An introduction to the principles of statistics, with emphasis on the statistical design and analysis of experiments

R. LOWELL WINE

STATISTICS FOR SCIENTISTS AND ENGINEERS

STATISTICS
for
SCIENTISTS and ENGINEERS

R. LOWELL WINE
Professor of Statistics
Hollins College

PRENTICE-HALL, INC., Englewood Cliffs, N. J.

Library of Congress Catalog Card Number 64-17363
Printed in the United States of America
84611 C

PREFACE

This book is designed as a beginning one year textbook in modern statistics with elementary calculus a prerequisite. It should be useful for *anyone* wanting to learn statistics starting with first principles. It is an expanded version of lecture notes used in a two-quarter course (three lectures per week) in statistics taught to juniors, seniors and first year graduate students in all areas of science and engineering at Virginia Polytechnic Institute and the University of Virginia. The material was also used several times with industrial research groups.

In writing this book I have tried to keep a balance between mathematical (theoretical) statistics and applied statistics. Many of the concepts are introduced by examples from applied statistics after which the concepts are formulated in mathematical terms and given a theoretical treatment. By presenting the material in this way, it is hoped that the reader will gain some real insight into the nature of statistics and at the same time learn how to apply the statistical procedures to actual experimental situations.

There are good statistical books for research workers in science and engineering, but generally speaking they are not very useful as a first introduction to statistics. In the first chapters of this book the reader is given an introduction and grounding in the foundations of statistics along with many examples and exercises. After this, some of those topics which engineers and scientists find most useful are introduced and developed. It is not intended that the topics selected for discussion be given an exhaustive treatment, but rather that they be developed to the point where they are useful to the practitioner of statistics. For the reader interested in special techniques or more sophisticated methods, some exercises and references have been given at the end of each chapter. (For example, very little is said about sampling techniques, quality control and acceptance sampling. There are already good textbooks in these areas.)

The examples are selected to illustrate the principles presented. That is, they are selected so that the principles of statistics can be understood without special knowledge of a particular subject matter field. These so-called common-sense examples refer to such things as heights, weights, tensile strengths, teaching methods and scores, and measures of objects which should be familiar to readers with at least two years of college.

Each chapter contains a copious supply of exercises which are applicable to many and widely different fields of science and engineering. Any reader should find numerous exercises of special interest. The exercises are designed to give the reader practice in applying the material presented, to extend both his practical and theoretical concepts of statistics, and to encourage him to look carefully at new concepts which are outlined only in exercises.

The book may be used as a text for either a theoretical or applied course in statistics. As a theoretical text emphasis should be given to Chaps. 3, 4 and 5, to those sections of the remaining chapters which deal primarily with the mathematical development, and to those exercises which stress proofs and new concepts. As an applied text the proofs and mathematical development may be cut considerably and the sections and exercises on application stressed.

This is a book on statistics. It should be useful to anyone learning the problems of numerical analysis in experimentation or planned investigations. The book should also be useful for anyone seeking some of the foundations of statistics and the way in which part of the statistical structure may be developed.

Many groups and individuals have helped in developing this book. First, I wish to gratefully acknowledge my thanks to Dr. John E. Freund for reading the manuscript and making numerous helpful suggestions, many of which have been included. I wish to acknowledge my gratitude to the several classes of students who read and made helpful suggestions on parts of the manuscript; to the United States Weather Bureau for data used in several exercises; to friends in the Celanese Corporation of America for data (coded) and advice on special problems; to friends in the West Virginia Pulp and Paper Company, United States Steel, and White Sands Missile Range for the opportunity to see applications to special problems; to the Statistics Department of Virginia Polytechnic Institute for typing several chapters of the first draft of the manuscript; to my colleagues in the Mathematics and Psychology Departments at Hollins College for valuable criticism and suggestions; to Miss Margaret Shinnick for reading most of the manuscript and commenting on style and content; to David and Suellen Wine, my son and daughter, for obtaining the data in Tables 5.2 and 5.5; and to many other persons who have published data which are reproduced and acknowledged at the appropriate places in the text. Further, I am indebted to the Danforth Foundation for a travel grant and to the personnel of the Hollins College library for their kind assistance on numerous occasions.

Finally, I wish to express my appreciation to Professor E.S. Pearson for his kind permission to reproduce Tables III, IV and VII from *Biometrika*, Table IX from *Biometrika Tables for Statisticians*, and Table X from *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples;* and to McGraw-Hill, Inc. for kind permission

to reproduce Table VIII. I am indebted to Professor Sir Ronald A. Fisher, F.R.S., Cambridge and to Dr. Frank Yates, F.R.S., Rothamsted, also to Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint Table VI from their book *Statistical Tables for Biological, Agricultural and Medical Research.* Also, I wish to express my appreciation for permission to reproduce Tables 8.1, 8.2 and 10.9 from the *Journal of the American Statistical Association;* Tables 8.3 and 8.6 from books published by McGraw-Hill, Inc.; Tables 10.10 and 16.3 from *Biometrics;* Table 16.1 from a publication by the American Cyanamid Company; and Table 16.5 from the *Annals of Mathematical Statistics.*

R. LOWELL WINE

Roanoke, Virginia

CONTENTS

1

INTRODUCTION

1.1. A DEFINITION OF STATISTICS

Modern statistics is a new and vigorous discipline. It is so new that some of the men who were most instrumental in establishing statistics as it is known today are still actively engaged in research and teaching. Its vigor can be attested from the fact that statistics is growing so rapidly that it is impossible to incorporate many of the latest techniques in a textbook, for by the time the last section is written the first chapters already need revision.

Statistics is playing an increasingly important role in research activities. For this reason it is necessary that special training in statistics be given as early as possible so that experimentation and scientific investigations do not suffer. The study of statistics should not be viewed as just another area of study which is merely desirable for the scientist and engineer; instead, statistics should be viewed as a very sensitive instrument which is capable of successfully coping with many of the most difficult problems posed by modern investigations. Ignoring the use of statistics in many of our research activities today should no more be tolerated than that of ignoring tractors and combines in the wheat fields of Kansas or of ignoring the latest drugs in the treatment of ailments.

The term "statistics" is old, but its present-day interpretation is very young. The term no longer simply refers to the collection and compilation of data; instead, *statistics* is often called *the science of decision-making in the face of uncertainty*. It has to do with both the deductive and the inductive process, that is, both mathematical and scientific procedures. Statistics currently deals with the theoretical development and application of methods suitable to numerical measurements.

Whenever data are collected, statistical methods may be used. In fact, anyone who attempts to work with data acts like or has occasion to act like a statistician. Statistics is a science, based upon mathematics, which deals with such problems as (1) planning a program or an experiment for obtaining data so that reliable conclusions can be drawn from the data, (2) tabulating and analyzing the data, (3) deciding what interpretations and conclusions can properly be drawn from the data, (4) determining to what extent the conclusions are reliable, and (5) justifying by mathematics the methods used in (1), (2), (3), and (4). *Statistical methods* are those procedures used in designing and planning experiments and in collecting, analyzing, and interpreting data. *Statistical theory* has to do with the mathematical development and justification of the methods used.

Statistical methods may be thought of as falling in two classes. Those methods which are used more meaningfully to describe a set of data but which do not involve generalizations are commonly called *descriptive statistical methods.* Those methods which are used on a relatively small set of data to generalize concerning the nature of a much larger set of possible data make up methods of *statistical inference.*

Descriptive statistical methods or, simply, *descriptive statistics*, include those methods which are used in making and describing such well-known objects of our everyday experience as graphs, charts, and tables. Such examples as the batting average of leading hitters, defense-spending graphs, airplane travel charts, stock market averages, census figures, production of automobiles by months, and the index of living costs represent only a few of the illustrations of descriptive statistics we see regularly. Thus, many of the results and techniques of descriptive statistics are known to most of us.

The methods of statistical inference are not so well-known, even though illustrations of their use are fairly common. We read, for example, that the Gallup poll makes a survey and predicts that Joe Brown will be elected governor instead of Sam Jones, or that Kinsey makes some inference about the sex habits of the American female, or that it has been proved that one brand of cigarettes contains less tar than other brands, or that a manufacturer claims that the average life span of a certain type of light bulb is 2500 hours. In each case, we read, and perhaps take issue with, the conclusion, but we know little or nothing about the methods used in arriving at these inferred statements.

The student is cautioned against thinking that methods of descriptive statistics and statistical inference are always distinct and clearly defined. As a matter of fact, most methods which are used in descriptive statistics are also applied in statistical inference. The two terms are generally used with reference to the *kinds of problems* we wish to consider, not with reference to particular formulas or series of formulas. For example, the "average number of defective parts" of a manufactured article might be used in either sense. The term "average" may *describe* the number of defective parts manufactured on a given day, or it may be used to *infer* the number of defective parts which will be produced per day during the remainder of the year.

Since statistics in some way touches so many of our daily activities, it is not possible to give a descriptive and short definition of the term. However, for our purposes it is probably adequate to think of *statistics as both a pure and an applied science which is involved in creating, developing, and applying procedures in such a way that the uncertainty of inferences may be evaluated in terms of probability.* It should be noted that deductive techniques, as used in mathematics, are required in developing the procedures.

1.2. SCIENTIFIC METHOD AND APPLIED STATISTICS

The student will soon realize that the selection and application of some statistical methods, particularly those used in the analysis and design of experiments, are similar to what the scientist and engineer do when setting up a hypothesis, planning and conducting an experiment, and testing the hypothesis by using the experimental data. In both the scientific and statistical disciplines one is concerned with such things as planning and analyzing an experiment in such a way as to establish a fact (the hypothesis) within the framework of a specific theory. In addition to this, the statistical discipline generally requires that a measure of the degree of uncertainty, in terms of probability, accompany the inference drawn from the experiment.

Even if the experimenter understands the theoretical framework and knows specifically what hypothesis he wishes to test, it is still not always obvious which collection of statistical techniques (methods) should be applied. For just as there are many ways to get from Denver to Boston, say, there are usually many ways to collect and use data statistically to justify a statement within a fixed theoretical framework. The investigator normally first looks for a relatively short and relatively simple procedure, but these are not the primary considerations. He wants to draw the correct conclusion within the framework of the experiment in the most efficient way, taking into account, among other things, time, money, and relative importance of the investigation. Thus, in addition to knowing several alternative statistical routes, the investigator must make a decision about the best one to select. In other words, he must decide which "statistical model," including a group of accompanying techniques, to use in the experiment, it being understood that *a model is an idealization of a particular experimental situation.*

There are many statistical models which may, or may not, be useful in solving real problems. For example, the normal curve may be used as a model in describing the distribution of the grades of all beginning mathematics students at a large engineering school. A model is not always so simple, nor is it always used in such a straightforward way, but the objective in using a model is always the same—to make the concept, analysis, and conclusion simpler to understand and to disseminate. Once the model is selected, the resulting conclusion can be relied on only to the extent that the model approximates the situation being studied.

Statistical models have not been constructed for many possible experimental situations. Thus, it may be desirable in a specific investigation to construct and develop the properties of a new model, and this is much easier when one already knows something about statistical models and the associated procedures.

1.3. A BRIEF HISTORY OF STATISTICS

Even though data have been compiled almost from the beginning of recorded time, the science of statistics has never been so broad in its scope as it is today. At first, statistics seems to have consisted in census taking. About 3050 B.C. the ancient Egyptians collected data concerning wealth and population before building the pyramids. There are two censuses of the Israelites recorded in the book of Numbers. In 594 B.C. a census was taken in Greece for the purpose of levying taxes. There are many other records of census taking in most countries of the world from early times to the present.

In addition to taking the census, the Romans prepared surveys of the entire country and kept records of births and deaths. After the Middle Ages, certain individuals as well as governments started keeping records on such things as wealth, armies, commerce, laws, and national resources.

About the middle of the eighteenth century, Gottfried Achenwall, a professor of philosophy in a German university, first used the word *statistik*, and the name *statistics* was introduced into England by E. A. W. Zimmerman about 1787. The Royal Statistical Society of London was founded in 1834 and the American Statistical Association in 1839; each of these societies holds meetings periodically and publishes papers of current interest. Thus, anyone who is interested may follow the growth of statistics over the last 125 years by looking at the records of these societies. In the first number of the *Journal of the Royal Statistical Society*, issued in 1838–39, we read, "Statistics may be said, in the words of the prospectus of this society, to be the ascertaining and bringing together of those facts which are calculated to illustrate the conditions and prospects of society."

The *theory of probability* and the *normal distribution* have been very important in the development of statistics, and they are now of primary importance in the theory and application of statistics. In the middle of the seventeenth century, some gambling experiences with a particular game of dice led Chevalier de Méré, a French nobleman, to consult the famous Blaise Pascal (1623–62) concerning the most advantageous way to bet. Pascal solved the problem, and de Méré posed another problem which Pascal investigated. This led to a private correspondence with the French lawyer-mathematician Pierre de Fermat (1601–65) and to the first foundations of the theory of probability. After becoming acquainted with the contents of this correspondence, Christiaan Huygens (1629–95) developed some new ideas and in 1654 published a first book on probability. Jacob Bernoulli (1654–1705), Abraham de Moivre (1667–1754), and Pierre Simon Laplace (1749–1827) made great contributions to the early theory and application of probability. Abraham de Moivre is responsible for the equation of the normal curve (1733). Much later, Laplace and Karl Friedrich Gauss (1777–1855) developed the same results independently of each other.

Laplace in his studies of the origin of comets seems to have been the first to attack problems relating to rules of "inductive behavior," that is, to the adjustment of our behavior to a limited number of observations. The geologist Charles Lyell (1797–1875), the biologist Charles Darwin (1809–82), and the monk Johann Gregor Mendel (1822–84) in his experimental breeding of plants based some of their work on statistical arguments. None of these men was a statistician, and they did not spend their energies in placing statistics on a firm foundation.

Karl Pearson (1857–1936), initially a mathematical physicist, after becoming interested in evolution spent nearly half a century in serious statistical research. He helped to found the journal *Biometrika* and gave the study of statistics its first great impetus. Sir Ronald Fisher (1890–1962) made many important contributions to statistics. Since the 1920's Fisher and his students have also stimulated great interest in applications of statistics in many fields, particularly agriculture, biology, and genetics. Some of the basic theory on hypothesis-testing was presented by J. Neyman (1894–) and E. S. Pearson (1895–) as late as 1936 and 1938.

Thus, it was not until early in this century that statistics started to be used to any large extent outside of census taking and other specialized areas such as genetics and astronomy. Since the late 1920's interest in the application of statistical methods to all types of problems has grown rapidly. It is interesting to note that prior to the 1920's applied statistics was predominately descriptive in nature, and that since the 1920's statistical inference has grown to constitute nearly all of statistics. Indeed, today purely descriptive methods play a very minor and almost incidental role in statistical applications.

It would be pointless to try to enumerate all the areas in which statistical