ELEMENTARY STATISTICS
WITH APPLICATIONS IN MEDICINE

CROXTON

Elementary Statistics

with

Applications in Medicine

FREDERICK E. CROXTON, PH. D.
Professor of Statistics, Columbia University

New York *1953*

PRENTICE-HALL, INC.

L. C. Cat. Card No. 52–12158

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

Although this book deals with elementary statistical methods, it is hoped that it will be widely useful to workers in the medical and allied fields. No previous study of statistics is assumed, and only a very modest knowledge of mathematics is required.

A vocabulary of symbols precedes each chapter. Each list of symbols includes all the symbols used in that chapter, not merely those employed in that chapter for the first time.

As in other books on statistics that I have written or on which I have collaborated, practically all numerical illustrations are based on actual, not hypothetical, data. Real illustrative data enhance the usefulness of a book and make it more interesting to the reader.

Thanks are due to many persons for helping in the various phases of the writing of this book. First, it is only fair to state that the able collaboration of Professor Dudley J. Cowden on previous books cannot help but be reflected, to some extent, in the present volume. Alfred J. Kana, Associate in Statistics, William A. Maloy, Instructor in Statistics, and Richard H. Ostheimer, formerly Instructor in Statistics (all of Columbia University), assisted in locating suitable illustrative material. Mr. Kana also read portions of the text and Mr. Maloy prepared some of the charts. Otto Dykstra, Jr. assisted in the preparation of charts and with the computations. Everett P. Messmer did some of the lettering of the charts. For the typing of the manuscript I am particularly indebted to Betty Ruth Austin and also to Anne M. Anderson and Eleanor A. Danker. Finally I am indebted to my wife, Rosetta R. Croxton, who helped make the lists of symbols, assisted with the preparation of the tables, and aided me with the reading of proof.

To all who have allowed me to use data from their articles and from their files, I express my appreciation. Each individual or organization is specifically mentioned at the point where the figures are used.

For permission to reproduce the tables of ordinates and areas of the normal curve, shown in Appendices I and II and taken from H. O. Rugg, *Statistical Methods Applied to Education*, I am grateful

to Houghton Mifflin Company. I am indebted to Professor R. A. Fisher and to Messrs. Oliver and Boyd, Ltd., of Edinburgh for permitting me to reprint in Appendices V, VI, and X portions of Tables III, IV, and V from *Statistical Tables for Biological, Agricultural, and Medical Research.* I am obligated to Professor Egon Pearson for allowing me to use, in Appendices V, VI, X, XI, and XII, tables or parts of tables originally published in Volumes XXII, XXXII, and XXXIII of *Biometrika.*

FREDERICK E. CROXTON

Leonia, New Jersey

CONTENTS

Chapter 1

INTRODUCTION; RATES, RATIOS, AND PERCENTAGES

Introduction

The term *statistics*, as used in this book, refers to the methods which have been developed for working with numerical data. Sometimes, in fact too often, *statistics* is used as a synonym for *data*, but it will not be so employed in this volume.

Nearly everyone involved in any aspect of medicine needs to have some knowledge of statistics. The practitioner and the medical student will better understand the conclusions of articles in the journals if they have some grasp of the statistical methods which are used. The research worker, seeking to present his results in a clear and effective manner, should know what statistical techniques are available to him and should use the appropriate procedure which will enable him to demonstrate the validity (or lack of validity!) of his findings. This is not to argue that everyone practicing medicine or doing medical research should also be a competent statistician. He should, however, be familiar with the rudiments of statistics. Beyond that he may wish to use the services of a professional statistician.

Recently a physician mentioned that he had attended a lecture on a medical topic and that the speaker had stated that his findings had been verified by means of a chi-square test. The doctor had no idea what the chi-square test might be and requested the writer to elucidate. Readers of this book will obtain a basic idea of some of the more important aspects of the chi-square test.

During World War II a pharmaceutical concern asked the writer to serve as statistical consultant in connection with research work under way on an antibiotic. Their approach concluded with the amazing statement: "We have come to the end of our rope, chemically, or we would not be looking for a statistician." The reason for mentioning this occurrence is that the organization should have had

a statistician on its staff, or, alternatively, should have had the services of a consulting statistician from the beginning. The importance of this lack was apparent when it was found that the experiments, seeking to ascertain the effectiveness of the antibiotic, had been so carelessly designed and performed that useful statistical analysis was impossible. The consulting assignment was declined.

In another instance, a group of physicians completed the laboratory portion of an extensive research undertaking involving scores of laboratory animals but, arriving at the point of analyzing their data statistically, found themselves completely helpless. Although none of them had a knowledge of statistics, their basic data were, luckily, sound and it was possible for a statistician to proceed with the statistical tests which were necessary in order to demonstrate that the findings were of value. The results of this study appeared in a leading medical journal.

The pharmaceutical company asked for the help of a statistician too late. Fortunately, statistical help was possible for the group of physicians. Sometimes, statistical analysis is not even attempted, although it is sorely needed. A published article undertook to describe a new device for measuring a certain characteristic of blood. After discussing the apparatus, the author compared the numerical results which he had obtained with those (for the same blood specimens) obtained by two other frequently used methods. That was the end of the article; there was no statistical analysis to demonstrate that his device did indeed give, as alleged, results essentially the same as those given by the customary procedures. A knowledge of the statistical methods described in this volume would have been more than adequate to enable him to analyze his data.

Illustrations of published studies which did not succeed in demonstrating the validity of their findings because of failure to use statistical methods are numerous. An article dealing with the use of antihistamines for treating head colds did not make the statistical tests which were necessary if the efficacy of the drug were to be determined. A discussion of the merits of early postoperative activity did not make the appropriate statistical tests to compare the results of early and late activity and even presented the data in such compressed form that the reader was not in a position to make the tests for himself. A moderately technical house organ of an insurance company compared similarly employed diabetics and

non-diabetics in regard to absenteeism and accidents but did not test the observed differences statistically.

The reader should not get the impression that all statistical work deals with tests of significance (the subject matter of Chapters 9–12). In many other situations simpler procedures are called for. The Veterans' Administration needed to estimate the amount of X-ray film of various sizes required for each hospital. After study of the data, it was found that the different sizes constituted relatively constant proportions at the hospitals, and all that was necessary was to keep a record of the number of patients and make use of the proper ratios for each size of film. Ratios will be discussed in the following section of this chapter. In later chapters illustrative material will involve the use of averages, measures of dispersion, curve fitting, correlation, and other procedures, as well as tests of significance.

Rates, Ratios, and Percentages

Absolute vs. relative comparisons. Before proceeding to discuss rates, ratios, and percentages it may be well to point out the difference between absolute and relative comparisons, since either may be used in statistical work and since rates, ratios, and percentages are all relative figures. Consider two cities: city A, which had a population of 100,000 at a given census and 200,000 at the time the following census was taken, and city B, which had 1,000,000 inhabitants at the first census and 1,300,000 at the second census. If we compare the absolute increases of the two cities, we find that city A showed an increase of 100,000 persons while city B increased 300,000. On a relative basis, city A increased 100 per cent, while city B increased 30 per cent. City B showed the greater *absolute* increase; city A showed the greater *relative* increase. There is nothing contradictory about these two statements. They merely represent different ideas.

Rates vs. ratios. Occasionally a distinction is made between the terms *rate* and *ratio*. A rate is sometimes considered as the amount or quantity of one variable considered in relation to one unit of a different variable. Thus, 30 miles per hour is a *rate* of speed. A ratio may be thought of as the relation existing between two similar variables. For example, if a community had 2,794 births and 1,411 deaths during a year, the birth-death ratio (2,794 ÷ 1,411) would be 1.98. Here the variables were both persons. This dis-

tinction between rate and ratio is not strictly honored in general usage and we shall make no forced attempt to do so in this book.

Ratios and percentages. We shall consider a ratio as a means of expressing the relationship that one magnitude bears to another. Thus, we may say 800 is to 400 as 2 is to 1, or using : to mean "is to" and :: to mean "as," we may write, 800:400::2:1. This is a ratio to the base one. We could also say 800:400::10:5, using a ratio to the base 5; or 800:400::20:10, using a ratio to the base 10; or 800:400::200:100, using a ratio to the base 100. A ratio to the base 100 states the first of the two magnitudes per hundred or *per cent* (per centum) of the second magnitude. Thus, we might say "800 is 200 per cent of 400." Note that a percentage is merely a ratio which has employed 100 as the base.

Although 800 is 200 per cent of 400, it is also correct to say that 800 is 100 per cent *greater than* 400. The reader must always note with care which of these two types of statements is being made.

Computation of ratios and percentages. Ratios involve extremely simple arithmetic, yet many mistakes result from the computation and use of them. The rule for computing a ratio is: *always divide by the base.* For those who, not infrequently, say: "But, I don't know which number is the base," the answer is: *The base is the number to which the other number is being compared.* Suppose that in a given year there were 300 cases of measles in a community and the following year 420 cases, and that the percentage of increase from the first year to the second is desired. The *amount* of increase is 420 − 300 = 120 cases, and 300 is the base. Therefore, the *relative* increase is

$$\frac{120}{300} = 0.40, \text{ or } 40 \text{ per cent.}$$

An alternative method of computation eliminates the initial subtraction to obtain the amount of increase, substituting the subtraction of 100 per cent at the end. Thus, for our figures,

$$\frac{420}{300} = 1.40, \text{ or } 140 \text{ per cent}$$

and the percentage of increase is 140 per cent − 100 per cent = 40 per cent.

When a decrease is involved, either method may still be used. If, in a city, 350 cases of lobar pneumonia in one year are to be com-

pared with 300 cases the following year, the relative change would be

$$\frac{300 - 350}{350} = \frac{-50}{350} = -0.1429, \text{ or } -14.3 \text{ per cent.}$$

Alternatively, we may compute

$$\frac{300}{350} = 0.8571, \text{ or } 85.71 \text{ per cent.}$$

This is the percentage that 300 is of 350, and the per cent of decrease is obtained by subtracting 100 per cent, as before, giving

$$85.71 - 100.00 = -14.3 \text{ per cent.}$$

Note that, while a percentage of increase may be indefinitely large, a percentage of decrease cannot exceed 100 per cent unless negative values are possible. For example, if 400 is increased to 1,600, the increase is 300 per cent; if 400 is increased to 4,000, the increase is 900 per cent. Now, if 400 is decreased to 0, the decrease is 100 per cent; a decrease of 120 per cent would mean that our original figure of 400 had declined to −80. Such a situation is possible when dealing with profit and loss data but not in many other situations. The failure of research workers, or of newsmen, to be fully aware of this has occasionally resulted in the publication of meaningless or misleading statements. In one instance, a new invention for the prevention of frostbite to high-altitude flyers was said to have cut cases of frostbite 1,500 per cent! The figures on which this was based were 60 per 100,000 before the use of the new device, and 4 per 100,000 after its employment. The figure of 1,500 per cent resulted from comparing 60 instead of 56 (the decrease) with the wrong base, 4. The correct percentage of decrease is 93 per cent. In another instance, the use of a new (but unspecified) treatment for malaria was said to have reduced the annual number of cases in a certain location 9,250 per cent from one year to the next. During the first year there had been 555 cases; during the second year, 6 cases. The correct percentage of reduction was 98.9. The incorrect figure was obtained by using the wrong base and relating it to 555 instead of to 549.

Rounding. When we computed the relative decrease involved in the decline of 350 to 300, we obtained 14.29 per cent, which we rounded to 14.3 per cent. The principles involved in rounding, as used in this book, are relatively simple. They are:

1. If the digit to be dropped is more than 5, the preceding digit is increased one. Thus, if we desire to show but one decimal,

16.78 becomes 16.8,
24.36 becomes 24.4, and
6.49 becomes 6.5.

Note that 12.3500000001 would be rounded to 12.4 if we wished to show but one decimal. In this case, we would not be likely to carry the division out as far as indicated, but would merely note that there was a remainder after the 5 in the second decimal place had been obtained.

2. If the digit to be dropped is less than 5, the preceding digit is unchanged. Thus, if we are to show one decimal,

43.61 becomes 43.6,
11.44 becomes 11.4, and
14.30 becomes 14.3.

3. If the digit to be dropped is *exactly* 5, it is customary to use a scheme which will result in half of the preceding digits being raised and half remaining unchanged. One way to accomplish this is to raise whenever the digit preceding the 5 is odd but make no change when the digit preceding the 5 is even.[1] Under this scheme

12.15 becomes 12.2,
7.35 becomes 7.4, and
37.95 becomes 38.0, but
4.85 becomes 4.8, and
22.25 becomes 22.2.

Misplaced decimals. When computing and using percentages, it is important that the decimal points be properly placed. A misplaced decimal point always involves a serious misstatement; the *least* mistake that can occur is for the decimal point to be *one* place out of position. This means that the figure is ten times as large as it should be if the decimal point is one place too far to the right; or one-tenth as large as it should be if the decimal point is one place too far to the left.

[1] It will be noted that this results in all final digits being even, when the dropped digit is exactly 5. There are other possible procedures, but none is thoroughly satisfactory.

An author of an article in a leading magazine[2] was making the point that the tests used by medical-legal experts are valuable but are not reliable if the technician who made the test was careless. In this connection he stated that court testimony by an expert to the effect that a man was drunk because his blood contained 2.5 per cent alcohol should not be considered conclusive. An alert state police trooper pointed out that if a man had 2.5 per cent alcohol in his blood, he not only would be dead but would hardly need to be embalmed.[3] The figure which the author intended to use was undoubtedly 0.25 per cent.

Misplaced decimal points are most likely to occur when very large or very small percentages of change are involved. To give an illustration of only the first of these: a news magazine, commenting on the decrease in the number of Chinese and the increase in the number of Japanese in the United States from 1880 to 1940, stated that the Japanese "increased from 148 to 126,947 or 845 per cent." Actually, the 1940 figure is 857.75 *times* the 1880 figure and the relative increase was 85,675 per cent.

Base should not be small. In general, percentages should not be computed unless the base is 100 or more. Percentages based upon small numbers are apt to be misleading. Consider the case of a college student who had planned to work in a lumber camp but who decided against it when he learned that in the camp "two per cent of the men were married to fifty per cent of the women." Actually, there were fifty men and two women cooks in the camp. One of the men was married to one of the women. The use of percentages with these figures could serve only to supply a joke of dubious merit.

Some medical ratios. Medical data are not infrequently relative figures. Cell volume, for example, is expressed in cubic centimeters per hundred cubic centimeters of blood and is, in fact, a percentage. Data of cell volume appear in various tables in Chapters 3, 4, and 5. In Table 8.1, red blood cell counts are shown. These are stated in terms of millions of cells per cubic millimeter of blood. Since this ratio is referred to rather frequently in Chapter 8, we have employed M^2/mm^3 to mean "millions ($M^2 = 1{,}000^2$) per cubic millimeter."

---

[2] See "How Murderers Beat the Law," by Pete Martin, *The Saturday Evening Post*, December 10, 1949.

[3] *Ibid.*, January 21, 1950, in a letter to the editor from Trooper M. Huffman of the Indiana State Police.

Table 6.1 includes data of intravenous injections of glucose, in terms of grams per kilogram of body weight per hour.

The ratios just mentioned are relatively simple in concept; there are others which are more complex. For example, a hemoglobin coefficient may be computed by stating the grams of hemoglobin which 100 cubic centimeters of blood would contain if the red cell count were 5 M^2/mm^3. This is obtained from the expression

$$\frac{\text{red cell count}}{5\ M^2/mm^3} = \frac{\text{hemoglobin in grams per 100 c.c.}}{\text{hemoglobin coefficient}}$$

Vital rates. Birth rates, death rates, morbidity rates, marriage rates, and divorce rates are occasionally of interest in medicine. To give an extended description of these and of allied figures such as fertility rates, stillbirth ratios, and so forth, is beyond the scope of this book.[4] However, we shall give brief and superficial attention to death rates.

Death rates. The *crude death rate* for an area is obtained by dividing the number of deaths occurring in that area during a year by the population of the area at the middle of the year. The resulting figure is expressed as a ratio per 1,000. As an example, the number of deaths from all causes during 1951 in the five boroughs of New York City was 79,109 and the July 1, 1951 population was estimated to be 7,976,000. The crude death rate for New York City was, therefore,

$$\frac{79{,}109}{7{,}976{,}000}\,1{,}000 = 9.9.$$

When death rates are computed for a period of less than a year, they are usually converted to an annual basis.

The crude death rate is sometimes referred to as a "recorded death rate" or "death rate by place," emphasizing that the figure for number of deaths (1) did not include the deaths of residents of the area who died elsewhere and (2) did include deaths occurring in the area of persons residing elsewhere. To rectify this situation, there may be computed a *resident death rate* which employs deaths

[4] For discussions, see F. E. Linder and R. D. Grove, *Vital Statistics Rates in the United States, 1900–1940,* Federal Security Agency, Public Health Service, National Office of Vital Statistics, 1947, and *Vital Statistics of the United States, 1949, Parts I and II,* Federal Security Agency, Public Health Service, National Office of Vital Statistics, 1951.

according to place of residence of the decedents. The computation of death rates according to place of residence instead of place of occurrence of the deaths does not affect rates for the entire United States and results in but little change in the rates for the individual states. The difference is important, however, for cities. Crude death rates for cities may be quite misleading. In 1935, when crude death rates for 1934 were announced, the borough of Queens, New York, was shown to have a crude death rate of 6.5. This was the lowest rate for any community during 1934 and there was newspaper comment to the effect that Queens was the healthiest place in the United States. Actually, the rate for Queens was low because Queens had proportionally fewer hospitals than the other nearby boroughs. Manhattan, with more hospitals, had a crude death rate of 16.3.

If a death rate, based at first on a preliminary population estimate, is recomputed in relation to a revised population estimate, the new death rate is referred to as a *revised death rate.* A preliminary population estimate is usually a post-censal estimate, as, for example, an estimate for July 1, 1953. A revised population estimate is ordinarily an inter-censal estimate; the July 1, 1953 estimates would become an inter-censal estimate after the 1960 census figures became available.

When death rates are computed for designated portions of the population or for particular causes of death, they are known as *specific death rates.* Thus, we may have "age-specific death rates," "sex-specific death rates," "race-specific death rates," and "cause-specific death rates." Explanations of these and of methods of adjusting death rates to "standard populations" will be found in the references given in footnote 4.