STATISTICS

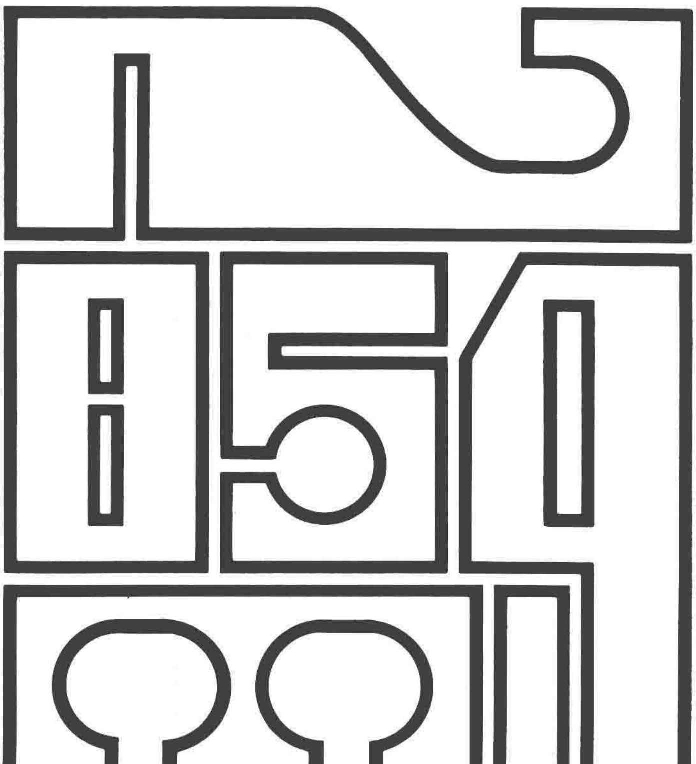
for the social sciences

WILLIAM L. HAYS

SECOND EDITION

STATISTICS
for the social sciences
SECOND EDITION

William L. Hays
University of Georgia



HOLT, RINEHART AND WINSTON, INC.
New York Chicago San Francisco Atlanta
Dallas Montreal Toronto London Sydney

Preface: To the Teacher

The first edition of this book was addressed to students in psychology, and most of the examples and explanations were designed with that audience in mind. Happily, the book has also appealed to a wider group, and has had a favorable reception by students and teachers in most of the other social and behavioral sciences. In view of this reception, I have been persuaded by the publisher to give the new edition a title which suggests its usefulness beyond the immediate psychological community.

On the other hand, I have not taken the further step of trying to rewrite the book so as to appeal to every conceivable user, since I felt this would destroy its essential unity of approach. Hence, the book often reads as though it were still addressed to students in psychology. I do not feel this to be any great barrier to its use by students in other fields, however. The examples cited are by no means exotic in their psychological content, and most of the issues discussed and the methods presented occur in all fields of social and behavioral research.

The aim of the book remains precisely the same as outlined in the preface to the first edition: to give the elements of modern statistics in a relatively nonmathematical form, but in somewhat more detail than is customary in such texts, and with considerably more emphasis on the theoretical than upon the applied and computational aspects of the methods. Indeed, I feel that the utility of such an approach may be even greater now than it was ten years ago. A large proportion of statistical analysis is now done on the computer, and the research worker relies considerably on a library of ready-made statistical programs. The problem is no longer so much "how to do it" as it is how to select an appropriate technique which will give results in the format needed. Nevertheless, the interpretation of a result remains just as big a problem as before. It now seems increasingly important for the student to understand the background of the large number of statistical methods available, and to know how and when to carry his problem to a statistician for advice and assistance. The author continues to hope that this book will help the student to learn to do these things.

The content of the book has been expanded in several ways. A good bit of additional space has been given to elementary distribution theory, in order to provide the student with somewhat more groundwork in this area. Several additional families of distributions, such as the Pascal, the gamma, and the beta, are discussed, both because of their intrinsic importance and because they provide some useful techniques. Exercises have been included at the ends of the chapters, along with solutions to odd-numbered problems. Hopefully, these will be useful to student and teacher alike, although teachers will likely wish to expand both the range and the content with additional exercises.

A major change from the first edition is the inclusion of a fairly long chapter on simple Bayesian methods. I believe that this approach and these methods are now too important for an introductory text to ignore. The placement of this chapter at the end of the book reflects my own opinion that a student should have a pretty fair grounding in the classical methods before the Bayesian approach is attempted.

It may be that much of the material on set and function theory is now superfluous for students brought up on the new mathematics. If so, well and good, let these students start with Chapter 2. However, if the new math is retained like the old, then much of this exposure may have worn off by the middle or late university years, and a quick review may be welcomed by many students.

Finally, in addition to all of those whose help I acknowledged in the preface to the first edition, I wish to express my thanks to Professor William Kruskal of the University of Chicago, who contributed many helpful suggestions toward this revision, and especially to Professor J. E. Keith Smith of the University of Michigan, whose careful review of the first edition had much to do with shaping the present version. I wish also to thank Professor E. S. Pearson and the trustees of *Biometrika* for their permission to reproduce selections from the *Biometrika Tables for Statisticians*, Vol. I (3d ed.), and to Professor R. S. Burington and the McGraw-Hill Company for permission to reproduce the table of binomial probabilities from R. S. Burington and D. C. May, *Handbook of Probability and Statistics with Tables* (2d ed.). Most of all, I wish to thank all of the many students and teachers who have so thoughtfully contributed corrections and suggestions over the years. I hope that my work is worthy of you.

W.L.H.

Ann Arbor
November 1972

Preface: To the Student

In its original version, this book was designed for students in experimental psychology. Therefore, many of the examples and much of the discussion deal with issues in that field. Most of these same problems occur in almost identical form in the other social and behavioral sciences, however, and you as a student, whatever your field, should not feel uncomfortable in translating psychological examples and issues directly into your own content area. The psychological examples are simple and easily understood, and every issue having to do with psychological research applies equally well to the other social and behavioral sciences.

In writing this book I assumed that its readers would be serious students just beginning their late undergraduate or early graduate studies. I have tried not to oversimplify or to write down to such students, and I believe that the kinds of students I have in mind will not be dismayed by some tough issues, by some algebraic manipulations, or by the prospect of learning more mathematics. In fact, many students find it interesting, and often exciting, to follow a logical argument and to try to anticipate what the next step will be. I have provided a great deal more explanation and discussion than is customary in statistics texts, largely because I believe that if a student is serious enough to contemplate a research career, he should be also serious enough to want to understand his research tools as fully as possible. I am not so naive as to believe that this will be true of all students, and some of you are going to find this book long-winded, complicated, and *deadly*. To you my sympathy! On the other hand, I am happy to say that many students find the content interesting, and even downright fascinating. These are the students I have in mind as I write. They are the "you" in this book.

Anyone who has had any exposure at all to the social and behavioral sciences does not need to be told that statistics is an important tool in these fields. Statistics serves in at least two capacities. First, it gives methods for organizing, summarizing, and communicating data. Second, it provides methods for making inferences beyond the observations actually made to statements about large classes of *potential* observations. The set of methods serving the first of these functions is generally called **descriptive statistics,** the body of techniques for effective organization and communication of data. When the man on the street speaks of "statistics" he usually means data organized by these methods. However, the major emphasis in this book is on **inferential statistics,** the body of methods for arriving at conclusions extending beyond the immediate data. A large part of the mathematical theory of statistics is concerned with the problem of inference, and with the development of inferential methods. Furthermore, most of the interesting and important applications of mathematical statistics to the sciences concern problems of inference. This book attempts to lay some of the groundwork for an understanding of the origins of inferential methods and their applications to data.

You will soon discover that the main concern in this book is with the mathematical theory underlying inferential methods, rather than with a detailed exposition of all the different methods psychologists and others find useful. The author had no intention of writing a "cookbook" that would equip the student to meet every possible situation he might encounter in his work. Many methods will be introduced, it is true, and we will, in fact, discuss most of the elementary techniques for statistical inference currently in use. However, in the past few years the concerns of the social scientist have begun to grow increasingly complicated. Psychological theory is growing, psychologists are turning their attention to new problems, and techniques for experimentation are becoming much more sophisticated than in the past. The same thing is happening in the other social and behavioral sciences. The statistical analyses required in many such experiments are simply not in the "cookbooks." From all indications, this trend will continue, and by the time that you, the student, are in the midst of your professional career it may well be the case that entirely new statistical methods will be required, replacing many of the methods currently found useful.

As social and behavioral research becomes more sophisticated and mathematical statistics produces more and more methods appropriate to particular situations, a point is rapidly being reached where the research worker simply cannot be familiar with all the statistical methods appropriate to his work. It seems unfair to demand that each competent researcher must also be a competent mathematical statistician as well, although a few gifted individuals (*not* including the author) have somehow found time and brain-power to be both. Furthermore, the advent of electronic computers has opened up new avenues of data-analysis, making it possible to answer questions that were formerly unanswerable because of the sheer arithmetical complexity of the analysis involved. In short, the days when each researcher was his own "do it yourself" statistician, relying on his handy cookbook, are about over.

What, then, is the research scientist to do? He wants the answers that statistical analysis can give him, but he may not know the range of methods open to him within theoretical statistics itself. The answer is very simple: when in doubt, ask a statistician, a man whose principal training and commitment is in mathematical statistics and the development of such methods. A large part of the work of most applied statisticians consists of consultation on problems of design and analysis of experiments, and many are available for such consultation on a professional basis. The statistician can usually provide answers to the research worker's questions, *provided that the statistician is asked about the problem before the data are collected, and can participate in the efficient and logical planning of the experiment.* It is most unreasonable to expect the statistician to reach in his hat and pull out a method that will extract meaning from a poorly designed or executed study.

In order to use the resources of mathematical statistics and statisticians the research scientist must know something about the nature of mathematical statistics. He should be able at least to talk to the statistician in terms they both understand. The statistician does not expect the scientist to know all about theoretical statistics, nor does the scientist expect the statistician to know all about his particular problem. But to work together effectively, each must have some idea of the basic concepts the other uses. This is the reason for the theoretical emphasis in this book. At the very outset, the student needs to know something about the nature of theoretical statistics if he is to appreciate the resources of statistics and not become lost in the complexities of using statistical methods effectively.

This book is not, nor does it pretend to be, a first course in mathematical statistics. Ideally, the serious student in the social or behavioral sciences should take at least one such course. However, there are two practical difficulties: The content and the organization of courses in mathematical statistics are framed for the training of statisticians, not behavioral scientists and the peculiar problems of these research areas are not emphasized in such courses: This is as it should be. In the second place, to become a really good researcher is a full-time job, and the student may not have the time to devote to the mathematical statistics courses and their prerequisites in order to gain the essential background he needs.

Thus, this book contains some of the concepts, results, and theoretical arguments that come from mathematical statistics, but these results and arguments are given at a far more intuitive and informal level than would be the case for a student in mathematics. Only very seldom will the level of mathematics used rise above the high school level, although the mathematical concepts used will occasionally be unfamiliar to most students. In particular we will use some results coming from the application of the calculus, especially results having to do with the idea of a "limit"; these ideas really cannot be treated adequately at an elementary level. From a mathematician's technical point of view, many of our statements are incomplete, poorly framed, or imprecise. On the other hand, many of these ideas can be grasped intuitively by the serious student, and the author feels that this intuitive understanding is better than no understanding at all, *provided* that the student understands the limitations of a presentation such as this.

A number of topics have been included that have little or no direct application to social or behavioral science at this time, because the author feels that these topics do help to clarify some theoretical point. On the other hand, a few topics ordinarily included in elementary statistics books have been omitted, largely because they have rather minor importance and the author preferred to devote space to other matters. Finally, some techniques are included simply because the author feels that you, as a research worker, might want to know these methods. These are techniques that are useful even in elementary experimentation, although they are usually given extensive coverage only in more advanced texts.

The student also should understand that the examples in this book are hypothetical. This, admittedly, goes against current practice in such texts. On the other hand, the author feels that it is more important to have a fairly simple and plausible problem that the beginner in research can understand and that illustrates the method clearly, than to try to provoke the student into exclaiming, "Gee, they really *do* use statistics in research!" Presumably, a student who has had an adequate introductory course in his field knows this already.

A glance at the table of contents reveals the topics covered, and there is little point in a detailed listing here. However, it should be pointed out that the chapters in this book fall roughly into two sections: Chapters 1 through 8 deal very largely with the essential ideas of probability and of distributions, the two central notions of theoretical statistics. The first chapter lays a foundation for these topics by introducing three very fundamental mathematical concepts: set, relation, and function. A clear idea of these concepts can do a great deal to clarify the remainder of the book. Chapters 2 through 8 are very closely related in the topics covered, and each succeeding chapter builds on the concepts introduced in the preceding ones.

Chapter 9 develops some of the issues connected with the actual use of results from theoretical statistics, particularly the problem of making up one's mind from data. Chapters 10 through 18 discuss particular methods for various kinds of inferences to be made in simple experimental situations. Finally, Chapter 19 gives some of the basic ideas of Bayesian statistics, an alternate approach.

A theme that runs throughout this book is the search for relationships through experimentation. A statistical relation will be said to exist when knowledge of one property of an object or event *reduces our uncertainty* about another property that object or event will show. A statistical relation occurs when things tend to "go together" in a systematic way. This theme will recur *ad nauseam* in the chapters to follow, but it is an important one.

Very many mathematical expressions occur throughout this text. These are of three kinds: algebraic equivalences serving as steps in some derivation, actual definitions or principles stated mathematically, and computational formulas useful in some method. Some of the mathematical expressions are numbered; ordinarily this occurs when some reference will be made to that expression at a later point. If the number for any expression is followed by an asterisk (*), then this is an important definition or relationship that is worthy of your special attention. If an expression is primarily a computing formula, then this will be given a dagger (†) following the number.

A few words must also be said about the symbols we will use. Generally, when a new symbol is introduced, it will be given an "on the spot" definition. However, there are a few symbols in such widespread use that the author may omit their definition; or you may have forgotten what the symbol meant on its first introduction. In either case you will find the glossary of symbols in the back of the book helpful. Furthermore, Appendixes A and B, rules for the manipulation of summations and of expectations, are very important, since we will use these rules to considerable extent in our simple derivations of results.

So far we have talked at length about the author's expectations about the student and the reasons underlying this book, but we have failed to say much about the topic itself. Next we will take an overview of what inferential statistics is about. In addition, some ideas about formal systems and mathematical models will be given, which may help the student understand how "statistics" can mean both a body of applied methods and a mathematical theory.

Applications of statistics occur in virtually all fields of research endeavor—the physical sciences, the biological sciences, the social sciences, engineering, market and consumer research, quality control in industry, and so on, almost without end. Although the actual methods differ somewhat in the different fields, the applications all rest on the same general theory of statistics. By examining what the fields have in common in their applications of statistics we can gain a picture of the basic problem studied in mathematical statistics. *The major applications of statistics in any field all rest on the possibility of repeated observations or experiments made under essentially the same conditions.* That is, either the researcher actually can observe the same process repeated many times, as in industrial quality control, or there is the *conceptual* possibility of repeated observation, as in a scientific experiment that might, in principle, be repeated under identical conditions. However, in any circumstance where repeated observations are made, even though every precaution is taken to make conditions exactly the same the results of observations will vary, or tend to be different, from trial to trial. The researcher has control over some, but not all, of the factors that make outcomes of observations tend to differ from each other.

In some areas of research, objects or phenomena viewed under the same conditions will vary only to a small extent. This is certainly true in some branches of physical science, where observations made under carefully controlled conditions give virtually identical results. On the other hand, in the biological, and especially the social, sciences, even though the experimenter exerts almost superhuman efforts to observe repeatedly under precisely the same conditions, some differences among his observations will occur, and these differences are ordinarily *not* negligible.

When observations are made under the same conditions in one or more respects, but they give outcomes differing in other ways, then there is some *uncertainty* connected with observation of any given object or phenomenon. Even though some things are known to be true about that object in advance of the observation, the experimenter cannot predict with complete certainty what its other characteristics will be. Given enough repeated observations of the same object or kind of object the experimenter may be able to formulate a good bet about what the other characteristics are likely to be, but he cannot be completely sure of the status of any given object.

This fact leads us to the central problem of theoretical statistics: *in one sense, mathematical statistics is a theory about uncertainty, the tendency of outcomes to vary when repeated observations are made under identical conditions.* Granted that certain conditions are fulfilled, theoretical statistics permits deductions about the *likelihood* of the various possible outcomes of observation. The essential concepts in statistics derive from the theory of probability, and the deductions made within the theory of statistics are, by and large, statements about the probability of particular kinds of outcomes, given that initial, mathematical, conditions are met.

Mathematical statistics is a formal mathematical system. Any mathematical system consists of these basic parts:

1. A collection of undefined "**things**" or "**elements,**" considered only as abstract entities;
2. A set of undefined **operations,** or possible relations among the abstract elements;
3. A set of **postulates** and **definitions,** each asserting that some specific relation holds among the various elements, the various operations, or both.

In any mathematical system the application of logic to combinations of the postulates and definitions leads to *new* statements, or theorems, about the undefined elements of the system. *Given* that the original postulates and definitions are true, then the new statements *must* be true. Mathematical systems are purely abstract, and essentially undefined, **deductive** structures. In the first chapter, the theory of sets will be used as an example of an abstract system of this sort, and the theory of probability also has this character, as we shall see.

Mathematical systems are not really "about" anything in particular. They are systems of statements about "things" having the formal properties given by the postulates. The mathematician does not, in fact, have to commit himself about what he really has in mind to call these abstract elements; indeed, he may have absolutely nothing in mind that exists in the real world of experience, and his sole concern may be in what he can derive about the other necessary relations among abstract elements given particular sets of postulates. It is perfectly true, of course, that many mathematical systems originated from attempts to describe real objects or phenomena and their interrelationships: historically, the abstract systems of geometry, school algebra, and the calculus grew out of problems where something very concrete was in the back of the mathematician's mind. However, as *mathematics* these systems deal with completely abstract entities.

When a mathematical system is interpreted in terms of real objects or events, then the system is said to be **a mathematical model** for those objects or events. Somewhat more precisely, the undefined terms in the mathematical system are identified with particular, relevant, **properties** of objects or events; thus, in applications of arithmetic, the number symbols are identified with magnitudes or amounts of some particular property that objects possess, such as weight, or extent, or numerosity. The system of arithmetic need not apply to other characteristics of the same objects, as, for example, their colors. Once this identification can be made between the mathematical system and the relevant properties of objects, then anything that is a logical consequence in the system is a true statement about objects in the model, *provided*, of course, *that the formal characteristics of the system actually parallel the real characteristics of objects in terms of the particular properties considered.* In short, in order to be useful as a mathematical model, a mathematical system must have a formal structure that "fits" at least one aspect of a real situation.

Probability theory and statistics are each both mathematical systems *and* mathematical models. Probability theory deals with elements called "events," which are completely abstract. Furthermore, these abstract things are paired with numbers called "probabilities." The theory itself is the system of logical relations among these essentially undefined things. The experimenter uses this abstract system as a mathematical model: his experiment produces a real outcome, which he calls an event, and he uses the model to find a probability, which he interprets as the relative frequency of occurrence for that outcome. If the requirements of the model are met, this is a true, and perhaps useful result. If his experiment really does not fit the requirements of probability theory as a system, then the statement he makes about his actual result need not be true. (This point must not be overstressed, however. We will find that often a statistical method can yield practically useful results even when its requirements are not fully satisfied. Much of the art in applying statistical methods lies in understanding when and how this is true.)

Mathematical systems such as probability theory and the theory of statistics are, by their very nature, **deductive.** That is, formal assertions are postulated as true, and then by logical argument true conclusions are reached. All well-developed theories have this formal, logico-deductive character.

On the other hand, the problem of the empirical scientist is essentially different from that of the logician or mathematician. Scientists search for general relations among events; these general relations are those which can be expected to hold whenever the appropriate set of circumstances exists. The very name "empirical science" asserts that these laws shall be discovered and verified by the actual observation of what happens in the real world of experience. However, no mortal scientist ever observes all the phenomena about which he would like to make a generalization. He must always draw his conclusions about what would happen for *all* of a certain class of phenomena by observing a very few particular cases of that phenomenon.

The student acquainted with logic will recognize that this is a problem of **induction.** The rules of logical *deduction* are rules for arriving at true consequences from true premises. Scientific theories are, for the most part, systems of deductions from basic principles held to be true. If the basic principles are true, then the deductions must be true. However, how does one go about arriving at and checking the truth of the initial propositions? The answer is, for an empirical science, observation and inductive generalization—going from what is true of some observations to a statement that this is true for *all possible* observations made under the same conditions. Any empirical science begins with observation and generalization.

Furthermore, even after deductive theories exist in a science, experimentation is used to check on the truth of these theories. Observations that contradict deductions made within the theory are prima-facie evidence against the truth of the theory itself. Yet, how does the scientist know that his results are not an accident, the product of some chance variation in procedure or conditions over which he has no control? Would his result be the same in the long run if the experiment could be repeated many times?

It takes only a little imagination to see that this process of going from the specific to the general is a very risky one. Each observation the scientist makes is different in some way from the next. Innumerable influences are at work altering—sometimes minutely, sometimes radically—the similarities and differences the scientist observes among events. Controlled experimentation in any science is an attempt to minimize at least part of the accidental variation or "error" in observation. Precise techniques of measurement are aids to the scientist in sharpening his own rather dull powers of observation and comparison among events. So-called "exact sciences," such as physics and chemistry, have thus been able to remove a substantial amount of the unwanted variation among observations from time to time, place to place, observer to observer, and hence are often able to make general statements about physical phenomena with great assurance from the observation of quite limited numbers of events. Observations in these sciences can often be made in such a way that the generality of conclusions is not a major point at issue.

In the biological and social sciences, however, the situation is radically different. In these sciences the variations between observations are not subject to the precise experimental controls that are possible in the physical sciences. Refined measurement techniques have not reached the stage of development that they have attained in physics and chemistry. Consequently, the drawing of general conclusions is a much more dangerous business in these fields, where the sources of variability among living things are extremely difficult to identify, measure, and control. And yet the aim of the social or biological scientist is precisely the same as that of the physical scientist—arriving at general statements about the phenomena under study.

Faced as he is with only a limited number of observations or with an experiment that he can conduct only once, the scientist can reach general conclusions only in the form of a "bet" about what the true, long run, situation actually is like. Given only sample evidence, the scientist is always unsure of the "goodness" of any assertion he makes about the true state of affairs. The theory of statistics provides ways to assess this uncertainty and to calculate the probability that he will be wrong if he decides in a particular way. *Provided that the experimenter can make some assumptions about what is true, then the deductive theory of statistics tells him how likely he is to observe particular results.* Armed with this information, the experimenter is in better position to decide, if he must, what he will say about the true situation. Regardless of what he decides from his evidence, he *could* be wrong; but using deductive statistical theory he can at least determine the probabilities of error in a particular decision.

In recent years, a branch of mathematics has been developed around this problem of decision-making under uncertain conditions. This is sometimes called "statistical decision theory." One of the main problems treated in decision theory is the choice of a decision rule, or "deciding how to decide" from evidence. Decision theory evaluates rules for deciding from evidence in the light of what the decision-maker wants to accomplish. As we shall see in later chapters, mathematics can tell us wise ways to decide how to decide under some circumstances, but it can never tell the experimenter how he *must* decide in any particular situation. The theory of statistics supplies one very important piece of information to the experimenter: the probability of sample results *given* certain conditions. Decision theory supplies another: optimal ways of using this and other information to accomplish certain ends. Nevertheless, neither theory tells the experimenter *exactly* how to decide—how to make the inductive leap from what he observes to what is true in general. This is the experimenter's problem, and he must seek the answer outside of deductive mathematics, and in the light of what he is trying to do.

These, then, are a few of the reasons for studying inferential statistics. The rest of this book will go into the background and details of how these methods are developed and used. I hope that you enjoy learning about them as much as I have enjoyed trying to explain them for you.

W.L.H.

Ann Arbor
November 1972

Contents

6. CENTRAL TENDENCY AND VARIABILITY 215

7. SAMPLING DISTRIBUTIONS AND POINT ESTIMATION 261