INTRODUCTION TO

STATISTICS

Robert G.D. Steel / James H. Torrie

Introduction to STATISTICS

Robert G. D. Steel
Professor of Experimental Statistics
School of Physical Sciences
North Carolina State University

James H. Torrie
Professor of Agronomy
College of Agriculture
University of Wisconsin, Madison

McGraw-Hill Book Company
New York St. Louis San Francisco Auckland Düsseldorf Johannesburg Kuala Lumpur London Mexico Montreal New Delhi Panama Paris São Paulo Singapore Sydney Tokyo Toronto

INTRODUCTION TO STATISTICS

1234567890DODO79876

This book was set in Times Roman.
The editors were A. Anthony Arthur, Alice Macnow, and Shelly Levine Langman;
the designer was Scott Chelius;
the production supervisor was Angela Kardovich.
The drawings were done by J & R Services, Inc.
R. R. Donnelley & Sons Company was printer and binder.

Library of Congress Cataloging in Publication Data

Steel, Robert George Douglas, date
Introduction to statistics.

(McGraw-Hill series in probability and statistics)
Bibliography: p.
Includes index.
1. Mathematical statistics. I. Torrie, James
Hiram, date joint author. II. Title.
QA276.S79 519.5 75-22235
ISBN 0-07-060918-7

Introduction to

STATISTICS

McGraw-Hill Series in Probability and Statistics

David Blackwell and Herbert Solomon, Consulting Editors

BHARUCHA-REID Elements of the Theory of Markov Processes and Their Applications
DE GROOT Optimal Statistical Decisions
DRAKE Fundamentals of Applied Probability Theory
EHRENFELD AND LITTAUER Introduction to Statistical Methods
GIBBONS Nonparametric Statistical Inference
GRAYBILL Introduction to Linear Statistical Models
HODGES, KRECH, AND CRUTCHFIELD StatLab: An Empirical Introduction to Statistics
LI Introduction to Experimental Statistics
MOOD, GRAYBILL, AND BOES Introduction to the Theory of Statistics
MORRISON Multivariate Statistical Methods
RAJ The Design of Sample Surveys
RAJ Sampling Theory
STEEL AND TORRIE Introduction to Statistics
THOMASIAN The Structure of Probability Theory with Applications
WADSWORTH AND BRYAN Applications of Probability and Random Variables
WASAN Parametric Estimation
WOLF Elements of Probability and Statistics

In Memoriam
Jonathan Chester Steel
1954–1975

Preface

This is an introductory text, presenting some of the basic concepts and techniques of statistical inference. These concepts and techniques have played an important role in a long list of very diverse fields, in fact, wherever one finds observational or experimental data, whether it be counts or measurements. It is because of the chance element in the origin and nature of data that statistics has become such an important tool, since statistics depends upon probability, which is concerned with chance.

The first four chapters of the book deal primarily with the presentation and summarization of large amounts of data; they indicate the need to distinguish between samples and populations and to consider the possibilities of drawing inferences about the latter from the former. At the same time, some of the procedures given are regularly applied to small samples.

The introductory chapters are followed by probability theory. It is here that the notions of sample space, random variable, and probability function are introduced. Several frequently encountered distributions are considered at some length.

The general concepts of statistical inference, namely the estimation of parameters and testing of hypotheses, are next discussed, with special reference to and examples for the binomial and normal distributions. However, each aspect of inference once applied to a specific distribution is then looked at from a nonparametric point of view. Consequently, the students' understanding of the topic as originally introduced is reinforced.

The analysis of variance and regression analysis are included, as are nonparametric alternatives to these techniques.

At no time should one find a need for more than a working knowledge of high school algebra. Even when such algebra is used, as in finding the expectation of a binomial sum, one can generally go straight to the results. For those who enjoy algebra, there are a number of problems with algebraic content.

The book was written to provide sufficient material for a two-semester course. However, Chapters 1 to 13, with some topics deleted, together with topics selected from later chapters, perhaps 15 and 17 in particular, can serve as a one-semester

course. For those instructors with an interest in nonparametric statistics, Chapters 14 and 18 can substitute for 15 and 17. Section titles should help with one's choice of topics to be deleted or included.

We are indebted to the literary executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint Table III from their book "Statistical Tables for Biological, Agricultural and Medical Research."

We are also indebted to E. S. Pearson and H. O. Hartley, editors of "Biometrika Tables for Statisticians," vol. I, and to the *Biometrika* trustees for permission to make use of Tables 8 and 18; to John W. Tukey and to Thomas E. Kurtz and Prentice-Hall, Inc., for permission to reprint Tables B-4 and B-5; to the late Frank Wilcoxon, Roberta A. Wilcox, and the American Cyanamid Company for permission to reprint Table B-6; and to Dr. James H. Goodnight for producing Table B-8 for us.

In addition, we are indebted to many authors, journal editors, and publishers for their generous permission to use the data that appear in the text and exercises. Their names are listed in the references.

Finally, we are grateful to a number of people for their contributions during the preparation of the manuscript, in particular to Mrs. Dorothy Green for her careful typing, and to Marcia and Jonathan Steel for their helpful suggestions toward a final draft.

ROBERT G. D. STEEL
JAMES H. TORRIE

Contents

Chapter 1
INTRODUCTION

1-1 YOU AND STATISTICS

Statistics plays a part in our daily lives in numerous ways—ways we may not always appreciate. This is a relatively new phenomenon, and is a consequence of the many problems that science and technology have chosen to study and deal with using statistical methods. Notice that the word "statistics" is used in the singular. The term no longer refers simply to the collection and presentation of data in tables, graphs, and descriptive figures. Today, statistics is a subject or discipline. As such, it is very much involved in the physical and social sciences, in business and industry, and indeed in most human activities that require decisions involving uncertain outcomes.

Statistics is used extensively in the production of many of our foodstuffs. The flour you use is most likely the result of breeding experiments with wheat, resulting in the selection of varieties or cultivars possessing high quality, and of baking tests requiring careful measurements, at various stages, of the appropriate

properties. Both the breeding techniques and the baking tests are based on sound statistical procedures. The drinking water we use so casually for so many purposes is almost certainly under continuous statistical quality control. For much of our clothing, rigorous statistically analyzed trials have determined what mixtures of fibers and what insulating materials provide a reasonable compromise between maximum comfort and minimum care.

The paint on your house may have been selected for the market as a consequence of a weathering trial conducted according to statistical guidelines at a forest products or industrial chemistry research laboratory. The building code of your community may have been changed recently on the basis of a statistical comparison of a long-used building material with one developed since the code was adopted.

Consumer protective agencies evaluate statistical evidence that may lead to the removal from the market of clothing that could prove to be a fire hazard, food additives that could cause genetic damage, and drugs with undesirable side-effects.

Unfortunately, statistics in the form of numerical "facts" is sometimes used to confound us. For example, we may be told that a gasoline with a special additive has given up to 3 miles per gallon more than the average of a number of leading brands without the additive. The advertisers have not said that their average mileage is any better than the average given by the leading brands, though they almost certainly hope we will conclude so. Sometimes our would-be confounders compare wage rates or crime rates where definitions have changed over time or place. They may compare sales for consecutive time periods when the only significant comparison would involve comparing the same period in consecutive years, and they may quote percentages without saying whether the figures were based on a small or a large number of observations.

As a consequence of the ever-present role that statistics plays in our lives, it is a virtual requirement that we attain some degree of statistical literacy. Without such knowledge, we will be unable to interpret satisfactorily a reasonable number of the statements and situations with statistical content with which we come in contact.

1-2 WHAT IS STATISTICS?

Already we have implied that statistics is concerned with knowledge in the form of numbers relevant to various questions. The numbers initially collected are organized, processed arithmetically, interpreted, and finally presented in tabular or graphic form or as summary numbers such as averages or percentages. Usually interpretation involves a conclusion or inference.

The originally collected numbers or *observations* are called, collectively, *data*. Clearly they are almost invariably a fraction or *sample* of a larger set of possible observations. For example, only a small amount of flour will be used to provide an observation or *sample value* in a baking test. The same is true in measuring water quality, house-paint weathering, or gasoline mileage.

The set of observations which constitute all possible measurements is called a *population*.

In the cases mentioned, it is easy to visualize the population as consisting of more observations than we can number, that is, as an *infinite population*. However, if we are sampling United States income tax returns for 1974, say for the purpose of auditing them, we can number them all in sequence and so have a *finite population*. Note that the term population applies to the data rather than to the people or objects which are observed to provide the data.

If a census or complete enumeration of the observations in a finite population exists, one can find the true average for that population. This is a constant and not subject to any variation. Thus, if we have complete census data on ages in the American population, the computed average age will be the true average for the time of the census. In the case of an infinite population, we cannot have the numbers to do the arithmetic of finding a true average, but we can conceive of such a value which, like that for the finite population, is a constant.

Notice that for a population to exist at all, it must first be capable of definition. Definition of any population is a matter for careful thought. It involves providing a rule that will identify every unit on which an observation might be made. For example, what will be our rule to identify "farmer" for a study of farm incomes in a county? How do we identify each and every dwelling unit in a city block for a study of housing density? Answers, of course, depend on the aims of the investigation.

Any inference drawn from a sample must, of course, be one concerning the population sampled. Thus, if we want a sample of registered voters, we go to voter rolls rather than to telephone listings. Since we are reasoning from the part or sample to the whole or population, the end result is an *inductive* or *uncertain inference*. Because of sample-to-sample variation, even within the same population, it is evident that our inferences cannot be completely dependable; they must be qualified in some manner. To make sure that most of our inferences about descriptive measures like averages will be valid, we must have good samples, samples resulting from sound sampling schemes and experimental designs.

What is a good sample? We would certainly like any sample to be typical or representative of the population. However, there is no way to determine what typifies a population unless we can examine all of it, a course of action not available to us. Consequently, we need a technique to tell us, at least, how well our sampling and inferential procedures work. When the laws of probability theory apply, this theory provides such a technique.

Random sampling is done when each possible sample has the same probability of being selected or when its probability of being selected is known. As a consequence, the laws of probability can be applied.