Introduction to Statistics

Robert A. Hultquist

INTERNATIONAL SERIES IN DECISION PROCESSES

Introduction to Statistics

Robert A. Hultquist

Pennsylvania State University

HOLT, RINEHART AND WINSTON, INC.

New York · Chicago · San Francisco · Atlanta · Dallas
Montreal · Toronto · London · Sydney

Library of Congress Catalog Card Number: 69-17654
AMS 1968 Subject Classifications 6201, 6260, 6255

SBN: 03-075590-5

Printed in the United States of America

1 2 3 4 5 6 7 8 9

Introduction to Statistics

INTERNATIONAL SERIES IN DECISION PROCESSES

INGRAM OLKIN, Consulting Editor

A Basic Course in Statistics, 2d ed., T. R. Anderson and M. Zelditch, Jr.
Introduction to Statistics, R. A. Hultquist
Applied Probability, W. A. Thompson, Jr.
Elementary Statistical Methods, 3d ed., H. M. Walker and J. Lev
Reliability Handbook, B. A. Koslov and I. A. Ushakov (edited by J. T. Rosenblatt and L. H. Koopmans)
Fundamental Research Statistics for the Behavioral Sciences, J. T. Roscoe

FORTHCOMING TITLES

Introductory Probability, C. Derman, L. Gleser, and I. Olkin
Probability Theory, Y. S. Chow and H. Teicher
Statistics for Business Administration, W. L. Hays and R. L. Winkler
Statistical Inference, 2d ed., H. M. Walker and J. Lev
Statistics for Psychologists, 2d ed., W. L. Hays
Decision Theory for Business, D. Feldman and E. Seiden
Analysis and Design of Experiments, M. Zelen
Time Series Analysis, D. Brillinger
Statistics Handbook, C. Derman, L. Gleser, G. H. Golub, G. J. Lieberman, I. Olkin, A. Madansky, and M. Sobel

TO MY WIFE, LYNNE
AND MY MOTHER, ALICE

When I applied my heart to know wisdom,
and to see the business that is done upon the earth
then I beheld all the work of God,
that man cannot find out the work that is done
under the sun: because however much a man labor
to seek it out, yet he shall not find it;
yea moreover, though a wise man think to know it,
yet shall he not be able to find it.

ECCLESIASTES 8: 16, 17

PREFACE

This text is the outgrowth of a careful editing of lecture notes, laboratory exercises and quiz problems given, during the past ten years, to students in first courses in statistical methods. It is a text which perhaps reflects more of what faculty members, outside of mathematical science, feel is important than what I personally like to teach. It is a text for students with motivation to perform experiments and to analyze the results, but I have found that it is also a text for beginning students of mathematical statistics. A professional statistician without a good knowledge of the concepts presented in this text would be considered by many of his colleagues to have a serious gap in his training.

Great effort has been put forth to make the text one unified story. The restriction to consider only sampling from normally distributed populations was made in order not to interrupt the plot. This restriction helped make it possible to present many concepts in a relatively small volume. I believe that in a first course it is better to examine many concepts with reference to one parent population than to examine a few concepts with respect to many parent populations. It is hoped that through thoroughly understanding the role of the normal density the student will, by transfer of knowledge, be in a position to efficiently learn statistical procedures for situations associated with other density functions.

The text is devised so that classical high school algebra is the only mathematical prerequisite. Early in the text the reader is introduced to summation notation, and this notation is used frequently throughout the remainder of the text. The topics are those which can be easily handled without matrix notation although the use of matrix notation would in many instances make the presentation more elegant. Although the topics are for the most part those taught in most beginning methods courses, the emphasis given to many of the topics is different from that given to these topics in most courses. For example, the formal presentation of probability is confined to a few pages.

The text starts with a discussion of some very basic concepts which lead into an introduction of inference. The story of statistical inference proceeds rather rapidly until it reaches analysis of variance and regression, whereupon topics are more carefully surveyed. This approach is preferred by most experimental and consulting statisticians. After the basic concepts are presented, the text deals with point estimation methods, interval estimation methods, and tests of statistical hypotheses in association with many different experimental situations. Some emphasis is placed on choosing a model to adequately represent the data, and in this regard the text employs a rather different illustration technique. For several models, observations are created from assumed parametric values and then the created observations are analyzed. Finally, the assumed values of the parameters are compared with the estimate obtained from created data.

The text can be used in a course with or without a weekly laboratory session. Numerous problems are placed in each chapter, and some of the problems are designated "Laboratory Type Problems." Most of the problems are an integral part of the text, and it is strongly recommended that all or nearly all of them be assigned as homework or worked in class. Since expected mean squares play a very important role in inference theory, numerous problems associated with determining expected mean squares have been included in the exercise sections. Students generally find these problems exceedingly difficult; hence, it is advised that some of these problems be worked by the instructor. If most of the problems are worked, the text contains material for a two-term course.

The text is not specifically designed for any discipline-oriented class. It contains illustrative examples and exercises from a variety of application areas. Students usually appreciate a statistical method to a greater extent when they see the power associated with wide applicability of a procedure. Few if any of the illustrations are so technically discipline-oriented that a student outside that discipline gets lost in the jargon of the illustration.

I would like to thank each senior member of the Statistical Laboratory at Oklahoma State University for educating me in experimental design and instilling in me a desire to help experimenters. Thanks go to the faculty and staff of Pennsylvania State University and to the chairman of that department for providing an academic climate in which I had the time and desire to prepare the manuscript. I appreciate the efforts of about a dozen individuals at Pennsylvania State University and Holt, Rinehart and Winston, Inc. who typed and in other ways worked on the manuscript. Thanks go to my immediate family and to my in-laws, who encouraged me to write this book. Finally, I acknowledge with deep appreciation my parents, who among many things made my professional education possible.

Boalsburg, Pennsylvania
March 1969

Robert A. Hultquist

CONTENTS

1

BASIC CONCEPTS

1.1 INTRODUCTION

Statistics is a science that concerns itself with experimentation and the collection, description, and analysis of data. The *statistical layout* displayed in Table 1.1 contains data typical of that considered by many

Table 1.1 Strength Test Scores of Sophomore Boys by High School and Class

High School	1	2	3	4	5
Scores from class (1, *j*)	70.3	78.9	52.3	69.8	80.5
	80.9	83.1	59.8	58.5	88.3
	59.7	85.2	61.3	75.7	81.2
	63.8	64.7	70.6	59.6	79.5
	58.5		71.7	68.7	74.4
	75.3		75.4	81.3	
Average for class (1, *j*)	68.1	78.0	65.2	68.9	80.8
Scores from class (2, *j*)	60.5	73.8	72.3	59.8	*No second class in this school*
	61.3	78.5	77.6	65.4	
	62.8	83.7	78.9	83.1	
	69.7	84.5	69.0	57.6	
	58.3	78.9	83.2	55.2	
	65.5	77.5	69.0	48.8	
	72.8	79.6		57.3	
Average for class (2, *j*)	65.2	79.5	75.0	64.5	

statisticians. The test scores recorded in Table 1.1 correspond to strength tests given to boys in nine different sophomore physical education classes. The boys and the scores that they made are examples of what more generally are called *experimental units* and *observations*, respectively. Five different schools were represented with two classes from each of four schools and different teachers for each class. The teacher (class) designation is as follows: (1, 3) indicates the first teacher from the third high school; (2, 4) denotes the second teacher from the fourth high school; and, in general, (i, j) denotes the ith teacher from the jth high school.

With reference to the data, we might ask: Do the scores provide evidence that one school is better than the rest with respect to its physical education program? Are there significant differences in teachers? Do some classes have significantly more variability than other classes? What is a typical score for the class with teacher (i, j)? There are of course many other questions that might be asked. Some of the questions are even more basic, for they deal with concepts at the very heart of experimentation. Are there enough boys to give good answers to questions such as those above? Is there a better layout for displaying the results of the experiment? What type of new experiment should be run if more information is desired relative to the teachers, the high schools, and the physical fitness of the students? This book presents methods that attempt to answer questions such as those posed here. It presents statistical methods for describing and analyzing results of experiments, augmented by discussions of why a technique is a good one and why it is preferred over another technique in a given situation.

Statistical methods are tools for examining data. The statistician dealing with business trends, the statistician keeping the baseball average up to date, the statistician conducting public opinion polls, these and many others, all possess a common characteristic in that they deal with data. Data handlers do not, of course, use the same tools all of the time, but statistical methods preferred by different types of statisticians do overlap, and it is the overlapping concepts that are stressed in this treatment of statistics. With these comments in mind, we shall consider some of the basic concepts underlying the study of data.

1.2 TYPICAL VALUES

Data can be and are presented in several ways. We consider here a numerical description of the outcome of an experiment. The word "experiment" is used to include the case where we merely observe nature. Let the numbers (observations), say n of them, that come from an experiment be denoted by indexed letter symbols such as $X_1, X_2, \ldots, X_n$ or $Y_1, Y_2, \ldots, Y_n$. In the study of numerical data, perhaps the most elementary and, in many respects, the most important concept is that of the *arithmetic mean* or average of n observations. Convention dictates that the *arithmetic mean,* henceforth referred to as the mean, be denoted by the symbol in use, covered by a bar, and the *mean* of $Y_1, Y_2, \ldots, Y_n$ is defined to be $\bar{Y} = (Y_1 + Y_2 + Y_3 + \cdots + Y_n)/n$.

It is convenient to denote a sum such as $Y_1 + Y_2 + \cdots + Y_n$ by the symbols $\sum\limits_{i=1}^{n} Y_i$ which are read, "the sum of Y_i from i equals 1 to i equals n." The Greek letter Σ (sigma) is referred to as the *summation sign,* and in summation notation the mean of $Y_1, Y_2, \ldots, Y_n$ becomes

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

$\bar{Y}$ is a measure of the *typical value* for the observations $Y_1, Y_2, \ldots, Y_n$, and although there are other measures of the typical value, we shall confine our attention to the mean.

The term *population* is used in statistical literature to denote the set of all possible experimental units or the set of all numerical values corresponding to the experimental units. A subset of a population of experimental units, or the observations $Y_1, Y_2, \ldots, Y_n$ corresponding to a characteristic that they possess, is called a *sample* from the population. The number n of experimental units in the sample is called the sample size. To illustrate the idea, consider the three distances, recorded in feet, for a shot putter in a track and field event: $Y_1 = 57.6$, $Y_2 = 52.4$, and $Y_3 = 59.2$. Conceivably, there are a great number of different distances that might have been attained. The three sample values recorded constitute a subset of the set of all possible distances that might have been attained. $\bar{Y}$ in this illustration is 56.4. Behind the scenes there is another number playing the role of the typical value for the population of all possible distances. This number we denote by the Greek letter μ and it is called the *population mean* or the *expected value* of the population.

In statistical studies, the word *parameter* is used to convey the idea of a fixed quantity in a given experimental situation, with the understanding that if the experimental situation is changed, the quantity may take on another fixed value. The population mean μ is an example of a parameter. If the experimental situation is changed, for example, by changing the weight of the shot in the track and field meet, then the mean of the population of all distances may take on another fixed value.

1.3 SOME PROPERTIES OF SUMMATION NOTATION

As indicated earlier, $\sum_{i=1}^{n} Y_i$ denotes $Y_1 + Y_2 + \cdots + Y_n$. Contemplating the state of affairs, a mathematician one day commented that in a sense, much of mathematics appears to be a science of position. What he had in mind can be illustrated by referring to the symbols $\frac{X}{2}$, X_2, X^2, $2X$, and $2/X$. The relative positions of the 2 and the X are indeed very important to the understanding of what the writer has to say. It might also be observed that we really are not dealing with concepts here but with definitions and language.

In discussing the symbols $\sum_{i=1}^{n} Y_i$, we refer to i as the index of summation and say that the index runs from 1 to n. Some other shorthand notations that are really extensions or different ways of looking at summation notation are the following:

$$\sum_{i=1}^{k} X^i = X + X^2 + X^3 + \cdots + X^k$$

$$\sum_{i=a}^{b} i = a + (a + 1) + (a + 2) + \cdots + (b - 1) + b$$

$$\sum_{i=1}^{n} (X_i + Z_i) = (X_1 + Z_1) + (X_2 + Z_2) + \cdots + (X_n + Z_n)$$

$$\sum_{i=k}^{m} cX_i = cX_k + cX_{k+1} + \cdots + cX_m.$$

After the meaning of summation notation is understood and before we try to use the notation efficiently, it is important to learn and then to recognize basic arithmetic properties in this notation. For example,

$$\sum_{i=1}^{n} (X_i + Z_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Z_i.$$

This property will now be verified after which several similar properties, listed in the following exercises, should be verified by using basic arithmetic properties of numbers.

By definition,

$$\sum_{i=1}^{n} (X_i + Z_i) = (X_1 + Z_1) + (X_2 + Z_2) + \cdots + (X_n + Z_n).$$

By the associative law of arithmetic,

$$(X_1 + Z_1) + \cdots + (X_n + Z_n) = X_1 + Z_1 + X_2 + Z_2 + \cdots + X_n + Z_n = (X_1 + X_2 + \cdots + X_n) + (Z_1 + \cdots + Z_n).$$

But

$$\sum_{i=1}^{n} X_i = (X_1 + X_2 + \cdots + X_n)$$

and

$$\sum_{i=1}^{n} Z_i = (Z_1 + Z_2 + \cdots + Z_n).$$

Thus, by substitution, we have

$$\sum_{i=1}^{n} (X_i + Z_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Z_i.$$

1.4 EXERCISES

Verify that

(1) $$\sum_{i=1}^{n} (X_i - Z_i) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Z_i$$

(2) $$\sum_{i=1}^{n} cX_i = c \sum_{i=1}^{n} X_i$$

(3) $$\sum_{i=1}^{n} c(X_i + Y_i + Z_i) = c \sum_{i=1}^{n} X_i + c \sum_{i=1}^{n} Y_i + c \sum_{i=1}^{n} Z_i$$

(4) $$\sum_{i=1}^{n} c = nc$$

(5) $$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$