Statistics: A Foundation For Analysis

Hughes — Grawoig

ANN HUGHES · DENNIS GRAWOIG
Georgia State University

Statistics: A Foundation for Analysis

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts · Menlo Park, California · London · Don Mills, Ontario

 Library of Congress Catalog Card No. 76–133891.

Statistics:
A Foundation for Analysis

Preface

Statistics: A Foundation for Analysis is addressed primarily to an audience of advanced undergraduate or beginning graduate students in business and economics. Its objectives are (1) to develop in these students an understanding of the basic concepts and techniques of both traditional and modern statistical methodology and the fundamental probability theory upon which these are based, (2) to produce in the students a degree of expertise in the use of these basic techniques, (3) to cultivate an appreciation for statistical methods as tools for analysis, and (4) to establish the foundation from which more advanced work in statistical analysis may be pursued.

The text strives to present as briefly, yet as clearly, as possible a reasonably comprehensive discussion of the more commonly used statistical methods and to illustrate their application in a business context. The opening chapters discuss the theory of probability from both the mathematical and the logical points of view. A strong effort has been made to develop a presentation of probability that is elementary and orderly, palatable to the nonmathematician, yet consistent with the mathematical theory of probability. Upon completion of these chapters the student will feel at ease with the concepts of a random experiment, a random variable, and likelihood-of-occurrence. The inclusion of this topic makes the book self-contained for those students who have not previously studied probability theory. Those students who have completed a course in probability should find that these chapters provide a useful review and reference source for those aspects of probability theory that are most useful in the study of statistics.

Several of the more frequently used probability models are examined in Chapters 5 and 6. Next the role played by sampling and sampling distributions in the classical approach to statistical inference is explored in Chapter 7. The concepts of estimation (Chapter 8) and hypothesis testing (Chapter 9) are carefully developed and then used in making inferences about means, proportions, variances, and about differences in parameters (Chapters 10, 11, and 12). Analysis of variance (Chapter 13) is treated as a logical extension of these techniques.

Both simple and multiple, linear and nonlinear regression and correlation models are presented (Chapters 14, 15, 16, and 17). Topics not always included in an introductory text—such as inference in correlation methods—have been included.

Nonparametric methods are covered in Chapter 18. These methods have been included so as to provide the student with a kit of tools which may be used when the more traditional parametric techniques of inference are not applicable.

Decision theory (Chapter 19) is introduced as an expansion of classical statistical methods. Concerned primarily with the logical analysis of choice between possible alternative courses of action when the consequences of these actions depend upon an unknown state of nature, decision theory explicitly incorporates in a decision model important assumptions that the classical theory of inference deals with only implicitly.

Throughout the text a determined effort has been made to emphasize the idea that each of the statistical techniques presented is closely related to the others rather than to treat each as separate or isolated subject matter. The various topics are treated as logical extensions of a basic foundation; it is the differences in the situations where the techniques are appropriately applied that are stressed.

The book contains ample material for a one-year course, whether it be a two-semester or a three-quarter sequence. However, the book may be used advantageously in many other types of courses. The following comprise only a few of the possible arrangements.

1. *In a three-quarter sequence.* The foundations of probability theory (Chapters 1–6) should be carefully built in the first quarter; statistical inference (Chapters 7–12) should be covered in the second quarter; and special topics in statistical analysis such as analysis of variance, regression, nonparametrics and decision theory (Chapters 13–19) should be covered in the third quarter. These latter topics provide fertile ground for individual term projects which always enhance the student's appreciation of statistical methodology.

2. *In a two-semester sequence.* Chapters 1 through 9, probability and inference, should be included in the first semester and given thorough treatment; the remaining chapters provide ample material for a rigorous second semester.

3. *For a two-quarter course where the student has not previously studied probability theory.* Probability and an introduction to inference will be sufficient material for the first quarter (Chapters 1–9, possibly omitting Chapter 4). In the second quarter, a further exploration of inference techniques should be undertaken with selected materials from Chapters 10–13, as well as Chapters 14 and 15.

4. *For a two-quarter course where the student has previously studied probability theory.* Statistical inference (Chapters 7–12) would be the subject matter for the first quarter. Analysis of variance (Chapter 13) could possibly be included

as the natural sequel to inference from two populations. Then selected topics from the remaining chapters would be covered in the second quarter.

5. A one-semester course for students with a firm grounding in probability fashioned along classical lines could use the first six chapters as a reference source and begin study with the introduction to sampling distributions, treating inference techniques in depth while eliminating selected topics in the concluding chapters.
6. A one-semester course for graduate students with undergraduate credits in statistics might cover Chapters 7–15, 18, and 19, which would give them a survey of inference, analysis of variance, regression, nonparametrics and decision theory.

Still other combinations might be selected to meet the requirements of specific groups of students.

Today's students in colleges of business administration are required to attain a degree of proficiency in the use of mathematical techniques. They are being taught set theory and matrix algebra; they are being introduced to the calculus. Yet they certainly do not view these techniques from the same level of sophistication as do students in engineering or mathematics departments. Thus, while this text attempts to capitalize on the student's mathematical background, it has as formal mathematical prerequisites calculus only through integration and differentiation of simple polynomial and exponential functions and the use of matrix algebra in solving equations. (Appendixes have been included which the student may use for review and reference on these topics.)

An effort has been made to reduce the shock of mathematical notation, often a difficulty for the beginner in statistics, by proceeding carefully and gradually in the introduction of new and unfamiliar symbols. We have sought to adopt notation of well-established usage. All symbols are defined when they are first introduced. A syllabus of symbols introduced in each chapter is included at the close of that chapter. Notation is sufficiently simple and precise so that even when the book has been used for only a short period of time, the student should gain confidence in the use of these symbols.

Even though mathematical notation is employed in a precise statement of the concepts, the basic ideas are always explained carefully and are illustrated by numerous examples. In this manner, the underlying statistical theory is fully exposed and the relation between theory and application thoroughly explored.

For each topic a large number of examples are worked out completely and step-by-step explanations are included. It is hoped that these examples will teach the student the methodology of statistics and its application to practical situations as well as the proper approach to problem solving.

Problem material designed to ascertain whether or not the student understands the material just presented is included at the end of each section within the chapters. The student should attempt to work these problems before continuing to the next sections. Many additional exercises and problems are given at the end of each chapter. Most of these problems are designed to draw together the most important concepts presented in the chapter and to develop a thorough comprehension of the subject matter. Realizing that our students are primarily business school students and proceeding on the premise that a first course in statistics should be within the subject matter field, we have taken our illustrations for the most part from the business arena and have attempted to relate material being presented to applications which are meaningful to students in business administration and economics.

In *Statistics: A Foundation for Analysis* we have attempted to provide a conceptual approach to statistics. We have avoided both a "cookbook" presentation and a "proof-for-proof's sake" approach. Believing that it is difficult for a student to use a formula with full confidence until he has been given some appreciation of its derivation, we have sought to enhance the reader's understanding of the subject matter by always explaining mathematical proofs when this is deemed feasible. Our goal is to enable the student to use statistical formulae with that level of confidence that attaches to a full understanding of what the formula is and why it works. We are not striving to give the student an appreciation for mathematical proofs. We use them only where we feel that they add meaningfully to the student's understanding of a concept. Yet the presuppositions regarding the student's mathematical competency make it possible for us to use those proofs which we feel are needed and to otherwise discuss the reasoning underlying many statistical procedures whereas this type of explanation would have been impracticable had we assumed a lower level of mathematical proficiency.

Nonetheless, we have eliminated all those facets of mathematical theory that do not illuminate the discussion at hand and have simplified whenever possible the mathematical theory that is presented, believing that simplicity and teachability can be achieved without compromise of theoretical correctness. Moreover, the more technical material has been placed in footnotes or in supplements at the ends of the chapters and only those mathematical explanations immediately required for the comprehension of the material have been incorporated in the body of the text. This manner of presentation has insured the ease of readability of the book.

It is in all of these ways that we have attempted to achieve, through this text, three major goals: (1) to give the student a clear understanding of the theory underlying statistical procedures while still maintaining a "practical" point of view, (2) to develop a statistics text for business school students that is mathematically challenging but not foreboding, and (3) to create a text from which it will be rewarding to teach, and to learn.

No goal is reached except through the interaction of myriad factors; no book is published except by the cooperation of many persons whose names cannot appear on the title page. It is our pleasure to achnowledge that many people have contributed, directly and indirectly, to the development of "our" book.

We have benefited immeasurably from the comments and advice of Robert Winkler (Indiana), Leonard Kent (University of Illinois at Chicago Circle), and Chris Theodore (Boston University). We are grateful also for guidance received from Henry Tingey (University of Delaware), Morris Hamburg (University of Pennsylvania), John Pratt (Harvard), Larry Richards (University of Oregon), and Meyer Belovicz (University of Massachusetts). Their perceptiveness and their helpful suggestions have enhanced the manuscript considerably.

We also wish to thank the staff of Addison-Wesley who made the actual production of the book a pleasure.

And to our colleagues in the Department of Quantitative Methods and the School of Business Administration of Georgia State University, our heartfelt gratitude for a stimulating environment in which to work.

Atlanta, Georgia
November 1970

A.H.
D.G.

Contents

1 The Foundations of Probability

The concept of probability occupies an important place in the decision-making process, whether the problem is one faced in business, in government, in theoretical physics, in astronomy, in the social sciences, or just in one's own everyday personal life. In very few decision-making situations is perfect information—all the needed facts—available. Most decisions are made in the face of uncertainty. Probability enters into the process by playing the role of a substitute for certainty—a substitute for complete knowledge.

Probability is especially significant in the area of statistical inference. Here the statistician's prime concern lies in drawing conclusions or making inferences from experiments which involve uncertainties. The theories of probability make it possible for the statistician to generalize from the known to the unknown and to place a high degree of confidence in these generalizations. Probability theory, then, is one of the most important tools of statistical analysis.

1.1 RANDOM EXPERIMENT DEFINED

Most of the raw data of statistical analyses are generated by processes called *random experiments*. These may be real or hypothesized, but they have certain characteristics that distinguish them from the laboratory experiments which pervade the physical sciences.

From the statistical point of view, a random experiment is a well-defined course of action that may result in one of two or more possible outcomes. Moreover, the outcome of the experiment is affected by chance factors to such an extent that it cannot be predicted in advance with certainty. In other words, even though the experiment is repeated under essentially the same conditions (and this is at least theoretically possible since the experiment is a well-defined course of action), the outcome from each trial will not necessarily be the same. Because of the element of chance operating on the particular outcome of a trial of the experiment, these processes are also called *chance processes* or *stochastic processes*.

A random experiment, then, is a process having these properties:

1. It can, at least theoretically, be repeated.
2. It has a number of possible outcomes, any one of which might occur on a single repetition of the experiment.
3. Because of chance factors affecting the process, the exact outcome of any particular trial of the experiment cannot be predicted with certainty.

Example Perhaps the simplest example of an experiment of chance is the toss of a fair coin. Of course, the experiment can be repeated under essentially unchanging conditions. If the rather unstable "on end" position is regarded as impossible, there are two possible outcomes—"heads" or "tails"—only one of which can occur on a single trial of the experiment. However, one cannot determine analytically—from the initial position of the coin, from the velocity of the toss, from the forces acting on it in flight—which of these two possible outcomes will result on a specific toss of the coin.

Ordinarily the statistician is not so much interested in which of the individual outcomes takes place on a particular performance of the experiment as he is interested in the relative frequency with which the various possible outcomes occur when the experiment is repeated a large number of times. So enters the theory of probability into the statistical process.

1.2 THE OBJECTIVE VIEW OF PROBABILITY

The traditional statistician defines the probability of the occurrence of a specified experimental outcome in terms of its *long-run relative frequency of occurrence.* The ratio of the number of times the outcome takes place to the total number of times the experiment is performed is called the relative frequency of the outcome. The relative frequency with which a given outcome takes place when the experiment is repeated a large number of times is called its *probability.*

Regrettably, such phrases as "in the long run" and "large number of times" are rather vague. It would be desirable if they could be avoided. Actually, a determined effort has been made to define probability of a certain outcome as the *limit* of the relative frequency with which it occurs.

Definition If an outcome occurs f times out of n trials, its relative frequency is f/n; the value which is approached by f/n when n becomes infinite is called the limit of the relative frequency. The probability of an outcome O_i is defined as the limit of its relative frequency; that is

$$P(O_i) = \lim_{n \to \infty} f/n.$$

Alas, such an attempt to so rigorously define probability runs into difficulty. The definition has only a conceptual interpretation and not an operational one since, in the real world, the experiment can be repeated only a finite number,