the area or total frequency as a zero moment since $yx^0 = y$, regardless of the value of x.

To illustrate the application of the method of equality of moments let us fit a straight line to the points (2, 4), (3, 3), (4, 6), (5, 7).

The equation of the required line is $y = mx + b$ *where* m and b are to be determined. The y's in terms of m and b are $2m + b$, $3m + b$, $4m + b$, $5m + b$ for the respective points. The total frequency is accordingly: $\Sigma y = 4 + 3 + 6 + 7 = 20$, from the data. We have $\Sigma y = \Sigma(mx + b) = m\Sigma x + Nb = m\ (2+3+4+5) + 4b$, from the assumed equation, $= 14m + 4b = 20$.

The first moments are $\Sigma xy = 2\times4+3\times3+4\times6+5\times7 = 76$, from the data, and $\Sigma(mx + b)x = m\Sigma x^2 + b\Sigma x$,

$$= m(4+9+16+25) + b(2+3+4+5$$
$$= 54m + 14b = 76.$$

We thus have the two equations to solve for m and b,

$$14m + 4b = 20$$
$$54m + 14b = 76.$$

These equations give $m = 1.2$ and $b = 0.8$.

Hence $y = 1.2x + 0.8$ is the equation which gives the straight line having the same frequency and the same first moments as the data.

Let us fit the parabola, $y = a + bx + cx^2$ to the same data.

The equality of the y's as computed from the data and from the curve gives the equation $4a + 14b + 54c = 20$.

The first moments give $14a + 54b + 224c = 76$.

The second moments give $54a + 224b + 978c = 314$.

By the usual methods of elementary algebra we find the following values for a, b, c:

$$a = +6.3$$
$$b = -2.3$$
$$c = +0.5$$

and hence $y = 6.3 - 2.3x + 0.5x^2$.

It is evident that an extension of the above methods would give an equation of the form $y = a + bx + cx^2 + dx^3 + \ldots\ldots$

The derived constants, a, b, c, d . . . would be such that the frequency, first moments, second, third, etc., moments computed from the ordinates under the curve would be equal to the same moments computed from the given frequencies.

An extended application of the method of moments to curve fitting is presented in the Appendix where the generalized normal curves of Pearson are treated.

Least Square Test of Fit. The basic idea of least squares is of interest as an alternative method of determining the "best fitting" straight line.

Let us take the straight line $y = mx + b$

and the points (2, 4), (3, 3), (4, 6) and (5, 7) which, for convenience, we may denote as (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4). On substituting,

$$\begin{aligned} y_1 &= mx_1 + b, \\ y_2 &= mx_2 + b, \\ y_3 &= mx_3 + b, \\ y_4 &= mx_4 + b, \end{aligned}$$

The values of y so computed will exactly agree with the actual y's only for points lying exactly on the line. In general there will be an error, say, e. Hence we may write

$$\begin{aligned} mx_1 + b - y_1 &= e_1 \\ mx_2 + b - y_2 &= e_2 \\ mx_3 + b - y_3 &= e_3 \\ mx_4 + b - y_4 &= e_4 \end{aligned}$$

According to the least square basic test, the best fit is where the sum of the squares of the errors is least. On squaring the e's and adding we have, $(mx_1+b-y_1)^2+(mx_2+b-y_2)^2+(mx_3+b-y_3)^2+(mx_4+b-y_4)^2$, which is to be a minimum. This expression is a minimum when the following two expressions each equated to zero are true:

$$x_1(mx_1 + b - y_1) + x_2(mx_2+b-y_2)+x_3(mx_3+b-y_3) + x_4(mx_4 + b - y_4) = 0,$$

$$\text{and, } (mx_1+b-y_1)+(mx_2+b-y_2)+(mx_3+b-y_3) + (mx_4 + b - y_4) = 0.$$

On collecting we have, $m(x^2_1+x^2_2+x^2_3+x^2_4)+b(x_1+x_2+x_3+x_4) -(x_1 y_1+x_2 y_2+x_3 y_3+x_4 y_4) = 0,$
and $m(x_1+x_2+x_3+x_4)+4b - (y_1+y_2+y_3+y_4) = 0.$

On substituting the values for the y's and x's we have,

$$\begin{aligned} 54m + 14b - 76 &= 0 \\ 14m + 4b - 20 &= 0. \end{aligned}$$

These two equations for determination of m and b are exactly the same equations as have been derived by using the methods of moments.

The student of calculus will recognize the foregoing equations as the result of equating to zero the partial derivatives with respect to *m* and *b*.

These results show that a straight line fitted to data by the method of moments conforms to the least square basic condition. Insofar as fitting a straight line is concerned the two methods are identical.

CHAPTER XII

FURTHER THEORY OF CORRELATION

Index of Correlation. The foregoing measures of correlation may be made use of in ways which can be applied to advantage. The concept of the *index of correlation* is based on a rearrangement of the formula for the correlation coefficient or the correlation ratio. This rearrangement involves a measurement of the divergence of the data from the regression line or curve. The ideas underlying the index may be developed as follows.

Let there be a series of points (X, Y) and let a line of regression be $Y' = a + bX$. The actual ordinates or values, the Y's will differ from the computed Y's by the differences $(Y - Y')$. The sum of the squares of all such differences is $\Sigma(Y - Y')^2$. This latter summation is divided by the sum of the squares of the deviations of the Y's from the mean, $\overline{Y}$, of all the Y's. We then have the expression $\dfrac{\Sigma(Y - Y')^2}{\Sigma(Y - \overline{Y})^2}$.

It is shown on a following page that the above expression is equal to $(1 - r^2)$ where r is the correlation ratio. Or, if the regression is not linear, equals $(1 - \eta^2)$. On transposing and changing signs we have $r^2 = 1 - \dfrac{\Sigma(Y - Y')^2}{\Sigma(Y - \overline{Y})^2}$.

Aside from the mathematical proof of the foregoing relation in r, it can be seen by general reasoning that the summation numerator is proportionally smaller as the actual points cluster closely about the regression line. That is, the more highly correlated the data the smaller is the summation fraction and hence the larger is the value of r.

On substituting in the second moment equation we have

$$Nr\sigma_x\sigma_y = b_{yx} \,.\, N\sigma_x^2$$

Therefore $b_{yx} = r \,.\, \frac{\sigma_y}{\sigma_x}$ and hence $\bar{y}_x = r \,.\, \frac{\sigma_y}{\sigma_x} \,.\, x$ is the required regression equation or $\bar{y}_x = b_{yx} \,.\, x$, where $b_{yx} = r \frac{\sigma_y}{\sigma_x}$.

Exercises

1. Derive the regression equation $\bar{x}_y = r \frac{\sigma_x}{\sigma_y} y$.

2. Prove in detail that $\Sigma\Sigma n_{xy} y = o$ where x and y are measured from the mean.

When x and y are measured from the original axes the regression equations become

$$\bar{y}_x - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$\bar{x}_y - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}).$$

Proof of Correlation Index Formula. The formula for the index of correlation $r^2 = 1 - \frac{\Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2}$ may be written $r^2 = 1 - \frac{\Sigma(y - y')^2}{\Sigma y^2}$ where x and y are measured from axes through the means.

From the regression equation, $y' = bx$, where $b = r \frac{\sigma_y}{\sigma_x}$.

We have $y - y' = y - bx$ and $\Sigma(y - y')^2 = \Sigma(y - bx)^2$.
That is, $\Sigma(y - y')^2 = \Sigma y(y - bx) - b\Sigma x(y - bx)$,
$= \Sigma y^2 - b\Sigma xy - b\Sigma xy + b^2\Sigma x^2$.

But $b\Sigma x^2 = \Sigma xy$, since $y = bx$ for points on the line of regression, and hence $b^2\Sigma x^2 = b\Sigma xy$.

Hence $\Sigma(y - y')^2 = \Sigma y^2 - b\Sigma xy = N\sigma_y^2 - \dfrac{(\Sigma xy)^2}{\Sigma x^2}$, from the regression equation, $= N\sigma_y^2 - Nr^2\sigma_y^2$

$$\text{or} \quad \frac{\Sigma(y - y')^2}{N\sigma_y^2} = 1 - r^2.$$

$$\text{That is,} \quad r^2 = 1 - \frac{\Sigma(y - y')^2}{\Sigma y^2}.$$

The Relation Between η and r. It was shown in Chapter IX that η and r have the same numerical value when the regression is truly linear. Hence a lack of agreement in the values of η and r is an indication of a divergence from linearity in the regression. The difference between η and r may be expressed by the two equations:

$$N\sigma_y^2(\eta_y^2 - r^2) = \Sigma n_x(\overline{Y}_x - \bar{y}_x)^2$$

and $N\sigma_x^2(\eta_x^2 - r^2) = \Sigma n_y(\overline{X}_y - \overline{x}_y)^2$, where $\overline{Y}_x$ and $\overline{X}_y$ are the regression line means as opposed to $\bar{y}_x$ and $\overline{x}_y$, as the actual means.

To prove the first of these formulas let us add and subtract $\bar{y}$ for each term in the summation $\Sigma n_x(\overline{Y}_x - \bar{y}_x)^2$. We then have after expansion,

$$\Sigma n_x(\overline{Y}_x - \bar{y}_x)^2 = \Sigma n_x \{\overline{Y}_x - \bar{y})^2 - 2\,(\overline{Y}_x - \bar{y})\,(\bar{y}_x - \bar{y}) + (\bar{y}_x - \bar{y})^2\}$$

On substituting from the regression equations the right hand form becomes

$$\Sigma n_x(x - \overline{x})^2 \,.\, r^2 \frac{\sigma_y^2}{\sigma_x^2} + \Sigma n_x(\bar{y}_x - \bar{y})^2 - 2\Sigma n_x \,.\, r \frac{\sigma_y}{\sigma_x}\,(\bar{y}_x - \bar{y})\,(x - \overline{x}),$$

which equals $N\sigma_x^2 \,.\, r^2 \dfrac{\sigma_y^2}{\sigma_x^2} + N\sigma_y^2 \,.\, \eta_y^2 - 2r\dfrac{\sigma_y}{\sigma_x} Nr\sigma_x\sigma_y$,

which equals $N\sigma_y^2 \,.\, \eta_y^2 - N\sigma_y^2 r^2$.

That is $\Sigma n_x(\overline{Y}_x - \bar{y}_x)^2 = N\sigma_y^2(\eta_y^2 - r^2)$.

Exercises

3. Prove the formula $\Sigma n_y \ (\overline{X}_y - \bar{x}_y)^2 = N\sigma_x^2 \ ({\eta^2}_z - r^2)$.

4. **Show from these formulas that $\eta > r$.**

5. **Show that the same pair of equations will be obtained for the regression lines if the assumed lines are fitted to the individual frequencies instead of to the means of the arrays.**

The Coefficient r for Non-linear Regression. After the further correlation theory of this Chapter it may be well to repeat that r is always too small in the case of a distribution not strictly linear. If the regression curve is carefully drawn a fair idea of the trustworthiness of r can be obtained by observing the departures of that curve from linearity. A more accurate way, of course, is to compute both η and r and observe the difference in value of the two measures of correlation.

Since $$r = \frac{\Sigma\Sigma n_{xy}(x - x)(y - \bar{y})}{N\sigma_x\sigma_y}$$

the size of r varies directly as the value of the summation in the numerator. In this summation the largest product values are when the points are along an x and y diagonal and hence r will be largest numerically when the values of n_{xy} are largest along a diagonal. If the frequencies tend to lie along one diagonal the value of r will be positive; along the other, negative. If the distribution should exhibit two tendencies,—to concentrate along both diagonals—the cancellation of terms with opposite signs would give rise to a small value for r.

Again the regression may be markedly non-linear, circular, or periodic as a sine curve, so that the straight line fitted to the means of the arrays is practically horizontal, resulting in a very small value for r. This may be true even for data which shows a definite tendency for the frequencies to cluster closely along the curve of means; that is, it is possible for r to have a small value even though the data shows the attributes to have in fact a high degree of correlation.

The Most Probable Value of a Characteristic can be determined from r. Let us first define the properties, homoscedasticity and homoclisy.

The squared standard deviation from the regression curve of the frequencies of an array has been denoted by the symbol (using larger type to distinguish the double subscripts)

$$\sigma^2_{a_y} \quad \text{where } \sigma^2_{a_y} = \sigma_y^2\,(1-\eta^2)$$

$$\text{or, in terms of } r\text{ , } \sigma^2_{a_y} = \sigma_y\ \ (1-r^2).$$

It must be remembered that these are mean values so that it may well happen that a computed standard deviation of an individual array may differ considerably from that obtained from these general formulas. A distribution in which all arrays of a given sense; that is, all y or all x arrays have the same standard deviation is said to be *homoscedastic* with respect to the arrays of that sense.

It has been assumed that the frequencies of the arrays are so distributed that the means and the modes coincide; that is, so that the mean is the most probable value of the array, but this may not always be true. The arrays of a distribution are said to be *homoclitic* when the mean is the most probable value of the array.

On the basis of the just preceding definitions it may be said that for homoclitic arrays the most probable value of y corresponding to a given value for x is found from the equation

$$y = r\,\frac{\sigma_y}{\sigma_x}\,x, \quad \text{or}$$

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}\,(x - \bar{x}).$$

A knowledge of the most probable values is of little importance unless accompanied by information of the dispersion about that value; that is, of the standard deviation and the probable deviation. Since the entire theory of estimating values of a characteristic is based on the coefficient of correlation the probable deviation of y when obtained from the regression curve is logically based on r instead of η and hence is $0.67459\ \sigma_y$

$\sqrt{(1-r^2)}$, and not 0.67449 $\sigma_y \sqrt{(1-\eta^2)}$, (provided the arrays are fairly homoscedastic, otherwise no general formula is possible and the dispersion of each array must be computed directly from the data of the respective arrays). Likewise, the probable error of x found from the regressions is 0.67459 $\sigma_x \sqrt{(1-r^2)}$, with the same restrictions as to homoscedasticity.

If the three conditions of linearity of regression, of homoscedasticity, and of homoclisy are satisfied the just preceding theory of estimating the value of a variable characteristic is complete and practically valuable. In ordinary distributions these conditions are likely to hold, at least approximately, so that when intelligently applied the theory is of importance. In every case the regression curve should be determined graphically and both η and r computed and the difference in their values noted, and the test for linearity applied. If there is doubt as to the homoscedasticity, the standard deviations can be computed directly from the arrays in question and the probable deviations determined from the resulting values instead of from the preceding formula. The question of homoclisy is usually disregarded though wide departures should be noted and taken into consideration.

Exercises

8. What is the most probable weight of a student of height 70 inches using tho data of Chapter III?

9. What is the most probable height of a student of weight 132 pounds?

10. From the Chicago Live Stock prices of an earlier Chapter, what is the most probable top beef cattle price for a month with a top hog price of $8.25?

11. Compute the probable deviations from the most probable values of Exercises 11, 12, 13.

12. Discuss the practical reliability of the preceding estimates. In how far is the probable deviation a trustworthy index of this reliability?

Spurious Correlation. By dividing each deviation by a third variable, it is possible to introduce correlation into strictly uncorrelated material to as great an extent as 0.5.

From the following formula for the correlation between index numbers where $\frac{x}{y}$ and $\frac{z}{y}$ are the two series, but where there is no correlation among x, y and z, it appears that $r = 0.5$ if $\frac{\sigma_y}{y} = \frac{\sigma_x}{x} = \frac{\sigma_z}{z}$,

$$r = \frac{\dfrac{\sigma_y^2}{\bar{y}^2}}{\sqrt{\dfrac{\sigma_x^2}{\bar{x}^2} + \dfrac{\sigma_y^2}{\bar{y}^2}} \; \sqrt{\dfrac{\sigma_z^2}{\bar{z}^2} + \dfrac{\sigma_y^2}{\bar{y}^2}}.}$$

Hence care must be taken in dealing with index numbers that the full value of r is significant for the absolute values of the measurements. By computing r from the above formula by substituting the values for the symbols the value of the greatest possible degree of spurious correlation is obtained. A value of r greater than this value is certainly significant; a value less may be significant but must be accepted with caution.

Since by the formula the spurious correlation is zero when the standard deviation or variability of y is zero, it follows that the base of a system of index numbers should be as nearly constant as possible.

A theory of spurious correlation might be developed for the correlation ratio but the algebraic details are so much more workable for the correlation coefficient that it would hardly be worth the additional effort. It is conceivable that such a theory would be practically necessary but it is unlikely because after all only approximate results are valuable. There would be little of value in attempting to measure the degree of spurious correlation with precision.

It must be remembered that the matter of spurious correlation is essentially one of interpretation. The question is what does correlation mean. The correlation is actual and real for the indices but it may be spurious insofar as the absolute values of the measurements are concerned.

CHAPTER XIII

SIGNIFICANCE OF DIFFERENCES

The Significance of a Difference. The analytical statistician must evaluate differences in order to appraise the significances of the data as an indicator of basic differences. If, for illustration, the mean height of one student distribution is 67.9 inches and for a second, the mean is 66.5 inches does this 1.4 inches difference indicate that the two distributions are from populations of basically different height characteristics, or is the difference merely a chance difference from random sampling?

In this Chapter the developments of the preceding chapters together with certain additional items are brought together with reference to methods of determining the significance of differences. The purpose here is to present the underlying ideas rather than rules for their application. For the applications reference is made to the mathematical treatises and the available manuals. Certain theorems will first be stated.

The variance of the difference between two variables is the sum of the variances less twice the covariance.

Let x_a and x_b represent the two series of variables, the two sets of height measurements, say, and set, $x_d = x_a - x_b$. Then

$$\sigma^2_{x_d} = \frac{1}{N} \Sigma x_d{}^2$$

$$= \frac{1}{N} \Sigma (x_a - x_b)^2 = \frac{1}{N} \quad [\Sigma x_a{}^2 + \Sigma x_b{}^2 - 2\Sigma x_a \ x_b]$$

$$= \sigma^2 x_a + \sigma^2 x_b - 2r\,\sigma x_a\,\sigma x_b \text{ (since } \Sigma x_a x_b = N r\,\sigma x_a\,\sigma x_b$$

That is, the mean squared deviation of the differences from the mean of the differences, the *variance of such differences,* is equal to the sum of the two variances less twice the covariance.

It is interesting to note that for perfectly correlated data where $\sigma x_a = \sigma x_b$ and $r = 1$, $\sigma x_d = o$, as it should be. On the other hand for uncorrelated data where $r = o$, $\sigma x_d^2 = \sigma x_a^2 + \sigma x_b^2$. We accordingly have the variance of the difference between two uncorrelated variables is the sum of the variances.

The formula $\sigma x_d^2 = 2 \; . \; \sigma x_a^2$ *or* $\sigma x_d = \sqrt{2} \; . \; \sigma x_a$ shows that for uncorrelated data the differences between two samples from the same population are less stable than one sample, this variance being increased by the factor $\sqrt{2}$ which is 1.4142.

The variance of the difference between two uncorrelated means from the same population is

$$\sigma \bar{x}_d^2 = \sigma \bar{x}_a^2 + \sigma \bar{x}_b^2 = \sigma^2 \left(\frac{1}{N_a} + \frac{1}{N_b}\right),$$

where $\sigma x_a^2 = \sigma x_b^2 = \sigma^2$, and from a preceding formula the variance of the mean is $\frac{1}{N} . \sigma^2$.

It thus appears that the variance of the population is in the ratio of 1 to $\left\{\frac{1}{N_a} + \frac{1}{N_b}\right\}$ to the variance of the difference of the means: for samples of 100, this is 1 to $\frac{1}{50}$.

If it be assumed that the variances are each the same as that of the general population we have

$$\sigma x_d^2 = \sigma^2 + \sigma^2 - 2r\sigma^2 = \sigma^2 \left\{\frac{1}{N_a} + \frac{1}{N_b} - \frac{2r}{\sqrt{N_a N_b}}\right\}$$

$$= \frac{2\sigma^2}{N} (1 - r), \text{ if } N_a = N_b = N.$$

The just derived formula shows that the variance of the differences of the means is zero for perfectly correlated data.

Difference Between Two Proportions. A measure of the significance of the difference between two proportions is of practical use in many investigations. In deriving a formula for this measure let us assume a population of individuals possessing a certain characteristic in the proportion of p.

Repeated drawings from this population would yield values for p varying from 0 to 100 per cent. If N is the number drawn in each sample it is reasonable to assume that the mean proportion from a great number of samples will be p so that the mean number of individuals drawn with the characteristic in question will be Np. That is $\overline{X} = Np$. It may be readily shown mathematically that the standard deviation will be $\sigma_x = \sqrt{Npq}$ where $q = 1 - p$ and the standard deviation of the mean $\overline{X}$ will be

$\sqrt{N\,pq}/N$, which may be written $\left(\dfrac{pq}{N}\right)^{\frac{1}{2}}$

The matter of interest becomes one of testing whether two series of samples with mean proportions $\overline{p}_1$ and $\overline{p}_2$, say, are most likely to come from the same population. In symbols this is whether the difference $(\overline{p}_1 - \overline{p}_2)$ does not differ significantly from zero by more than might arise merely from random sampling variations.

From preceding theorems the standard deviations of the difference in means is $(\sigma^2_{\overline{p}_1} + \sigma^2_{\overline{p}_2})^{\frac{1}{2}}$, provided there is no correlation between the proportions.

But $\sigma^2_{\overline{p}_1} = \dfrac{pq}{N_1}$, where p is the true proportion in the population and $q = 1 - p$ and N_1 is the number of individuals in the samples, and $\sigma^2_{\overline{p}_2} = \dfrac{pq}{N_2}$.

Hence the standard deviation of the difference in the means is $(pq/N_1 + pq/N_2)^{\frac{1}{2}} = (pq)^{\frac{1}{2}}(1/N_1 + 1/N_2)^{\frac{1}{2}}$.

Some assumption must be made for the value of p and hence q. A reasonable assumption is to take p equivalent in value to the weighted mean of $\overline{p}_1$ and $\overline{p}_2$. Hence set $p' = \dfrac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$.

We now measure the observed difference in the means in terms of the standard deviation of this difference and have $\frac{p_1 - p_2}{(p'q')^{\frac{1}{2}}\,(1/N_1 + 1/N_2)^{\frac{1}{2}}}$. It may be safely assumed that this coefficient is normally distributed, regardless of the form of distribution of the characteristics in the population and hence that the probabilities of values to exceed any derived value can be obtained from a table of areas under a normal curve.

As an illustration of the foregoing discussion of the significance of the difference between means of proportions let us take two series of tossings of 100 pennies and 64 dimes. Assume that p_1 (pennies) is 0.52 and p_2 (dimes) is 0.48.

On referring to a table one finds that the probability of a difference numerically greater than 0.50 is 0.309, which is 309 times 1000 or almost 1 in three times. Whether the difference of .04 in the sample means indicates a structural difference in the two series of coins may be a matter open to individual opinion. The point here is that a measure or index has been computed showing the chances that a *numerically greater* deviation might arise from purely random sampling from the same population.

Differences in General. The foregoing discussion of the significance of differences presents the basic ideas and some of the more important special ideas. The differences of correlation coefficients, or correlation ratios, regression equations, multiple correlation indexes, partial correlation ratios, may each be discussed for measures of significance.

The mathematical literature is becoming extensive on the subject of the significance of differences. A statistician without the benefit of a thorough training in advanced mathematics need not hesitate in an attempt to understand the underlying ideas.

CHAPTER XIV

CORRELATION IN DISTRIBUTIONS WITH A SMALL NUMBER OF CLASSES

Correlation from a Few Classes. A statistician working with data arising from experimental science, as biology or agriculture, or any data where the classes are very wide and hence few in number, has use for methods of measuring correlation adapted to this type of data. Since such a distribution has only a few classes the volume of detailed information presented is relatively small and hence the significance of any mathematical index from such data must be interpreted with caution. It must always be realized that by the use of any amount of mathematics one can not obtain from data more information than the data actually contains.

Distinctive Ideas for Correlation from Few Classes. The indexes which have been derived for the measurement of correlation in data distributed in few classes are based on the elimination of variations due to random sampling, or else are based on comparisons with the variations which might arise in samples drawn from the same population. Analysis of variance, randomized blocks, Latin squares, measures of contingency, Chi-square distributions each are more or less derived from some relationship with the probabilities from random sampling.

This basic idea of the variations from random sampling is also at the bottom of the idea of the probable error and the standard error, and in any measure of correlation. The use of it is only more direct in deriving the indexes of the present chapter.

Analysis of Variance. The theory of correlation can be made more convenient for working with a small number of classes by following the ideas of the *analysis of variance.* The development of these ideas is facilitated by certain preliminary descriptions.

Let us assume a point with ordinate Y and assume further that the corresponding point on the regression line has an ordinate of length Y'. The difference between the actual value of the ordinate and the value computed by the regression line is then $(Y - Y')$.

It is necessary to define certain other distinctions. The difference between the mean ordinate $\overline{Y}$ and the actual ordinate Y is, of course, $(Y - \overline{Y})$. Likewise we need the difference between the regression line ordinate and the mean of all of the Y's, $(Y' - \overline{Y})$.

The Y', the ordinate of the point on the regression line is the mean of the Y array. Hence the mean squared deviation of these Y' points about the horizontal line through the center, is, of course, the mean squared deviation used in the correlation ratio and measures the variation in the means of the arrays. Hence the sum of all the squares $(Y'-Y)^2$ is the significant factor in the correlation ratio.

The points themselves do not all lie on their respective regression means. This variation from the mean, however, is of the general nature of accidental variations arising from random sampling. The expression, $\Sigma(Y - Y')^2$ gives the sum of squares of deviations of the points from their respective regression line means.

We have finally the result obtained by taking the difference between each ordinate and the mean ordinate of all of the Y's and squaring and totaling so as to give $\Sigma(Y - \overline{Y})^2$.

It will be presently proved that

$$\Sigma(Y - \overline{Y})^2 = \Sigma(Y - Y')^2 + \Sigma(Y' - \overline{Y})^2.$$

It is understood that these summations cover each individual of the distribution.

From the above equation it is apparent that the correlation part, namely, the last term, can be obtained by first computing the sum of the deviations from the mean and then the sum of squares of deviations from the regression line and then subtracting these quantities. The remainder will be the sum of squares of deviations of the regression line from the mean ordinate of all of the Y's, or each term may be computed independently.

This equation is the principal equation of the analysis of variance.

Demonstration of Principal Equation of Analysis of Variance. We are to prove:

$$\Sigma(Y-\overline{Y})^2=\Sigma(Y-Y')^2+\Sigma(Y'-\overline{Y})^2.$$

By adding and subtracting Y' to the left hand member of this equation we have,

$$\begin{aligned}\Sigma(Y-\overline{Y})^2&=\Sigma\left[(Y-Y')+(Y'-\overline{Y})\right]^2\\&=\Sigma(Y-Y')^2+\Sigma(Y'-\overline{Y})^2+2\Sigma(Y-Y')\,(Y'-\overline{Y})\end{aligned}$$

Now the third term on the right hand side vanishes because for each value of X, $(Y'-\overline{Y})$ is constant and hence with the summation on Y, only, we have $2(Y'-\overline{Y})\ \Sigma(Y-Y')$.

Since, Y' is assumed to be the mean of the Y's in the Y array of each type X the sum of the differences between each variate and the mean is zero, as was shown in the Chapter on Averages.

Since this term is zero in each array the sum for all values of X must accordingly be zero.

We have therefore proved that the sum of the squares of the difference of each Y from the mean of all the Y's is equal to the sum of the squares of the differences of each Y from the mean of its Y array plus the sum of the squares of the differences of each array mean and the mean of all the arrays.

It should be noted that the foregoing proof holds strictly only when the regression is strictly linear. It is easy to suspect that if the regression is only sensibly linear the errors from assuming that all array means lie on the regression line will be both positive and negative, and hence the sum of the errors will be small. That is, the term $\Sigma(Y-Y')\ (Y'-\overline{Y})$ is very small, if it does not vanish altogether, for sensibly linear regression.

A more mathematical demonstration of the vanishing of the term $\Sigma(Y-Y')(Y'-\overline{Y})$ is based on the least square derivation of the regression equations which is given in the following paragraph. For this proof let us substitute $a + bx$ for Y'. We then have,

$$\begin{aligned}\Sigma(Y - Y')\ (Y' - \overline{Y}) &= \Sigma(Y - a - bX)\ \ (a + bX - \overline{Y}) \\ &= a\Sigma(Y - a - bX) + b\Sigma X(Y - a - bX) \\ &\qquad - \overline{Y}\Sigma(Y - a - bX).\end{aligned}$$

But each summation factor, that is, each term in the right hand side of the equation, is one or other of the two regression normal equations and hence is zero.

Derivation of Equation of Regression Line. A least square derivation of the regression line equations is helpful in deriving a number of formulas in the analysis of variance. In this derivation we assume that $Y' = a + bX$ is a regression line with a and b so chosen that the sum of the squares of the errors made by substituting Y' for Y is a minimum. That is, so that $\Sigma(Y - Y')^2$ is a minimum. On substituting we have

$$\Sigma(Y - Y')^2 = \Sigma(Y - a - bX)^2.$$

For this to be a minimum the least square normal equations must hold:

$$\Sigma\,(Y - a - bX) = o,$$
$$\text{and } \Sigma\, X\,(Y - a - bX) = o.$$

Expanding these normal equations, we have,

$$\Sigma Y - aN - b\Sigma X = o$$
$$\text{and} \qquad \Sigma XY - a\Sigma X - b\Sigma X^2 = o.$$

These two equations can be solved for a and b in terms of X and Y. We give the solution only for the case where the deviations are measured from the means only. Then $\Sigma X = \Sigma Y = o$,

and hence $a = o$, and $b = \dfrac{\Sigma XY}{\Sigma X^2}$.

Some Other Theorems. It is useful to have $\Sigma(Y - Y')^2$ in terms of Y and X. On expanding and substituting $a + bX$ for Y', we have,

$$\begin{aligned}\Sigma(Y - Y')^2 &= \Sigma(Y - a - bX)^2 \\ &= \Sigma Y(Y - a - bX) - a\,\Sigma(Y - a - bX) - b\Sigma X\ \ (Y - a - xX) \\ &= \Sigma y^2 - a\,\Sigma y - b\Sigma xy,\end{aligned}$$

since both $\Sigma(Y - a - bX)$ and $\Sigma X\ \ (Y - a - bX)$ vanish, being the normal equations for the regression line.

If we measure Y about the mean Y and, as is usual, replace Y by y we have, since $\Sigma Y = o$ when taken about the mean,

$$\Sigma(y-y')^2 = \Sigma y^2 - b\Sigma xy$$
$$= \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}.$$

Since $b = \frac{\Sigma xy}{\Sigma x^2}$ directly from the least square derivation or

$$b = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x\sigma_y}\cdot\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x{}^2} = \frac{\Sigma xy}{\Sigma x^2}.$$

Again $\Sigma(Y'-\overline{Y})^2 = \Sigma y'^2$, from the mean,

$$= b^2\Sigma x^2 = \frac{(\Sigma xy)^2}{(\Sigma x^2)^2}\cdot\Sigma x^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}.$$

Of course, $\Sigma(Y-\bar{y})^2 = \Sigma\, y^2$ when measured from the mean.

A Formula for the Correlation Coefficient. In a previous Chapter it was seen that

$$r^2 = 1 - \frac{\Sigma(Y-Y')^2}{\Sigma(Y-\overline{Y})^2}$$
$$= \frac{\Sigma(Y-\overline{Y})^2 - \Sigma(Y-Y')^2}{\Sigma(Y-\overline{Y})^2},$$

from the principal equation, $= \frac{\Sigma(Y'-\overline{Y})^2}{\Sigma(Y-\overline{Y})^2} = \frac{\Sigma(y')^2}{\Sigma y^2}$, about the mean,

$$\text{or } \Sigma(y')^2 = r^2\Sigma y^2.$$

It has already been proved that

$$\Sigma(y-y')^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$$
$$= \Sigma y^2 - r^2\Sigma y^2$$
$$= (1-r^2)\ \Sigma y^2$$

We may then write the basic theorem of the analysis of variance in the form,

$$\Sigma y^2 = (1-r^2)\ \Sigma y^2 + r^2\Sigma y^2.$$

This last form brings out mathematically the fact that the correlation is significant as the sum of the squares of the variations of the regression means, that is, the second left hand term, is large in comparison with such variations from the regression.

Further Mathematical Development of Analysis of Variance is beyond the scope of this book, neither can we extend the general ideas here presented to the splitting up of co-variance with its relations to the correlation ratio. Reference is made to R. A. Fisher "Statistical Methods for Research Workers", Oliver and Boyd, Edinburgh and London.

Method of Contingency. A method of measuring correlation which is readily adapted to distributions having broad classes is the method of contingency. This method is based on a measurement of the divergence of the frequency of each class from strictly uncorrelated frequencies. We may return to the student height-weight data for an illustration.

When there is no tendency for certain weights to be most often associated with certain heights, the frequency of a subgroup should be proportional to the total frequencies of its two arrays. Thus imagine the frequencies of the sub-groups erased from the Correlation Table on page 76 and then filled in entirely at random; that is, without bias or selection. Since 110/750 of the total frequency of the distribution appears in the height array of weight type 137; that is, since 110 individuals out of 750 are of weight 137, it is logical to assume that this height array contains 110/750 of the frequency of each array which it crosses. The frequency of the sub-group (68 — 137), for instance, should be 110/750 of 126. And in general, when the individuals are placed at random, the frequency of a sub-group is given by the formula $\frac{n_x \cdot n_y}{N}$. For the *y* array of type *x* contains $\frac{n_x}{N}$ of the frequencies of each array which it crosses. The frequency of the *x* array of type *y* is n_x. Hence the subgroup of intersection has a frequency $\frac{n_x}{N} \cdot n_y$, which equals $\frac{n_x \cdot n_y}{N}$.

Now, if in the actual distribution, the frequency of a sub-group, n_{xy}, is larger or smaller than the random selection frequency given by the formula $n_x \,.\, \frac{n_y}{N}$ the divergence must be due to the presence in the data of a tendency for certain values of the attributes to be most often associated and hence the total extent of this divergence is a measure of the degree of the association or correlation in the data. This method of measuring correlation is called the *method of contingency.*

The difference $n_{xy} - \frac{n_x \,.\, n_y}{N}$ is squared to prevent the cancelling of positive and negative values. Since only the relative size of the difference is significant, this square is divided by the above random selection frequency $\frac{n_x \,.\, n_y}{N}$. On summing all such values, we have the mean square contingency Φ^2 where

$$N\Phi^2 = \Sigma\Sigma \frac{\left(n_{xy} - \frac{n_x \; n_y}{N}\right)^2}{\frac{n_x \; n_y}{N}} .$$

On expanding and reducing, this summation is arranged in a more convenient form for computation. We have

$$\frac{\left(n_{xy} - \frac{n_x \; n_y}{N}\right)^2}{\frac{n_x \; n_y}{N}} = N \frac{n_{xy}{}^2}{n_x n_y} - 2n_{xy} + \frac{n_x \, n_y}{N}$$

and hence, $\Sigma\Sigma \frac{\left(n_{xy} - \frac{n_x \; n_y}{N}\right)^2}{\frac{n_x \; n_y}{N}} = \text{N} \; \Sigma\Sigma \frac{n^2{}_{xy}}{n_x \, n_y} - 2N + N,$

since $\Sigma\Sigma \frac{n_x \, n_y}{N} = \frac{1}{N} \Sigma n_x \, \Sigma n_y = \frac{1}{N} \Sigma n_x \; N = \Sigma n_x = N.$

Therefore $\Phi^2 = \Sigma \frac{n^2{}_{xy}}{n_x \; n_y} - 1.$

APPENDIX I

DERIVATION OF EQUATION OF NORMAL PROBABILITY CURVE

The equation of the Normal Probability Curve follows from the basic statement of a normal probability distribution. The frequencies, or the ordinates, are the resultants of a large number of elemental attributes which are individually small and each are equally likely to be positive or negative. It may be assumed that each of these elements are equal in force or strength.

The height of a student, to refer again to the much used illustration of student heights, depends on the thickness of cartileges in the spinal column; shape of the head; length of thigh bones; posture; and so on and so on. Each of these influences may be further broken down into small influences. It is not difficult to see that if analyzing the elemental influences involved in the stature of a student is continued far enough the conditions just laid down for a normal probability distribution would seem to be conformed to closely. In other words, it is not surprising that the equation of student heights is of the form of a normal probability curve.

There are a number of mathematical lines of reasoning which lead to this equation. The method here used is an extension of the binomial expansion.

Since plus and minus are equally likely, as are heads and tails in coin tossing, the probability of the occurrence of a positive value is $\frac{1}{2}$. The chance that all elements are positive in *m* cases is $(\frac{1}{2})^m$, just as the chance that 3 heads result in three throws is 1 in 8 or $(\frac{1}{2})^3$.

Let the numerical magnitude of an element be represented by (delta x) Δx. Then if all m are positive the resultant is $m \,.\, \Delta x$ and the probability of this value, $m \,.\, \Delta x$, is $(\frac{1}{2})^m$. That is, we here have a point with $x = m \,.\, \Delta x$ and $y = (\frac{1}{2})^m$.

Let $m - 1$ elements be positive and 1 be negative. Then the resultant is $[(m - 1)\Delta x - \Delta x]$ or $(m - 2) \cdot \Delta x$ while the probability is $m\ (\frac{1}{2})^m$. These coordinates are hence $x = (m - 1)\Delta x$ and $y = m\ (\frac{1}{2})^m$.

Let $(m - 2)$ elements be positive and 2 be negative. Then the resultant is $[(m - 2 \cdot \Delta x - 2\Delta x)]$, or $(m - 2.2)\Delta x$ and the probability is $\dfrac{m(m-1)}{1.2}$, hence $y = \dfrac{m(m-1)}{1.2} \cdot (\frac{1}{2})^m$.

On extending the foregoing reasoning, we have the general term,

$$x = (m - 2n)\ \Delta x,$$

$$y = \frac{m\ (m-1)\ (m-2 \ldots\ldots (m-n+1)}{1\ .\ 2\ .\ 3 \ldots\ldots n} \cdot (\tfrac{1}{2})^m$$

The coordinates of the next following point are

$$x' = (m - 2n - 2)\ \Delta\ x$$

$$y' = \frac{m(m-1)\ (m-2) \ldots (m-n+1)\ (m-n).\ (\frac{1}{2})^m}{1\ .\ 2\ .\ 3 \ldots\ldots n\ .\ (n+1)}$$

The ratio of the two y's is,

$$\frac{y'}{y} = \frac{m-n}{n+1}$$

Subtracting both sides from 1, we have

$$1 - \frac{y'}{y} = 1 - \frac{m-n}{n+1}.$$

On reducing,

$$\frac{y - y'}{y} = \frac{n + 1 - m + n}{n+1} = \frac{2n - m + 1}{n+1}$$

On subtracting x's, we have,

$$\begin{aligned} x - x' &= (m - 2n)\ \Delta x - (m - 2n - 2)\ \Delta x \\ &= (m - m - 2n + 2n + 2)\ \Delta x \\ &= 2\ .\ \Delta x. \end{aligned}$$

On substituting for x' we have,

$$x = m\ .\ \Delta x - 2n\ \Delta x.$$

On solving we have,

$$n = \frac{m \,.\, \Delta x - x}{2 \Delta x}.$$

Hence,

$$2n - m + 1 = \frac{m \,.\, \Delta x - x}{\Delta x} - m + 1$$

$$= \frac{m \,.\, \Delta x - x - m \Delta x + \Delta x}{\Delta x}$$

$$= \frac{\Delta x - x}{\Delta x},$$

and

$$n + 1 = \frac{m \,.\, \Delta x - x}{2 \,.\, \Delta x} + 1$$

$$= \frac{m \,.\, \Delta x - x + 2 \,.\, \Delta x}{2 \,.\, \Delta x}$$

$$= \frac{(m+2) \,.\, \Delta x - x}{2 \,.\, \Delta x}$$

On substituting, in $\dfrac{y - y'}{y} = \dfrac{2n - m + 1}{n + 1}$

we have,

$$\frac{y - y'}{y} = \frac{\Delta x - x}{\Delta x} \cdot \frac{2 \Delta x}{(m+2) \Delta x - x}$$

$$= \frac{2 (\Delta x - x)}{(m+2) \; \Delta x - x}$$

On further combining and dividing by $x - x'$, we have,

$$\text{(A)} \qquad \frac{y - y'}{x - x'} = \frac{y \,.\, 2(\Delta x - x)}{(m+2) \Delta x - x} \cdot \frac{1}{2 \Delta x}.$$

This latest expression should be carefully noted even if the reader may not care to follow the intermediate expressions in detail.

Let us now assume that there have been a large number of the elements, and that these elements are assumed to be smaller and smaller. In other words, let m become large and Δx become small. Now, in (A), Δx is negligible in comparison with x and 2 is negligible in comparison with m while $m \,.\, \Delta x$ is the maximum possible value which is very large in comparison with x.

Hence (B)

$$\frac{y - y'}{x - x'} = y \cdot \frac{-2x}{2m\Delta x^2} = \frac{-yx}{m\Delta x^2}$$

Expression (B) means that the difference in consecutive y's divided by the difference in the corresponding x's is equal to the negative product of x and y divided by the maximum positive value of the x, times Δx.

Let us set down $\frac{1}{m\overline{\Delta x^2}} = 2h^2$.

Then (C) $\frac{y - y'}{x - x'} = -2h^2 \cdot yx$.

Now (C) is true regardless of how close are the two points. Hence let the points become closer and closer and designate the limit of $y - y'$ by dy and of $x - x'$ by dx. Then we have

$$\frac{dy}{dx} = -2h^2 yx$$

or, (D) $$\frac{dy}{y} = -2h^2 x\, dx.$$

It is an elementary principle of the calculus that expression (D) is equivalent to

(E) $\log y = -h^2 \cdot x^2 + k'$,

where k is constant and the logarithm is to base e.

On passing from logarithms to number, we have,

$$y = e^{-h^2x^2} + k = e^{-h^2x^2} \cdot e^{k'}$$

or (F) $y = K \cdot e^{-h^2x^2}$, since $e^{k'}$ may be taken equal to the constant K (where $^{h^2x^2}$ is the exponent of e).

We finally have in (F) the normal probability curve,

$y = K \cdot e^{-h^2x^2}$, which is of the same form as,

$$y = \frac{N}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \frac{x^2}{\sigma^2}} \text{ where}$$

$$K = \frac{N}{\sigma\sqrt{2\pi}} \text{ and } h^2 = \frac{1}{2\sigma^2}.$$

The derivation of the forms

$K = \frac{N}{\sigma\sqrt{2\pi}}$ and $h^2 = \frac{1}{\sigma^2}$ proceeds directly from the assumption that the area under the curve must equal the total frequency N, and that the standard deviation of the ordinates under the curve is equal to the standard deviation computed from the observed points. The actual derivation which requires a further knowledge of integral calculus may be found in Elderton: "Frequency Curves and Correlation."

No attempt is made here to prove that the constant is actually equivalent to $N/\sigma \, . \, \sqrt{\pi}$ nor that

$$h^2 . = \tfrac{1}{2}\,\sigma^2.$$

There are various ways of presenting these two proofs but all of which are so detailed that it seems best to leave the actual derivation to special references as those interested may consult the various writings of Karl Pearson, or Merriman "Methods of Least Squares."

Even though the demonstration given here might be somewhat lengthy *the point is that the equation follows directly from the basic statement of the normal probability curve.*

APPENDIX II

Introduction. The generalized frequency curves of Pearson are so diverse in shape that a curve of this class can be found to fit any ordinary statistical distribution. By the following methods the fitting of a Pearson curve is reduced almost entirely to a matter of routine substitution in formulas, so that the practical statistician can make extended use of the curves without great familiarity with their theory.

This discussion is designed both to present the working methods of the generalized frequency curves and to give the statistician who has a minimum of acquaintance with the higher mathematics some degree of familiarity with the underlying theory. The demonstrations are, for the most part, omitted.

In developing the theory of the generalized frequency curves it is logical as well as practically convenient, to start with the normal curve and consider the *general distribution as a modification of the normal type of distribution.*

The Slope Property. The particular modification which leads to the frequency curves of Pearson is obtained by generalizing the slope condition of the normal curve. The slope of a curve at a given point is the tangent of the angle which the line touching the curve at that point makes with the X-axis. In the case of the normal curve, the ratio of the slope to the ordinate is negatively equal to the abscissa of the point, as is shown in the derivations of Appendix I.

This slope property is generalized by taking the ratio equal, not to $-x$ but to $-\dfrac{x+a}{b+cx+dx^2}$ where a, b, c, d, are constants. The slope of a curve is ordinarily denoted by the symbol $\dfrac{dy}{dx}$ and hence, we have the following equation

$$\frac{1}{y}\,\frac{dy}{dx} = -\frac{x+a}{b+cx+dx^2} \, .$$

The Constants, a, b, c, d. The statistical significance of each of the constants, *a, b, c, d,* can be readily determined.

In Chapter IV, it is shown that the slope of a frequency curve is zero at a mode. Since $\frac{dy}{dx}$; that is, the slope, is **zero** when $x = -a$, the constant a determines the position of the mode. The mode is therefore at a distance, $-a$, from the mean. As explained in Chapter V, a is thus a measure of the skewness, of the lack of symmetry of the distribution. For a symmetrical distribution a is evidently o.

When both c and d are zero the generalized slope equation is merely the normal slope equation with x replaced by $\frac{x+a}{b.}$.

This leads to the normal curve,

$$y = k\,.\,e^{-\frac{(x+a)^2}{b}}, \text{ where } k \text{ is a constant.}$$

Comparing this equation with the standard normal equation,

$$y = k\,e^{-\frac{x^2}{2\sigma^2}},$$

we see that b equals $2\sigma^2$ multiplied by a constant.

The degree of symmetry of the curve is indicated by the value of c as well as by the value of a. For, when x is positive, the term cx is added in the denominator and when x is negative it is subtracted. This tends to make the frequency curve steeper to the left than to the right of the origin, and hence the curve must extend farther to the right, that is, the curve must be skew.

But it was seen in Chapter V, that β_1 is the fundamental measures of skewness. Therefore both a and c must contain β_1 as a factor.

When x^2 is small the constant d has little effect on the slope, but for the extremities of the curve where x and hence $d\,x^2$ is large the slope is reduced by a large value of d. It will be seen that d depends largely on β_2.

The Types of Curves. We may now discuss the distinct types of curves that possess the slope properties of the generalized slope equation. *Distinct types of curves result according as the denominator, $b + cx + dx^2$, has two distinct factors, two coincident factors, or has no factors.*

With two distinct factors the slope equation can be written

$$\frac{1}{y}\,\frac{dy}{dx} = -\frac{x+a}{b+cx+dx^2} = k\,.\,\frac{x+a}{(r + x)\,(r^2 - x)}$$

where k is a constant.

By the usual mathematical methods we then have

$$y = y'\;(r_1 + x)\;\frac{k(a - r_1)}{r_1 + r_2}\;.\quad (r_2 - x)\;\frac{-k(a + r_2)}{r_1 + r_2}\qquad \text{(A)}$$

where y' is the constant of integration.

By a simple transformation and rearrangement, this equation can be reduced to the form of Pearson's first type, namely:

$$y = y_0\left(1 + \frac{x}{a_1}\right)^{m_1}\left(1 - \frac{x}{a_2}\right)^{m_2}. \qquad \textbf{Type I.}$$

Exercises.

1. Carry through in detail the necessary transformations to determine the equation of Type I from equation (A).
2. Perform the integrations to obtain the curve of Type I.

When a_1 and a_2 are equal it is readily shown that $m_1 = m_2$ and the equation takes the form of Type II:

$$y = y_0\left(1 - \frac{x^2}{a^2}\right)^{m} \qquad \textbf{Type II.}$$

When one root of the denominator $b + cx + dx^2$ is indefinitely large, that is, when d is zero, we have, from the theory of the exponential e, the third type:

$$y = y_0 e^{-\gamma x}\left(1 + \frac{x}{a}\right)^{\gamma a}. \qquad \textbf{Type III.}$$

This equation may be looked upon as that of Type I with a_2 indefinitely large.

The curves of Type III are especially serviceable because the equations are simple in form and convenient for computation. They are the most elementary skew curves.

By transforming expression (A), in a manner somewhat different from that to obtain Type I, the form of Pearson's sixth type is readily obtained. It is

$$y = y_0 \ (x - a)^{m_2} x^{-m_1}. \qquad \textbf{Type VI.}$$

Exercises.

3. Obtain the equation of Type II by direct integration from the differential equation.
4. Compare Type II with the normal curve.
5. Obtain Type III directly by integration.
6. Obtain Type III from (A).
7. Compare the shape of Type III with that of the normal curve.
8. Obtain the equation of Type VI directly from the differential equation.
10. Is Type VI geometrically distinct from Type I?

When **two roots are indefinitely large** we have the normal curve:

$$y = y_0 \ e^{-\frac{x^2}{2b}}.$$

which is called simply "Normal" in Pearson's scheme of classification.

With two coincident roots, the slope equation becomes

$$\frac{1}{y} \cdot \frac{dy}{dx} = k \ \frac{x + a}{(x + r)^2}$$

This leads to the form $y = y_0 x^{-p} e^{-\frac{\gamma}{x}}$, Type V.
which is Pearson's type V.

Exercises.

11. Derive in detail the equation of Type V.

When the denominator of the slope equation cannot be factored the integration is performed by writing

$$\frac{1}{y} \cdot \frac{dy}{dx} = - \frac{x+a}{b+cx+dx^2},$$

$$= - \frac{x + \frac{c}{2d} + a - \frac{c}{2d}}{d\left\{x^2 + \frac{c}{d}x + \frac{c^2}{4d^2} + \frac{b}{d} - \frac{c^2}{4d^2}\right\}}.$$

This gives

$$y = y_0\left(1 + \frac{x^2}{a^2}\right)^{-m} e^{-\nu \tan^{-1}\frac{x}{a}} \qquad \text{Type IV.}$$

which is the form of Type IV.

Exercises.

12. Derive in detail the equation of Type IV.

13. Derive the equation of Type IV by transformation from the equation of Type I.

14. Compare the form of the equation of Type IV to that of Type III.

If ν is zero in the immediately preceding equation we have Pearson's Type VII.

$$y = y_0\left(1 + \frac{x^2}{a^2}\right)^{-m} \qquad \text{Type VII.}$$

The Intercepts. The intercepts made on the X-axis by the various types of curves can now be examined. The ordinate y Type I is zero when $x = -a_1$ or $x = a_2$ hence it can be proven mathematically that there can be no real values of y beyond these two intercept points. That is, a Type I curve stops at the points on the X-axis at distances $-a_1$ and a_2 from the mean.

In Type II the intercepts are of the same length and numerically equal to a, so that a Type II curve, like the curves of Type I is a limited curve. Unlike the curves of the first type, Type II curves are symmetrical, that is, have zero skewness.

In Type III one intercept is $-a$ and the other is indefinitely large, thus these curves are limited to the left and indefinitely extended to the right. It should be taken as the type of a sort of basic skew curve.

In Types IV and VII there are no intercepts, because there are no values of x which reduce y to zero.

In Type V one intercept passes through the origin and the other is indefinitely large.

In Type VI both intercepts are positive or both are negative.

Ordinarily the type of curve selected should have intercepts harmonizing with the natural limits of the range of the data. For instance, data necessarily limited in either direction should be smoothed with a curve correspondingly limited. However, nearly all the curves are practically limited in range because the ordinates soon become negligible, so that the matter is not one of great importance; although a somewhat better fit is likely to be obtained with a curve limited in accordance with the data.

Exercise

15. Of what type is the normal curve a limiting curve?

16. Distinguish between a curve with indefinitely large intercepts and a curve with imaginary or non-existent intercepts.

17. Show that there are indefinitely more curves of Type I, VI and IV than of Types III, V, II or VII, or of the normal curve.

18. Show how Type I can be said algebraically to include Type IV.

19. Show that Types I and VI are not fundamentally distinct.

20. Show that by taking all combinations of sign into account there are three distinct classes of curve under Type I.

21. Show that there are two sub-classes under Type II according as the exponent m is positive or negative.

22. Show that there are two classes under Type III.

23. Is there more than one general form of curve under Type IV? Under Type V?

24. Discuss the curves of Type VI as to the existence of sub-classes within the Type.

25. What types of these curves have asymptotes?

26. Do all the curves have a mode?

27. Find the point of inflexion fo reach type.

The Criterion K. Since the separation into types depends primarily on the nature of the roots of the quadratic, $b + cx + dx^2$, the discriminant of this quadratic constitutes a criterion of the type of curve which fits the distribution. The values of

a, b, c and d are first determined by the method of moments and then the discriminant expressed in terms of the computed expressions for b, c and d.

The formula for K, the discriminant obtained in this way is

$$K = \frac{\beta_1 (\beta_2 + 3)^2}{4(2\beta_2 - 3\beta_1 - 6)\ (4\beta_2 - 3\beta_1)}$$

The Value of K and the Type of Curve. The following table gives the types of curves corresponding to the different values of K.

$K < o$,	i. e. negative	Type I.
$K = o$,	$\beta_1 = o,\ \beta_2 = 3$	Normal Curve.
	$\beta_1 = o,\ \beta_2 < 3$	Type II.
	$\beta_1 = o,\ \beta_2 > 3$	Type II.
$K > o$	< 1	Type IV.
$K =$	1	Type V.
$K >$	1, but not indefinitely large,	Type VI.
K	indefinitely large	Type III.

It is to be noted that the types of curve for any given statistical distribution can now be determined by strictly arithmetic methods.

The only restriction on the generality of the theory of the criterion K is that the quantity $x^n\ (b + cx + dx^2)\ y$ must vanish at both ends of the range. This condition marks the pairs of values of β_1 and β_2 for which no curve of the generalized differential equation can be found. The limiting values of β_1 and β_2 are $\beta_2 > \frac{3}{4}\beta_1$ and $\beta_2 > \beta_1/8 + 9/2$.

The Computation Formulas. The computation formulas for the several types of Pearson's frequency curves are derived in accordance with the method of moments. For each type as many moment equations are written as there are constants in the equation of a curve of the type. In some of the type equations, as in Type I where $a_1/m_1 = a_2/m_2$, the constants are connected by equations so that the number of moment equations is reduced. *The moment equations are the result of equating the theoretical moments of the curve obtained by integration to the moments computed directly from the data.*

It might be expected that the differential equation would be integrated to give the equations directly, but the present process is more convenient. The chief purpose, therefore, of the slope or differential equation is for the determination of the type forms of the equations. After the algebraic forms of the equations are determined each type is worked out without making use of its connection either with the slope equation or with other type forms.

The expression $\Gamma\ (p)$, called the *gamma function,* occurs in the following formulas. This function is defined by the relation

$\Gamma\ (p) = (p-1)\ \Gamma\ (p-1).$

If p is an integer, $\Gamma\ (p) = \underline{|p-1}.$

If p is not an integer, $\Gamma\ (p) = (p-1)\ (p-2)\ \ldots\ (p-p+2)\ \Gamma\ P$ where P is the remainder after subtracting a sufficient number of 1's to bring p down to between 2 and 1 in value. The values of $\Gamma\ (P)$ are given in Table XXXI of "Tables for Statisticians and Biometricians."

The probable errors of K as well as of β_1 and β_2 are given in "Tables."

The derivation of the following computation formulas, except the moment formulas, is not possible without an extensive acquaintance with the calculus.

After the constants in the equation are composed the smoothed frequencies are obtained by computing the areas under the curve and between the bounding ordinates. Thus the frequency of the first class is the area between the ordinate $x = \frac{1}{2}$ and $x = 1\frac{1}{2}$. Simpson's quadrature formula is ordinarily used for finding the class area. According to this formula the area is $1/6\left\{y_{x-\frac{1}{2}} + 4y_x + y_{x+\frac{1}{2}}\right\}$ where $y_{x-\frac{1}{2}}$ and $y_{x+\frac{1}{2}}$ are the bounding ordinates and y_x is the mid-ordinate of the class.

Formulas for the Moments.

$S_2 = d.$

$\nu_2 = 2S_3 - d\,(1 + d)$

$\nu_3 = 6S_4 - 3\nu_2\,(1 + d) - d\,(1 + d)\,(2 + d)$

$\nu_4 = 24S_5 - 2\nu_3 \{2(1 + d) + 1\} - \nu_2\{6(1 + d)\,(2 + d) - 1\}$
$- d\,(1 + d)\,(2 + d)\,(3 + d).$

$\mu_2 = \nu_2 - \frac{1}{12}$

$\mu_3 = \nu_3$

$\mu_4 = \nu_4 - \frac{1}{2}\nu_2 + \frac{7}{240}$

$\sigma = \sqrt{\mu_2}$

$\beta_1 = \mu_3{}^2 \div \mu_2{}^3$

$\beta_2 = \mu_4 \div \mu_2{}^3$

$$K = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)\,(2\beta_2 - 3\beta_1 - 6)}$$

The computation formulas for Type I are as follows:

The equation is,

$$y = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 - \frac{x}{a_2}\right)^{m_2}$$

where $a_1/m_1 = a_2/m_2$.

We have

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{3\beta_1 - \beta_2 + 6}$$

$$\epsilon = \frac{4r^2}{16(r+1) + \beta_1(r+2)^2}$$

$$b^2 = \frac{\mu_2(r+1)r^2}{\epsilon}.$$

m_2 and m_1 are given by the formulas

$$\tfrac{1}{2}(r-2) \pm \tfrac{1}{4}(r+2)\sqrt{\beta_1\epsilon}.$$

The constant m_1 is taken with the negative root when μ_3 is positive and with the positive root when μ_3 is negative.

$a_1 + a_2 = b.$

a_1 and a_2 can be found from the relations $a_1 + a_2 = b$ and $a_1/m_1 = a_2/m_2$.

$$y_0 = \frac{N}{b}\,\frac{m_1{}^{m_1} m_2{}^{m_2}}{(m_1+m_2)^{m_1+m_2}}\,\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}$$

$$\text{The skewness is } \tfrac{1}{2}\sqrt{\beta_1}\left\{\frac{r+2}{r-2}\right\}$$

$$\text{Mode} = \text{mean} - \tfrac{1}{2}\frac{\mu_3}{\mu_2}\left\{\frac{r+2}{r-2}\right\}$$

The formulas for Type II are as follows. The equation for this type is

$$y = y_0\left(1 - \frac{x^2}{a^2}\right)^m$$

The formulas are

$$m = \frac{5\beta_2 - 9}{2(3 - \beta_2)}$$

$$a^2 = \frac{2\mu_2\beta_2}{3 - \beta_2}$$

$$y_0 = \frac{N\,\Gamma(2m+2)}{a.2^{2m+1}\{\Gamma(m+1)\}^2}$$

Type III. The equation is

$$y = y_0\,e^{-\gamma x}\left(1 + \frac{x}{a}\right)^{\gamma a}.$$

The formulas are,

$$\gamma = 2\frac{\mu_2}{\mu_3},$$

$$a = \mu\gamma - \frac{1}{\nu},$$

$$y_0 = \frac{N}{a}\,\frac{p^{p+1}}{e^p\Gamma(p+1)},\ \text{where } p = \nu a.$$

$$\text{Mode} = \text{mean} - \frac{1}{\gamma}$$

$$\text{Skewness} = \frac{1}{\sigma\gamma}$$

Type IV. The equation is

$$y = y_0\left(1 + \frac{x^2}{a^2}\right)^{-m} e^{-\nu\tan^{-1}\frac{x}{a}},$$

The formulas are:

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6}$$

$$m = \tfrac{1}{2}(r+2),$$

$$e = \frac{4r^2}{16(r-1) - \beta_1(r-2)^2},$$

$$\nu = \tfrac{1}{2}(r-2)\sqrt{\beta_1 e,}$$

$$a = \frac{1}{2}\,\frac{r\sigma}{\sqrt{e}}.$$

$$y_0 = \frac{N}{a}\sqrt{\frac{r}{2\pi}}\ \frac{e^{\frac{\cos^2\phi}{3r} - \frac{1}{12r} - \nu\phi}}{(\cos\phi)^{r+1}}\ , \text{ where } \tan\phi = \frac{\nu}{r}.$$

$$\text{Origin} = \text{mean} + \frac{\nu a}{r}$$

$$\text{Mode} = \text{mean} - \frac{1}{2}\frac{\mu_3(r-2)}{\mu_2(r+2)}$$

Type V. The equation is

$$y = y_0 x^{-p} e^{-\gamma/x}$$

The formulas are:

$$p = 4 + \frac{8 + 4\sqrt{(4+\beta_1)}}{\beta_1}.$$

$\gamma = (p-2)\sqrt{\mu_2(p-3)}$, with sign same as that of μ_3.

$$y_0 = \frac{N\gamma^{p-1}}{\Gamma(p-1)}.$$

$$sk. = \frac{2\sqrt{p-3}}{p}.$$

$$\text{Origin} = \text{mean} - \frac{\gamma}{p-2}.$$

$$\text{Mode} = \text{mean} - \frac{2\gamma}{p(p-2)}.$$

Type VI. The equation is

$$y = y_0 (x-a)^{q_2} x^{-q_1}.$$

The formulas are:

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{6 + 3\beta_1 - 2\beta_2}.$$

$$e = \frac{4r^2}{16(r+1) + \beta_1(r+2)^2}.$$

$$1 - q_1 = -\frac{r}{2} + \frac{r+2}{4}\sqrt{e\beta_1},$$

$$1 + q_2 = -\frac{r}{2} - \frac{r+2}{4}\sqrt{e\beta_1},$$

$$a = \frac{r\sigma}{\sqrt{e}}.$$

$$y_0 = \frac{Na^{q_1 - q_2 - 1}\,\Gamma(q_1)}{\Gamma(q_1 - q_2 - 1)\,\Gamma(q_2 + 1)}.$$

$$\text{Origin} = \text{mean} - \frac{a(q_1 - 1)}{q_1 - q_2 - 2}.$$

$$\text{Mode} = \text{mean} - \frac{1}{2}\,\frac{\mu_2}{\mu_3}\cdot\frac{r + 2}{r - 2}$$

Normal Curve. The equation, as was proved in Chapter VI, is

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

and the curve was discussed in that chapter.

APPENDIX III

MAKEHAM'S LAW OF MORTALITY

Makeham's Law of Mortality. An application of the idea of rates under Makeham's Law of Mortality is of interest as an illustration of the general idea of the geometric mean. It is also of interest as showing how comparatively simple ideas may be developed into a powerful mathematical formula. This formula states a trend in precise mathematical language.

The forces which bring about death may be looked upon as belonging to two classes. One class may be called the *accident* class, made up of purely accidental forces which affect the young and old alike. The totality of all such forces can then be represented by a constant, A, let us say.

The second class of forces may be thought of as the totality of the wear and tear, as the result of the degenerative diseases. These latter forces have been illustrated by the action of an air pump where each stroke removes a fixed percentage of the air. Under these progressive decreases in the force of life each year takes away a fixed percentage of the ability to withstand death.

Under this fixed ratio of decline idea, let us denote the constant ratio by the letter, c. Then the cumulative force according to age will be the result of applying this ratio once for each year, that is, c^x. This latter expression, c^x, is only the cumulative ratio and hence should be multiplied by a constant which we may denote by B.

Bringing all these expreessions together we have $A + Bc^x$ as the force of mortality.

At this point it is necessary to bring in the usual expression for the force of mortality without reference to the law of mortality. The symbol μ_x is found in actuarial literature for the force of mortality. If the symbol l_x denotes the number living at age x then the number dying may be denoted by Δl_x. Assume these deaths take place in Δt years, then the number dying in one year would be $\frac{\Delta l_x}{\Delta t}$. If 100 deaths occur in one month there would be $100/\frac{1}{12}$ or 1200 deaths in one year. These yearly deaths must be expressed as a percentage of the number living to get an average death rate; this gives $-\frac{l}{l_x} \cdot \frac{\Delta l_x}{\Delta t}$. The negative sign is used because all the decrements in l_x are negative.

In the expression $\frac{1}{l_x} \cdot \frac{\Delta l_x}{\Delta t}$ let us think of the interval of time as becoming smaller and smaller when, in the language of the calculus, this expression becomes $-\frac{1}{l_x} \frac{dl_x}{dt}$. That is, we have, $\mu_x = -\frac{1}{l_x} \frac{dl_x}{dt}$.

To return to the derived constant form for the force of mortality we have,

$$\mu_x = -\frac{1}{l_x} \cdot \frac{dl_x}{dt} = A + Bc^x .$$

From here on it is necessary to have an elementary knowledge of the calculus. Those who are not familiar with the solution of differential equations can accept the statement that the foregoing equation leads to:

$$\log_e l_x = -Ax - \frac{Bc^x}{\log_e c} + \log_e k,$$

Where A, B and c are our previously defined constants and $\log_e k$ is a constant of integration.

Let us set $-A = \log_e s$ · and $-B/\log_e c = \log_e g$

Then $\log_e l_x = \log_e k + x \log_e s + c^x \log_e g$

$$\text{or } l_x = ks^x g^{c^x}$$

which is Makeham's equation of mortality.

It should be noted the Makeham's equation is true to the extent that our basic assumptions as to the classifications of the forces producing death are true.

INDEX.