INTRODUCTION TO BIOSTATISTICS

By

HULDAH BANCROFT, Ph.D.

PROFESSOR OF BIOSTATISTICS
TULANE UNIVERSITY SCHOOL OF MEDICINE
NEW ORLEANS

CASSELL AND COMPANY LIMITED
LONDON, TORONTO, MELBOURNE, SYDNEY AND WELLINGTON

INTRODUCTION TO BIOSTATISTICS

For information address Paul B. Hoeber, Inc.,
Medical Book Department of Harper & Brothers,
49 East 33rd Street, New York 16, N. Y.

FIRST EDITION

Bound in Great Britain by Webb, Son and Company, Ltd.

This book is dedicated

to the students

who through the years

have made the teaching of

Biostatistics

a pleasure

PREFACE

THE present textbook represents the third revision of a series of mimeographed notes prepared for use in the teaching of biostatistics to sophomore medical students at Tulane University. The apparent usefulness to and acceptance by the students here during the past nine years has led the author to agreee to its publication. As originally written it has been used in a course covering approximately forty-eight hours, of which fifteen are spent in lectures, the balance in supervised laboratory work.

The text deals mainly with statistical methods appropriate for large samples. Frequency distributions, tabulation, graphing, use of centering constants and measures of variation, and other descriptive methods are treated in the introductory chapters. Further topics include the binomial, χ^2, and the use of the normal distribution in large sample approximate tests of significance for differences in sample means and proportions. Additional chapters deal with the *t* tests, vital statistics including the use of the modified life table techniques in follow-up studies, and the final chapter presents a brief discussion of quantal bioassay and the Reed-Muench Method.

Because the book was written for medical students and practicing physicians, who frequently have little training in mathematics, the presentation is in simple, nontechnical terms. No knowledge of mathematics beyond elementary high school algebra is required for understanding. Since the book will be used primarily by the medical profession, illustrations in the text and examples at the end of each chapter have been taken almost entirely from clinical medicine.

The author wishes to acknowledge her gratitude to Dr. John Fertig, Dr. Alan Treloar, and Dr. R. A. Fisher and their publishers for permission to print the tables dealing with size of sample, the χ^2 and t distribution respectively.

Dr. Lila Elveback has given many helpful suggestions in the preparation of the manuscript. To Miss Ethel Eaton credit should be given for her painstaking preparation of the illustrations as well as for many of the computations. To Mary Grace Kelleher, Patricia Spaid, and Arva Boesch thanks are due for the careful typing of the manuscript.

HULDAH BANCROFT

New Orleans

INTRODUCTION TO BIOSTATISTICS

CONTENTS

1

INTRODUCTION

A GROWING emphasis on the role of quantitative methods in the practice of medicine makes it imperative that the medical student as well as the practicing physician have some knowledge of statistics. The medical student while in school is taught the best method of diagnosis and therapy. After graduation he must of necessity depend on papers presented at medical meetings and in current journals to learn new methods of therapy as well as improvements in diagnosis and surgical technique. He, therefore, must be able to evaluate for himself the results of other workers. He must decide when a new technique or method shall supplement an older one. He must be able to answer the question of the mother as to preventive measures against polio with as much surety as he now advises her regarding vaccination against smallpox. He should be able to give the family intelligent assurance of the prognosis for a given patient. Such prognosis may depend on his ability adequately to appraise laboratory findings as well as on his knowledge of the relation of age, sex, and other conditions of the patient to a particular disease. New knowledge regarding facts such as these will come to the physician through research work done by himself or by others. He must, therefore, be able to select from masses of information that which is of high caliber and which will pass rigid scientific tests. He must develop a healthy skepticism toward everything he reads.

Just how can a knowledge of statistical techniques assist him in this problem? First and foremost, he must recognize that individuals vary, not only from each other, but that an individual varies from day to day, or from hour to hour. A certain amount of variation is normal, but the question facing the physician is to determine when a specific variation becomes pathologic. To determine this, the student must learn how variation in normal individuals is measured and what is the range of normal variation. He must learn that there is some error present in every measurement or count made. It is highly unlikely that two successive red blood cell counts made on the same specimen of blood will be identical. When therefore does a difference become greater than the error of measurement? For example, a patient has a red blood cell count of 4.3 million cells today. The laboratory tomorrow reports 4.2 million cells. Does this indicate that the count is decreasing? Or a patient is admitted with a white blood cell count of 6,000. Two hours later the laboratory reports a count of 6,200 on the same patient. Does this indicate a real rise in white count? In other words, can these differences be explained by the inaccuracy inherent in the method of counting blood cells? To treat his patient with the utmost skill he must know the answers to questions such as these. For every measurement or determination provided by the laboratory, the physician should know the variation that is a part of the method itself, in order to know when a given variation represents a real change in a patient.

Whenever new methods of diagnosis or therapy are introduced, the question to be answered is whether the specific new method under question is superior to the old method. Critical evaluation of the experimental study must be made. Questions that must be answered are:

1. Were controls used in measurement of results? If so, were they well selected? Were the experimental group and the control group as near alike as possible as regards all known factors that would affect the outcome? In other words, was the factor in question the only important one differentiating the groups?

2. Was the difference in the results obtained greater than could be attributed to chance or normal variation?

Only when these questions have been answered can conclusions regarding a new method be drawn. It must be recognized that there are no statistical techniques available that will give absolute proof that one technique is in all respects better than another. They will, however, give the probability of a given difference occurring if there is no real difference in technique so that the reader may conclude whether in a particular instance the difference is significant.

That there is ample need of a knowledge of some of these concepts is shown by Dr. O. B. Ross, Jr., in a report published in the *Journal of the American Medical Association, 145,* 1951, p. 72. Dr. Ross analyzed the first 100 articles dealing with therapy published betweeen January 1 and June 1, 1950, in the following five leading medical journals: *The Journal of the American Medical Association, Annals of Internal Medicine, The American Journal of the Medical Sciences, Archives of Neurology and Psychiatry,* and the *American Journal of Medicine.* Of these articles, and remember they are typical of articles from which conclusions must be drawn, 45 per cent were based on wholly uncontrolled studies. In 18 per cent the controls were poorly chosen and so inadequate; in 10 per cent because of the nature of the problem no controls could be used. In only 27 per cent was the study well controlled. No doubt these findings could be duplicated in other journals. This should emphasize the need for critical reading of the literature.

To aid the student in developing such an attitude certain basic statistical concepts are necessary. Some of the most important of these are presented in the following chapters. The student must recognize that statistics is a combination of logic and mathematics. An attempt is made to present concepts with a minimum of technical terminology and mathematical proof. The science of statistics cannot be mastered in this elementary presentation. Rather it is hoped that the student will become conscious of the way in which biometric techniques may be of value in the analysis and presentation of quantitative data. It is believed that this study will enable the student to read current medical literature more critically while in medical school as well as later in private practice. For those students who may intend to carry on research, it should help them to recognize the type of data that need statistical analysis and seek the aid of someone with more advanced training in the field of statistics.

2

CLASSIFICATION AND TABULATION

IT HAS been said that all knowledge arises through some process of observation. This observation may be as simple as noting whether an event occurs or does not occur, such as the fact that an individual wears glasses or does not wear glasses. On the other hand, the observation may involve a more complicated procedure, such as the measurement of the amount of hemoglobin in 100 cc of blood. Such observations, taken collectively, are called *data*. The first type of data is usually referred to as enumeration or counting data; the second type is known as measurement data.

Regardless of the type of information collected, it must be recorded in some manner. This may be done by use of elaborately planned printed individual record sheets, by use of a notebook with permanent or loose-leaf pages on which the data are recorded in parallel columns, or in many instances it may be recorded directly on a *card* record form. If the data are recorded on sheets or in books, they must be transferred to some type of a card for analysis. Therefore, if at all possible, the card system should be used.

For a simple study in which only a few items are to be recorded, often a blank 3- × 5-inch card will be found to be very satisfactory. For example, a young resident wishes to study the relation of allergy to bed-wetting in children. Items he wishes to consider are sex, race, and present age of child, age at which bed-wetting first appeared, age at which it ceased, whether the child has had any form of allergy, and if so what was the allergic diagnosis. The following master card would easily serve for recording these data.

Presence of allergy	Race	Present age
Allergic diagnosis	Sex	Age first wet bed
Personal identification		Age last wet bed

If for the first child studied no allergy was present, the word *No* would be written in the position indicated by *Presence of allergy;* race would be recorded in the position marked *Race* as *W* for white, *B* for Negro, *O* for other; present age would be recorded as 6 if six years at last birthday; similarly other positions would be filled by appropriate letters or numbers. In the lower left-hand corner, some form of identification should be given. This might be name, hospital record number, or clinic number. This is fundamental to good records since checking of data must be done at times.

This type of card can be used for any number of items from one or two up to eighteen, nine being recorded on the face of the card and nine on the back since there are nine positions which can easily be identified. If both front and back of the card are to be used, one corner must be cut so as easily to recognize front and back. By using a larger card, as many as sixteen positions on the card can be clearly marked off. Larger cards, however, are more difficult to handle when the final tabulations are made.

When more elaborate studies are to be undertaken, printed forms, either cards or sheets, must be used. These will be designed for the particular problem at hand. Care should be taken to see that adequate space is allowed for recording the specific information necessary since crowding of information into too small a space often leads to inaccuracy.

After the individual records are made out, some process of classification must take place before analysis of the data can be made. This process involves grouping the observations into mutually exclusive categories according to certain definite characteristics or attributes. These may be either qualitative or quantitative in nature. Qualitative attributes such as sex, race, or presence of a particular disease entity are easy to classify. Only a definite number of possibilities exist: the person is male or female, is white or nonwhite, has measles or does not have measles. When the selection of the category into which a record falls involves personal judgment, such as severity of disease, or skin texture, or color of eyes, the problem is not as simple. Even though definite criteria are set up, personal bias will enter into the classification and comparisons of classifications made by different persons must be made with caution.

Quantitative data may be of two types, *discrete* or *continuous*. Discrete measurements are those in which only whole units can be counted or measured. For example, in counting the white blood cells in a given field one might count sixty lymphocytes. The count must be a whole number of lymphocytes since fractional lymphocytes do not exist; if only a part shows in a field it is not included. Contrast this with the measurement of the hemoglobin in the blood. Here the measurement may be in whole grams or in any fractional part of a gram, depending on the fineness of the measurement used. Such data are continuous in character since an individual could have 12.1 gm., 12.15 gm., 12.155 gm., or any such fractional part of a gram.

In qualitative and in discrete measurement data the selection of the groups into which the data are to be placed is relatively simple since only a limited number of possible groups exist. The groups are sharply delimited and there can be no question as to the group to which a single observation belongs. A person is white or nonwhite; male or female; or has been admitted to a hospital one, two, or some other definite number of times.

With quantitative data in which the measurements are made on a continuous scale the question of how measurements are to be grouped becomes more of a problem. It must be recognized that in any type of grouping certain detail is lost. To classify in one group babies whose birth weight is from 1,000 to 2,000 gm. does not tell us how many weigh 1,000, 1,100, 1,200, etc., gm. In general, not too much detail will be lost if 8 to 15 groups are used. The limits of these groups—or class intervals as they are called—should be stated clearly and in such a way that there can be no doubt as to the interval to which an individual observation belongs. Suppose a series of determinations of the amount of nitrogen in a 24-hour specimen of urine has been made for a group of individuals. Measurements have been recorded to hundredths of a gram and the class interval has been selected as 1 gm. Correct statement of the interval limits would be 12-12.99 gm., 13-13.99 gm., not simply 12-13 gm., 13-14 gm., etc. With this latter method of stating class intervals, the individual with exactly 13 grams might be placed in the first group by one worker and in the second group by another worker.

In most instances it is best to select class intervals of equal size. However, in problems dealing with morbidity and mortality unequal class intervals are often better since disease affects different ages differently. To classify a group of patients with measles by ten-year age groups would be meaningless since practically all cases would fall in either the 0-9- or the 10-19-year group. Or to classify a group of patients with cancer by ten-year intervals would be equally as poor, since relatively few would fall in the early decades. The particular intervals used will depend on the distribution of the data. When dealing with morbidity and mortality *from all causes* the following age groups are commonly used:

Under 1 year	infant
1-4 years	preschool
5-14 "	school
15-24 "	adolescent
25-44 "	young adult
45-64 "	middle age
65+ "	old age

After the class intervals have been determined or the particular method of classification selected, the next problem that arises is that of actual tabulation of the data.

If the record form used is a complicated printed sheet, or if observations have been recorded in notebook form, there are two methods of tabulation available. One is that known as the tally method. A tabular arrangement is set up as in the accompanying diagram. Each observation is indicated by means of a tally mark in the appropriate space. The fifth mark in each row is crossed over the first four so as to make counting of tallies easy. For this reason it is often called the cross-five method. It is very simple when only one attribute is being classified, but checking for accuracy of classification can be done only by repeating the entire work independently.

Age in years	Tally mark	Frequency in group
5	卌 ///	8
6	卌 卌 //	12
7	卌 卌 卌 /	16
8	卌 卌 卌 //	17
9	卌 卌 //	12
10	卌	5
Total		70

When several attributes are being classified as in the second arrangement, there are many chances for error.

Age in years	White		Nonwhite	
	Male	Female	Male	Female
5				
6				
7				

The second method is to transfer the basic data from the record to some type of a card by means of a code. These cards can then be sorted mechanically. Two of the commonly used systems are the mechanical punch cards of the International Business Machine Corporation and of the Remington-Rand Corporation. A marginal punch card is also available, made by the McBee Company. The latter card has the advantage that the card itself may serve as a record, with the data coded on the four edges of the card by use of holes that may or may not be punched out. This card can be sorted fairly easily by hand.

Records may also be transfered to an unprinted 3- × 5-inch or somewhat larger card. This is often a timesaving device even if the original data have been assembled in notebooks or on sheets of paper, since sorting is much simpler if data have been collected directly on a card such as the one illustrated or have been transferred to such cards; then the cards may be divided according to a single attribute by placing them in piles according to the particular code for that attribute. For example, in sorting the cards for race all marked *W* would go in one pile, those marked *B* in a second pile, and those marked *O* in a third pile. It now is very easy to count the number of cards in each pile and record in the tabulation form. If one wishes to subsort the racial cards by sex, all that is necessary is to take the *W* pile and break it down into *M* and *F* cards. The sum of the counts for *M* and *F* should equal the number of *W* cards. Thus, there is an immediate checking for accuracy of sorting. As many subsorts as are desired can be made.

The questions of how the data are to be collected and how they are to be arranged for purposes of tabulation are important. In any problem certain questions must be answered. Among these are how many records are involved, how many items or variables are to be studied, what facilities are available for aid in tabulation, and how much help can be obtained. If many records are involved and a large number of variables are to be studied, some form of mechanical tabulation is necessary. If only a few variables are to be studied, even though a fairly large number of records are involved, some form of simple card with hand sorting will suffice. With only a small number of records regardless of the number of variables to be studied, no elaborate mechanical equipment is necessary.