An Introduction to
MODERN STATISTICAL METHODS

BY

PAUL R. RIDER

Washington University
Saint Louis

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED

THIRD PRINTING, MARCH, 1947

PRINTED IN U. S. A.

AN INTRODUCTION TO MODERN STATISTICAL METHODS

PREFACE

In the past three decades enormous advances have been made in the field of statistics. These advances have taken place so rapidly that it is not at all surprising that those employing statistical methods have found it difficult to keep pace with them, or that certain of the older methods, which are obsolete, and even in some cases erroneous—or, at the very best, crudely approximative in character—continue to be taught in the classroom and to be treated in textbooks which appear from time to time. A notable example is the use of the probable error or standard error in testing the significance of a correlation coefficient derived from a sample, although this method gives unreliable or incorrect results if there is a high degree of correlation in the population from which the sample is drawn, or if the number in the sample is small.

It is, of course, in the theory of small samples that the greatest progress has been made. The theory of sampling which assumes that the sample is composed of a large number of items is inadequate for many practical purposes. Biological, agricultural, and other scientific experiments frequently deal with comparatively few observations. Sometimes the cost of obtaining additional observations is prohibitive; sometimes, indeed, it is impossible to obtain more data, as might be true in the case of meteorological records. In manufacturing inspection, too, small samples are of frequent occurrence.

Most of this theory has been developed and unified by R. A. Fisher, who has shown how to make more accurate estimates and how to utilize the maximum amount of information contained in a set of data, and has provided exact tests of reliability and significance. Fisher's efficient methods, at first slow in taking hold, because not thoroughly understood, gradually began to gain momentum, and are now spreading rapidly.

In this book I have endeavored to explain the most widely used of these methods, illustrating their application by com-

paratively simple numerical examples, so that the underlying principles are not lost sight of in a maze of arithmetical computations. The earlier chapters develop the fundamental concepts of statistics, so that the book is suitable as a textbook for a first course in the subject. It is also planned for those with some knowledge of the subject who wish to gain an insight into the more modern methods, as it leads from the classical concepts, through such topics as "Student's" distribution and the chi-square distribution, to the analysis of variance and the design of experiments, which are the culminating features of Fisher's work.

Grateful acknowledgment is made to Professor Fisher, and to his publishers, Messrs. Oliver and Boyd, for their generous permission to reproduce, from "Statistical Methods for Research Workers," the tables of t, chi square, and the 5 per cent and 1 per cent points of the distribution of z.

I am deeply indebted to Dr. Churchill Eisenhart for a critical reading of the manuscript, and for many valuable suggestions. However, full responsibility for errors is, of course, my own.

I am also indebted to various persons and various sources for material used in the exercises, being particularly grateful to Messrs. A. G. Brooks and J. B. Gibson for supplying, and for granting permission to use, certain data of the Western Electric Company.

PAUL R. RIDER

WASHINGTON UNIVERSITY
SAINT LOUIS
September, 1938

CONTENTS

INTRODUCTION TO MODERN STATISTICAL METHODS

CHAPTER I

FREQUENCY DISTRIBUTIONS

1. Frequency tables. A *frequency table* is a table classifying a set of observations according to the numbers of them which fall within certain limits. It is a tabular method of exhibiting a *frequency distribution.* For example, Table 1 classifies the heights of a group of men. The table shows the frequency with which men of a given height, or rather between two given limits of height, occur in the group of 346 men under consideration. The values 58 inches, 60 inches, 62 inches, etc., are the *class limits,* and the difference between two consecutive class limits, here 2 inches, is the *class interval.* The mid-values of the classes are obviously 59 inches, 61 inches, etc. The *range* of the table, from 58 inches to 74 inches, is 16 inches.

TABLE 1

FREQUENCY TABLE OF HEIGHTS OF A GROUP OF MEN

Height in inches	Number of men within given limits of height (frequency)
58–60	1
60–62	2
62–64	9
64–66	48
66–68	131
68–70	102
70–72	40
72–74	13
Total . . .	346

A word regarding classification and class limits may be worth while at this point. It is assumed that in the construction of Table 1 the measurements of height have been made to a sufficient degree of fineness that no doubt exists regarding the class to which a man belongs. If, however, we had a set of

data in which measurements were made to the nearest inch, we could not employ the same class limits as those used in Table 1. For a height recorded as 62 inches would simply mean that the measurement was between 61.5 and 62.5 inches. In such a case, if we wished to use a 2-inch interval, we could set the classes as 57.5–59.5, 59.5–61.5, . . . The mid-values of these classes would be 58.5, 60.5, . . .

Some question arises as to what disposition to make of an observation or measurement which falls exactly on a class boundary. For example, if the classes are as in Table 1 and we have a measurement of 62 inches, it is not clear whether this should be assigned to the class 60–62 or to the class 62–64. In such an instance there are certain theoretical advantages in dividing the unit of frequency between the two classes, and assigning ½ to each class.

Difficulties in the classification of raw data can usually be avoided by a proper choice of class limits.*

2. Cumulative frequency tables. The above frequency distribution is equally well specified if we know the number of men below (or above) any given height, for if we know that there are 60 men below 66 inches in height and 12 below 64 inches, we can find at once that there are 60 − 12, or 48, men who are between 64 and 66 inches tall. If we convert Table 1 into a table showing the number of men below certain heights, we obtain the cumulative frequency table, Table 1A.

3. Continuous and discrete variables. The variable in the foregoing example is height. Theoretically it can be measured to any degree of fineness. Such a variable is called *continuous*. There are, however, variables which can have only integral values. Such variables are called *integral* or *discrete*. Examples of such variables are the number of petals on flowers (see Table 2), the number of spots obtained in throwing dice, the number of heads obtained in tossing coins, the number of children in a family.

* For a good discussion of these and related questions, see Yule and Kendall, "An Introduction to the Theory of Statistics," Charles Griffin & Co., Ltd., London, 1937.

TABLE 1A

CUMULATIVE FREQUENCY TABLE OF HEIGHTS

Height in inches	Number of men below specified height (cumulative frequency)
60	1
62	3
64	12
66	60
68	191
70	293
72	333
74	346

TABLE 2

FREQUENCY TABLE OF NUMBERS OF PETALS ON A CERTAIN SPECIES OF FLOWER

Number of petals	Number of flowers having a specified number of petals (frequency)
5	133
6	55
7	23
8	7
9	2
10	2
Total....	222

4. Graphic representation of frequency distributions. In the case of continuous variables, the accepted method of representing

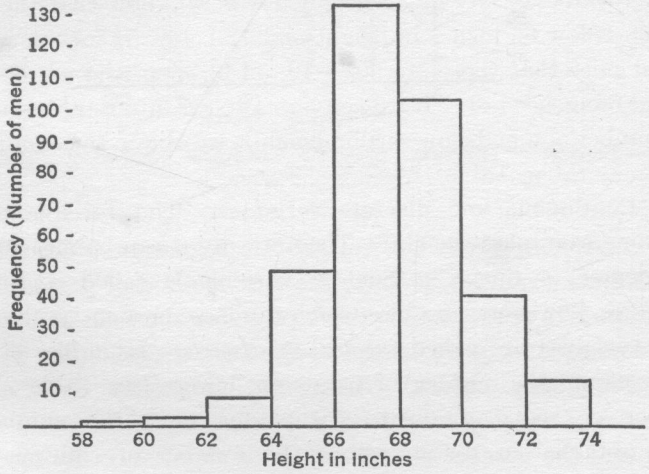

FIG. 1.—Histogram of Heights of a Group of Men.

a frequency distribution graphically is by means of a *rectangular frequency diagram* or *histogram*. This is constructed by marking

off a scale for the variable and erecting, at the appropriate positions on this scale, rectangles whose areas are equal or proportional to the respective class frequencies (See Fig. 1.) In the usual case of class intervals of equal size, such rectangles will obviously have heights equal or proportional to the respective class frequencies.

A cumulative frequency distribution can be represented by plotting points with ordinates equal to the cumulated frequencies, and with abscissas equal to the upper limits of the classes, and

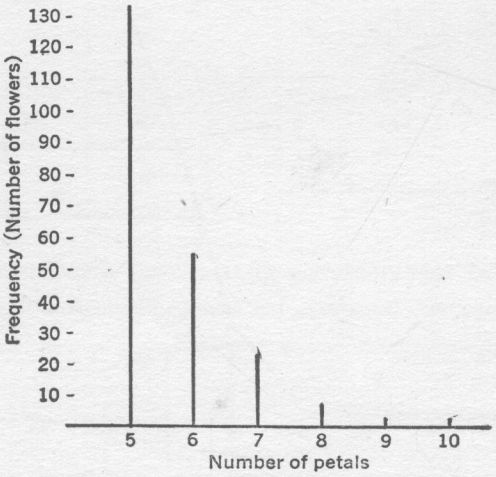

FIG. 2.—Frequency Diagram. Discrete Variable—Number of Petals on a Species of Flower.

then connecting these by straight-line segments. (See Fig. 3.) For discrete variables, perhaps the best method of graphic representation is to erect, at the proper places on the scale or base line, ordinates equal or proportional to the frequencies. (See Fig. 2.)

5. Frequency curves. Theoretical frequency distributions. If the size of the class interval of a distribution be decreased indefinitely and the number of individuals be simultaneously increased indefinitely, the histogram approaches a *frequency curve.* A frequency curve may be regarded as representing an idealized frequency distribution.

Suppose that the equation of the frequency curve is $Y = f(X)$.

The function $f(X)$ can always be multiplied by a constant factor which will make the area under the curve equal to unity, and we shall assume that this has been done. Then we have

$$\int_{-\infty}^{\infty} f(X)dX = 1 \tag{1}$$

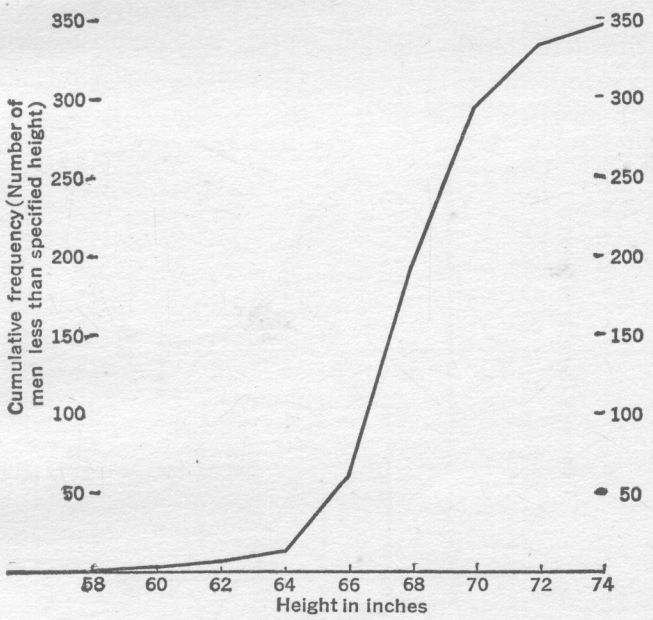

FIG. 3.—Cumulative Frequency Diagram.

(If the curve does not actually extend to infinity, but meets the X-axis in some point such as X_0 in Fig. 4, the value of $f(X)$ is to be regarded as zero outside of such points. Consequently we can always use $-\infty$ and ∞ as limits of the above definite integral.) The proportion of items between the values $X = a$ and $X = b$, $a < b$, is the shaded area in Fig. 4, and, provided (1) holds, is given analytically by

$$\int_{a}^{b} f(X)dX \tag{2}$$

By the artifice of defining $f(X) = 0$ outside the range of the curve,

we see that we have defined $f(X)$ so that (2) gives the proportion of the area under the curve lying between *any* two values $X = a$ and $X = b$, $a < b$. This is the exact mathematical interpretation of $f(X)$. As an aid to intuition, it is often convenient to regard $f(X)dX$ as giving approximately the proportion of items in the interval between X and $X + dX$, the closeness of the approximation depending on how small dX is. It is desirable to write the equation of a frequency curve in the form $YdX = f(X)dX$, because if any transformation is made in the variable X it must also be made in the differential dX. (See next paragraph.)

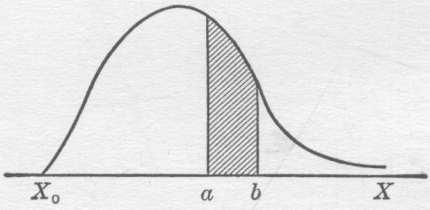

FIG. 4.—Frequency Curve.

The most widely used frequency curve is the *normal* curve, whose equation may be written

$$YdX = (2\pi)^{-\frac{1}{2}}e^{-(X-\mu)^2/2\sigma^2}\frac{dX}{\sigma} \tag{3}$$

If we make the transformation

$$\frac{X-\mu}{\sigma} = x, \quad \frac{dX}{\sigma} = dx$$

(3) goes into the simpler form

$$YdX = (2\pi)^{-\frac{1}{2}}e^{-x^2/2}dx \tag{4}$$

Another important curve is the *Pearson type III* curve

$$YdX = \frac{1}{k!}X^k e^{-X}dX, \quad 0 \leqq X < \infty \tag{5}$$

The symbol $k!$, called "k factorial," is defined by

$$k! = \int_0^\infty X^k e^{-X}dX = \Gamma(k+1) \tag{6}$$

If k is an integer, this reduces to $k(k-1)\ldots 3.2.1$.

Examples of ideal or theoretical discrete frequency distributions are the *binomial* distribution

$$Y = \frac{N!}{X!(N-X)!} p^X (1-p)^{N-X}, \quad 0 < p < 1, \tag{7}$$

$$X = 0,1,2, \ldots, N$$

and the *Poisson exponential* distribution

$$Y = \frac{e^{-\mu}\mu^X}{X!}, \quad X = 0,1,2, \ldots \tag{8}$$

EXERCISES

1. The frequency in a class of Table A is the number of students receiving grades between the limits indicated. (*a*) Draw a histogram for the data of this table. (*b*) Construct a cumulative frequency table from the data. (*c*) Draw a cumulative frequency diagram.

2. Table B gives the number of men of a certain group whose weights fall within specified limits. (*a*) Draw a histogram for this frequency table. (*b*) Construct a cumulative frequency table, and (*c*) draw a cumulative frequency diagram.

3. (*a*) Reduce Table C to a percentage frequency basis, and (*b*) draw the corresponding histogram. (*c*) Construct a cumulative percentage frequency table, and (*d*) draw the corresponding cumulative diagram.

4. Table D shows the number of lost articles turned in per day at the lost and found bureau of a store. The frequency is the number of days on which the specified number of articles were returned. (*a*) Draw a frequency diagram. (*b*) Construct a cumulative frequency table. (*c*) Draw a cumulative frequency diagram.

5. (*a*) Draw a frequency diagram for the data of Table E. (*b*) Construct a cumulative table, and (*c*) draw a cumulative diagram.

6. Form a frequency table from the data of Table F. First construct a tally sheet making a mark for each time that a temperature of a given number of degrees occurs. From this tally sheet make a frequency table of class interval 1°. Then choose what you consider an appropriate wider class interval and group the frequencies accordingly. Next construct a rectangular frequency diagram. Finally construct a cumulative frequency table and the corresponding cumulative frequency diagram.

TABLE A

GRADES RECEIVED BY A CLASS OF STUDENTS IN AN EXAMINATION

Grade	Frequency
10– 20	1
20– 30	
30– 40	4
40– 50	6
50– 60	7
60– 70	12
70– 80	16
80– 90	10
90–100	4

TABLE B

WEIGHTS OF A GROUP OF MEN

Weight in pounds	Frequency
100–110	2
110–120	3
120–130	11
130–140	34
140–150	84
150–160	65
160–170	48
170–180	33
180–190	20
190–200	11
200–210	4
210–220	3
220–230	1
230–240	1

TABLE C

DISTRIBUTION OF EMPLOYEES IN A CERTAIN INDUSTRY ACCORDING TO ANNUAL EARNINGS

Annual earnings, dollars	Number of employees
0– 200	88
200– 400	236
400– 600	396
600– 800	385
800–1000	412
1000–1200	341
1200–1400	208
1400–1600	113
1600–1800	68
1800–2000	68
2000–2200	33
2200–2400	18
2400–2600	15

TABLE D

LOST ARTICLES RETURNED

Number of articles	Frequency
0	84
1	67
2	37
3	16
4	5
5	1

TABLE E

NUMBER OF CHILDREN BORN PER FAMILY IN 735 FAMILIES

Number of children born per family	Number of families
0	96
1	108
2	154
3	126
4	95
5	62
6	45
7	20
8	11
9	6
10	5
11	5
12	1
13	1