MODERN ELEMENTARY STATISTICS

JOHN E. FREUND

MODERN ELEMENTARY STATISTICS

BY JOHN E. FREUND

Associate Professor of Mathematics, Alfred University

1952

NEW YORK • PRENTICE-HALL, INC.

Copyright, 1952
by
PRENTICE-HALL, INC.
70 Fifth Avenue
New York

L. C. Cat. Card No.: 52–11149

PREFACE

This book has been written for a general introductory course in statistics. A preliminary draft in mimeographed form was used for several years in a course designed to acquaint beginning students in the social, as well as the natural, sciences with the fundamentals of modern statistical methods.

The order and the emphasis of the material covered follows the modern trend in the teaching of statistics—to include *informally* topics that in the past have often been taught only on an advanced level. Although a large part of this book deals with the concepts and problems of inductive statistics, the standard techniques of descriptive statistics are amply covered in Part I and in Chapters 18 and 19.

Mathematical proofs and derivations have been keyed to the lowest level at which modern statistics can effectively be taught. Since the mathematical training assumed of the reader is a knowledge of arithmetic and, perhaps, some high school or college algebra, many theorems, for example, those relating to sampling distributions, are stated in the text without proof.

To acquaint the reader as early as possible with the idea of a theoretical or expected distribution, Chapter 3 contains a more or less intuitive introduction to the binomial and normal distributions. Problems relating to these distributions are actually not taken up until these distributions are discussed later in more detail in Chapters 6 and 7. This early introduction of theoretical distributions has made it desirable to treat them as percentage rather than as probability distributions, because the concepts of probability and chance are not discussed in Part I of this book.

The exercises which are given at the end of the various sections and chapters were distributed as impartially as possible among the various natural and social sciences. The reader should, therefore, not find it difficult to locate a sufficient number of examples and problems pertaining to his particular field of interest. To facilitate the reader's task of distinguishing between those formulas which are used for practical applications and those which are given primarily as definitions or as part of derivations, the first are marked with a large asterisk.

The author would like to express his appreciation to his many colleagues and students whose helpful suggestions, criticisms, and comments contributed considerably to the writing of this book. The author is greatly indebted to Professor M. Bernstein for his careful reading of the manuscript and proofs, to the editorial staff of Prentice-Hall, Inc. for courteous cooperation in the production of the book, and above all to his wife for her continuous assistance and encouragement.

Finally, the author expresses his appreciation to Professor R. A. Fisher and Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint Tables III and IV from their book *Statistical Methods for Research Workers;* to Professor E. S. Pearson and the Biometrika Office for permission to reproduce Tables III and VI; to Dr. C. E. Eisenhart for permission to use the material which is given in Tables Va and Vb; to Henry Holt and Company, Inc. for permission to reprint Table I; and to The Macmillan Company for permission to reprint Table VIII.

JOHN E. FREUND

CONTENTS

PART ONE: DESCRIPTIVE STATISTICS

PART TWO: PROBABILITY AND SAMPLING

PART THREE: PREDICTION AND CORRELATION

PART FOUR: SPECIAL TOPICS

PART ONE:

DESCRIPTIVE STATISTICS

1. INTRODUCTION AND MATHEMATICAL PRINCIPLES

1.1 Introduction

The beginning student in the natural and social sciences, for whom this book is written, often faces the study of statistics with mixed emotions. He knows, or at least he is told, that on the one hand he cannot proceed to advanced studies in his chosen field without a thorough understanding of statistical methods, while he remembers distinctly, on the other hand, the difficulties which he may have encountered in his previous contacts with mathematics.

There can be no doubt that it is practically impossible to understand the meaning and implications of the work which is being done in the social sciences, and for that matter also in the natural sciences, without having at least a speaking acquaintance with statistical theory. The reason for this is that scientists obtain knowledge from experimentation, measurements, and observations. Consequently, they must know how to squeeze usable information from their data, or at least know how to present their data in a form that lends itself to further study and displays the most important features of their results.

Numerous books have been written on business statistics, statistics for psychologists, educational statistics, and on statistical methods in sociology. It is true, of course, that these diversified fields of scientific inquiry demand somewhat different and specialized techniques in particular problems; yet the fundamental principles which underlie all the various methods are identical regardless of the field of application. This will become evident to the reader once he accepts the idea that *statistical methods in general are nothing but a refinement of everyday thinking.*

The approach which we shall follow in this elementary study of the subject of statistics is keynoted by the above statement, because it is our goal to acquaint the beginning student in the natural and social sciences with the ideas and with the concepts which are fundamental to the understanding of statistical methods. This in turn—at least this is our hope—will give him a sounder understanding of scientific principles and will enable him to see the scope and limitations of empirical knowledge.

As we have said before, the study of statistics may be directed towards its application in a particular field of scientific inquiry. Furthermore, it may also be presented in varying degrees of mathematical refinement and in almost any balance between theory and application. Since it is, in our opinion, much more important to understand the *meaning and implications* of a few basic concepts than to be able to recite verbatim a large assortment of impressive sounding formulas, we shall have to sacrifice some of the mathematical detail which is customarily covered in introductory texts in statistics. This is unfortunate in some respects, but rather than miss seeing the forest because of the trees, we shall compromise by chopping down a few of the trees, i.e., eliminate some of the less essential detail. With the use of this approach we hope to avoid the dangerous effect which often results from the indiscriminate application of so-called standard statistical procedures without a thorough understanding of the fundamental logical principles which are involved.

1.2 Descriptive and Inductive Statistics

Everything which deals even remotely with the collection, analysis, interpretation, and presentation of numerical data may be classified as belonging to the domain of statistics. The task of computing the batting average of a ballplayer is delegated to a team statistician; births, marriages, and deaths are recorded as vital statistics; and one of the most advanced branches of atomic physics goes by the name of quantum statistics.

The word *statistics* itself can be used in a variety of interpretations. We may use it, for example, in the plural to denote simply a collection of numerical data. Such statistics can be found in the *World Almanac,* the U.S. census reports, in the records of county clerks, and wherever numerical data are collected and reported. The second meaning which we give to the word, also in the plural, is that of the totality of methods which are employed in the collection, study, and analysis of numerical data. In this sense, statistics is a branch of applied mathematics, and it is this field of mathematics which is the subject matter of this book. In order to complete this linguistic study of the word *statistics,* we also mention that the word *statistic* in the singular is used to denote a particular measure or formula like an average, an index number, or a coefficient of correlation.

In order to clarify the basic differences between the various types of statistical problems which we shall discuss, let us first differentiate between the ideas of *descriptive* and *inductive* statistical methods. Although the term *descriptive statistics* is often limited to denote merely the tabular or graphical presentation of numerical data, we shall use it in a much wider sense. By descriptive statistical methods we shall understand any treatment of numerical data which does *not* involve inductive generalizations. In contrast, we shall speak of inductive statistics the very moment that we make generalizations, predictions, or estimations.

This distinction can easily be explained by means of the following simple illustration. Let us suppose, for example, that students Smith and Jones take three tests each. Smith received the grades of 63, 41, and 55, while Jones's scores were 60, 58, and 59, respectively. From this information we can easily find that Smith had an "average" of 53, whereas Jones averaged 59. The word "average" was advisedly given in quotes because, as we shall see later, its meaning is not without ambiguity. The computation of an average is an elementary statistical technique which accomplishes the task of putting a large volume of numerical data into a (for certain purposes) more *usable* form.

So far we have merely computed two numbers, the two averages, which are in a sense *descriptive* of the numerical information which we were given at the start. Consequently this part of our work belongs to what we have defined as descriptive statistics. Now let us suppose that on the basis of the given information we are asked to decide which is the better student, Smith or Jones. The very moment that we generalize in this fashion and make a statement of the type that Jones is the better student, we are employing inductive methods, since we are saying *more* than we were given in the original data. We now find ourselves in the domain of inductive statistics. This distinction is not difficult to see; as long as we restricted ourselves to the computation of the two averages, we added nothing to the given information, but merely rearranged it in a different and possibly more convenient form. It does not follow by any means that Jones is necessarily the better student. Smith might have had an off day, he might have been ill when he took the test, and it could even have happened that Jones was just plain lucky in finding the correct solutions to the problems. Therefore if we claim that the given evidence implies that Jones is the better student, we are making a generalization which may or may not be correct, and we are consequently taking a gamble. Careful evaluation and analysis of the chances which must be taken whenever we make such inductive generalizations are the main task of inductive statistics.

The terms *descriptive* and *inductive statistics* apply, therefore, to the kind of problem which we want to solve, rather than to the particular formula or statistic which we may choose to employ. We can compute, for instance, an average solely for the purpose of description. On the other hand, we can compute it in order to make generalizations and predictions. This is true for practically all statistical measures which will be discussed in later chapters.

Since the primary objective of science is to discover predictable generalizations concerning observable phenomena, we shall emphasize in this text the types of problems which belong to what we have called inductive statistics. The problem of arriving at general hypotheses, "good" estimates, and "fair" predictions requires the type of careful statistical thinking, which is the cornerstone of science. It is, therefore, our goal to acquaint the reader with the type of thinking which is essential for an understanding of the meaning of scientific statements in particular, and scientific theories in general.

Let us now acknowledge the fact that a limited amount of mathematics is indeed necessary as a prerequisite for any course in statistics on the college level. A thorough study of the theoretical principles which underlie statistics would require a knowledge of mathematical subjects which are commonly described as graduate courses even for a student of mathematics. Since this book is written for students who are not only undergraduates, but also usually nonmathematicians, our aims and therefore also our requirements are considerably more modest.

Actually the mathematical background which is necessary for this study of elementary applied statistics is amply covered in college algebra or any equivalent freshman course in mathematics. Besides having a reasonable skill in the elementary arithmetic of addition, subtraction, multiplication, and division, the reader should be familiar with the most common types of problems which are studied in college algebra, such as solving simple equations, substitutions, and use of the functional notation, logarithms, and tables.

1.3 The Use of Summations

Since many of the formulas which we shall develop in the following chapters must be applicable to different sets of numerical data, we shall have to represent the scores, measurements, or observations with which we shall deal by means of some general symbols such as x, y, and z. Unless we introduce a slight modification, however, this symbolism will lead us to immediate difficulties because, if we were to represent, for example, the age of every inhabitant of New York City by means of a different letter, we would easily use up the English, Greek, Russian, and Hebrew alphabets, without accommodating even as much as 1/500 of 1 per cent of the population of New York City. This makes it clear why we shall have to adjust our symbolism to fit the treatment of mass data, following the customary practice of using subscripts. The three scores which Jones received in our previous illustration can now be written as x_1, x_2, and x_3, respectively. If we want to discuss any one of these scores in general, we shall call it x_i, where i is, so to speak, a variable subscript, which can in this example take on the values 1, 2, and 3. Instead of the letter i we could, of course, have arbitrarily used any other letter such as k, l, m, . . . , and instead of x we could also have used any arbitrary letter or symbol. Therefore x_{31} and x_{73} might stand for the intelligence scores of the thirty-first and seventy-third individuals, while in a different problem x_{12}, y_{12}, and z_{12} might represent the age, height, and weight of the twelfth guinea pig. It is customary to use different letters for different kinds of measurements and different subscripts if we are speaking about different individuals (items).

In order to simplify the many formulas which will involve great quantities of numerical data, we shall now introduce the symbol Σ (capital Greek sigma), which is what might reasonably be called a *mathematical shorthand notation.* We have by definition

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4 + \cdots + x_n \qquad (1.3.1)$$

which is read as: "the summation of x_i, i going from 1 to n." It means that we take the sum of the x's which have the subscripts 1, 2, . . . , n. Similarly we might have, for example

$$\sum_{i=1}^{4} y_i^2 = y_1^2 + y_2^2 + y_3^2 + y_4^2$$

or
$$\sum_{i=1}^{5} x_i f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4 + x_5 f_5$$

Since the summation sign will appear in many formulas, it will prove to be very helpful to study some of the fundamental theorems concerning summations. These theorems, a few of which will be given below, are not difficult to understand and prove.

THEOREM A: *The summation of the sum (or difference) of two or more variables or terms is equal to the sum (or difference) of their separate summations.*

Symbolically we can write this theorem in the case of three variables as

$$\sum_{i=1}^{n} (x_i + y_i + z_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} z_i \qquad (1.3.2)$$

If we had wanted to use minus instead of plus signs, we could have done so on both sides of the equation. The proof of this theorem, which consists of showing that the two sides of the equation are identical if they are written in full without the use of summation signs, will be left as an exercise to the reader.

THEOREM B: *The summation of a constant times a variable is equal to the constant times the summation of the variable.*

Symbolically this can be written as

$$\sum_{i=1}^{n} kx_i = k \sum_{i=1}^{n} x_i \qquad (1.3.3)$$

This theorem follows immediately from the following considerations:

$$\begin{aligned}\sum_{i=1}^{n} kx_i &= kx_1 + kx_2 + kx_3 + \ . \ . \ . + kx_n \\ &= k(x_1 + x_2 + x_3 + x_4 + \ . \ . \ . + x_n) \\ &= k \sum_{i=1}^{n} x_i\end{aligned}$$

THEOREM C: *The summation of a constant k, from 1 to n, equals the product of k and n.*

This means that

$$\sum_{i=1}^{n} k = kn \qquad (1.3.4)$$