STATISTICAL MODELS IN BEHAVIORAL RESEARCH

WILLIAM K. ESTES

Statistical Models in Behavioral Research

William K. Estes
Harvard University

LEA
1991 LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Hillsdale, New Jersey Hove and London

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Library of Congress Cataloging-in-Publication Data

Estes, William Kaye.
Statistical models in behavioral research / W.K. Estes.
p. cm.
Includes bibliographical references and index.
ISBN 0-8058-0686-5 (cloth). — ISBN 0-8058-0688-1 (pbk.)
1. Psychometrics. 2. Psychology—Research. 3. Psychology-Statistical methods. 4. Psychology—Mathematical models.
I. Title.
BF39.E72 1991
150'.1'5195—dc20 90-43392
CIP

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

For George and Rhey

Preface

This book evolved during my teaching of courses in statistics and quantitative methods to advanced undergraduate and graduate students in psychology and social science off and on for nearly 40 years. Apologies are due the students who happened to fall in earlier classes, before I had learned how to teach this kind of material effectively, and appreciation is due to many collaborators and assistants. In the latter connection, I owe a special debt to Robert R. Rosenthal, with whom I shared a graduate course at Harvard University for 8 years and from whom I have learned a good deal about how to help students get across the gap between theory and application. I also want to mention a succession of outstanding teaching assistants at Harvard, including Beverly Chew, Jean MacMillan, and most notably, Kris Kirby, who caught seemingly innumerable glitches in draft chapters of this volume and brought my attention to many possibilities for improving communicability. Finally, I wish to thank Nancy Rury, who accomplished the all but impossible task of converting my handwritten drafts into neat typescript, Kay Estes, who prepared the index, and my longtime friend and publisher, Lawrence Erlbaum, who personally supervised the final transition from typescript into print.

W. K. E.

Contents

1 Introduction

WHAT IS NEW?

The teacher of a graduate course in statistics for psychologists or other behavioral scientists now has so many excellent texts available (for example, Hays, 1988; Hildebrand, 1986; Howell, 1987; Kirk, 1982; Loftus & Loftus, 1988) that one may well ask, "Why should another be needed?" The atypically small size of this volume may suggest the answer: I set out to supplement, not to duplicate, available textbooks. I assume that the reader has had or is taking a course that covers the elements of probability, sampling distributions, and the computation of analyses of variance and regression on balanced data sets obtained from simple, standard designs. I assume further that the reader shares with this author at least the following needs:

1. Capability of doing statistical analyses by means of statistical programs with some insight into what is going on behind the scenes.
2. Understanding of the basis for the various rules given in textbooks about admissable tests in various common designs.
3. Guidance in the calculation and use of statistics never fully covered in standard texts (e.g., effect size measures, standard errors, and contrasts in various types of designs).

4. Deeper insight into the relationship between analysis of variance and regression and the ways of getting the best of both approaches out of statistical packages.
5. Help in dealing with the hazards of unbalanced data sets.

The state of the art in statistics for psychological research is changing rapidly and with it what the researcher needs to learn. The prototype of a second course for prospective research workers in behavioral science is a concentration on mastering methods of calculating analyses of variance (henceforth ANOVAs) for a variety of common research designs. However, the hand calculation of ANOVAs is rapidly going out of style. Often (unhappily), even before learning a modicum of statistics, the student in psychology enters research data into computer programs and then seeks help in decoding a massive output of summary tables. The reward is a large increase in output for a given amount of time and effort on the input side. The negative aspect is similar to that of putting powerful machines into untrained hands. The ability to obtain statistical analyses soon outruns the ability to interpret them. Thus, the present-day student needs more theory than his predecessors were even allowed to see in order to be able to cope with the outputs of computer programs.

Meeting these needs requires some understanding of the models that underlie statistical methods and how the models can be applied to guide the solution of new problems not covered by textbook examples. However, although the mathematics needed for practical purposes is not very deep, the difficulties of coping with mathematical formalisms and details of derivations have tended to make the needed approach inaccessible to all but a few behavioral scientists. In this connection, I have been struck by the findings of current research in cognitive science that the early development of children's understanding of science and mathematics is greatly facilitated if the qualitative understanding of physical or mathematical models precedes the task of dealing with computational details. Thus, I have wondered whether it may be possible to convey the essentials of statistical models by means of simplified representations that eliminate most of the complexities of notation and focus on qualitative understanding of the models. An individual who learns to think in terms of models in this way will not be able to do much in the way of new derivations for novel problems but perhaps will be equipped to find his/her way intelligently through many of the difficulties of interpreting research data and in particular to reap the advantages of statistical packages with some confidence in how to interpret the outputs.

I have no magic bullet to offer, but over many years of teaching research design and quantitative methods in psychology, I have developed some ways of simplifying the presentations of concepts and derivations so as to make the substance of important statistical results available to the mathematically unskilled research worker. An important boost to this effort has come quite recently in the appearance of some readily available microcomputer programs, most notably SYSTAT, that encourage, even in some instances require, the user to plan the analysis of a statistical design in relation to the mathematical model that underlies the computations. In this book, I follow a path somewhere between the level of the SYSTAT manual (Wilkinson, 1986), in which models appear only in the form of highly simplified and stylized equations that serve as instructions to the program, and that of treatises like Winer (1971), in which models are presented fully but with so much detail of notation and derivation as to be inaccessible to all but a few users.

There is no new statistical theory in this book. The basic theory I have drawn on is well covered in Graybill (1961), Searle (1987), and Winer (1971). My contribution has been mostly to abstract, reorganize, and apply the theoretical results to problems that arise frequently in psychological research and especially to find ways of simplifying the presentation of models and operations with models so as to make them readily available to students and investigators who lack either mathematical background or taste for doing derivations, or both. For the experienced investigator, I include material on model testing and related topics that is not covered in textbooks or other readily available sources.

A COMPARISON OF STATISTICAL AND SCIENTIFIC MODELS

My view of statistical models and their applications as simply a special case of the uses of formal models in scientific theory and research may seem unconventional to some readers. The close relationship of the two types of models can, however, be pointed out in terms of an illustration. A simple theoretical model that has become very familiar in cognitive psychology is the function relating reaction time to set size in short-term memory search. In a paradigm made famous by Sternberg (1966), an experimental subject is presented with a small set of items, typically randomly selected digits, letters, or short words, then is presented with a test item and responds yes or no as quickly as possible, yes indicating that the test item was in the set of items presented (the memory set) and no indicating that it was not. On the assumption that the memory set is represented in the subject's short-term memory system in a list-like

format and that the task of responding to the test item is achieved by comparing it successively to each of the items in the memory representation, reaction time for the yes or no response can be predicted from the function

$$Y = a + bi \tag{1.1}$$

where Y denotes mean reaction time to respond to the test item, a is the time required to generate a response, b is the time required for a single comparison of the test item with an item in memory, and i is the number of items in the set.

This scientific hypothesis seems straightforward, but how are we to decide whether or not it is supported by data? The upper panel in FIG. 1.1 presents the problem more concretely. The open squares in the figure represent data from a hypothetical memory search experiment in the form of mean reaction time at each set size, and the straight line represents the theoretical hypothesis with a particular choice of values of the constants *a* and *b*. It appears to the eye that the data and the hypothesis are in fair agreement, but for many scientific purposes one wishes to be able to say something more specific and preferably quantitative about the goodness of fit. In the course of research in this paradigm, investigators have, for example, wanted to make comparisons of the goodness of fit in different studies that have used different kinds of items, different conditions of item presentation, or the like, or they have wanted to compare the goodness of fit of this hypothesis with others that differ from it in some particular way. To serve these purposes, it is necessary to replace our initial general and somewhat vague statement of the problem of goodness of fit with more structured and specific questions to which definite answers can be obtained.

To start this process, we note that the data points in FIG. 1.1 do not fall on the plotted straight line exactly but vary somewhat irregularly around it. This result is to be expected if there is experimental error in the situation so that different experimental subjects differ somewhat from each other in performance and so that the reaction times obtained from a particular subject on a particular set size may vary somewhat from one occasion to another. Thus, we need to face the question whether we can confidently rule out the possibility that the upward trend of the data points in FIG. 1.1 is simply a chance result of the error in the situation. Two steps that we take in order to deal with this question are illustrated in the middle panel of FIG. 1.1. The first step is to augment the equation for the theoretical hypothesis by a term, e, representing experimental error. We have, in a manner of speaking, imbedded the theoretical model in a statistical model, in particular a linear regression

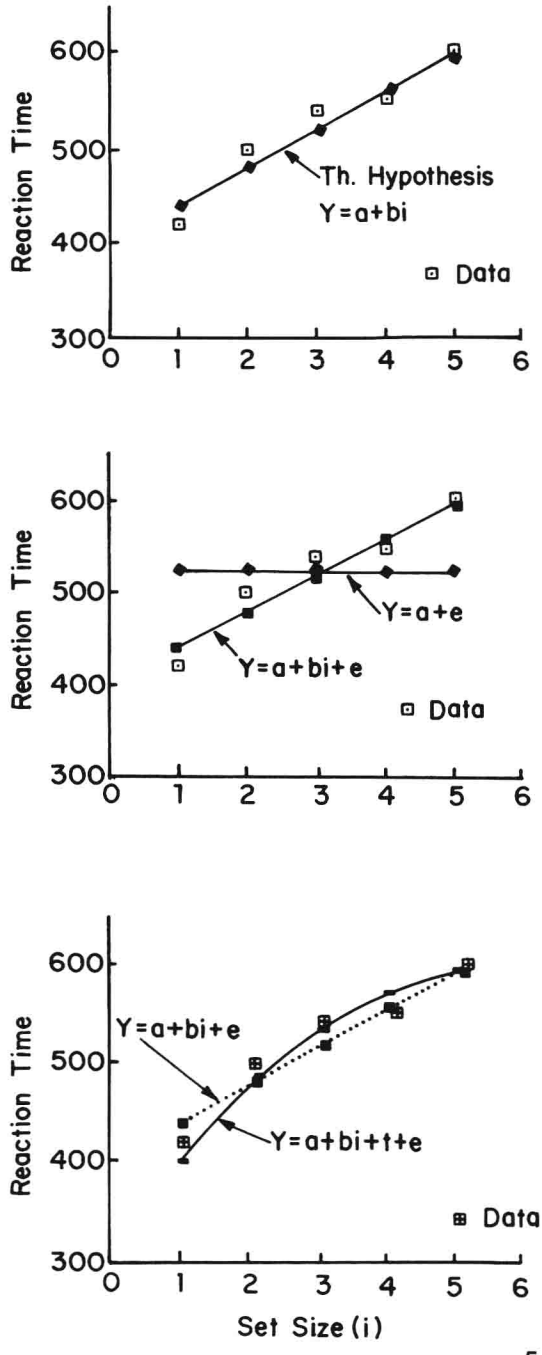

FIG. 1.1 Steps in statistically evaluating a scientific hypothesis. The upper panel presents a set of observed data in the form of mean reaction times versus size of a set of items to be remembered (in an experimental paradigm described in the text) together with a theoretical hypothesis that takes the form of a linear function relating reaction time to set size.

For a first test of the hypothesis, it is compared with the simpler hypothesis that reaction time is constant over set size, represented by the horizontal line in the middle panel. If the result of the first test demonstrates a significant positive slope for the observed trend, thus rejecting the simpler model, we ask next whether the trend is specifically linear. To answer that question, we compare the model representing the linear trend with the augmented model illustrated in the bottom panel, in which a quantity t is added to or subtracted from the linear function at each set size.

model. The equation for the latter expresses the idea that the observed data should represent a combination of the true theoretical function relating reaction time to set size with an error term that may result in observed points falling randomly above or below the true function at each set size.

The second step is to formulate an alternative hypothesis, represented by the horizontal line and accompanying regression equation in the middle panel of FIG. 1.1, expressing the possibility that there is no true relation between reaction time and set size and that the observed trend might be simply a combination of a horizontal straight line plus random error. As will be described in detail in Chapter 7, a computer program for regression analysis, given the full set of data values as input, will form an estimate of the magnitude of experimental error, select values of constants a and b that are the best possible in the sense that they reduce the variability of the data points around the theoretical function to a minimum, and will enable us to make a quantitative statement about the confidence with which we can conclude that the upward sloping function provides a better account of the data than the horizontal function.

If the results of this analysis enable us to conclude that the upward trend in the data points is real (i.e., not due to chance) so that the scientific hypothesis is preferable to the alternative one, we may wish to proceed to the specific question whether the particular assumption of a linear relation between reaction time and set size, assumed in the theoretical hypothesis, provides a better description of the data than some alternative function that would have an upward trend of a different form, such as the curvilinear function shown in the bottom panel of FIG. 1.1. At various times in the history of research in this paradigm, investigators have in fact surmised that the true relation is better assumed to be a logarithmic or power function than a straight line.

In order to be able to make a decision about these alternative functions as a class, we define the statistical model represented by the lowermost equation in FIG. 1.1, which will be seen to be the linear regression function plus an additional term t. The values of t, which in general would be different from one set size to another, represent the differences between the values predicted by the augmented hypothesis and the values predicted by the linear hypothesis. Working in the regression framework, we do not attempt to guess what the values of t should be but rather let the statistical computer program determine the values that are best in the sense of yielding the smallest *error estimate*, that is, the smallest variation of the observed data points around the

function. The final step is to compare this error estimate for the linear and the augmented hypothesis and produce a quantitative statement as to whether the augmented model yields a significantly better account of the data than the linear model.[1]

The specific statistical concepts and tools needed to understand how the regression program accomplishes these purposes will be developed in the next few chapters. At this point, I wish only to emphasize the overall strategy of imbedding the idealized relationship derived from a scientific hypothesis (as the one expressed in Equation 1.1) in a model general enough to allow for the possibility that either this hypothesis or some alternative hypothesis is true. One can then determine in a systematic way whether the hypothesized relation or the alternative is better justified by the data. This strategy of comparing a more general model with a simpler one that is in a sense included within it underlies all of the types of statistical hypothesis testing that will be covered in this volume.

ORGANIZATION OF THE VOLUME

The following four chapters develop the tools needed to understand a variety of ANOVA and regression designs and the models underlying them. These chapters review the elements of probability and decision theory, sampling distributions, contrast analysis, and hypothesis testing. The tools are then applied to a sample of research problems chosen to illustrate the various aspects of design and analysis represented in the following outline, all of the treatments being integrated within a common theoretical framework known in the statistical literature as the *general linear model.*

- Balanced Designs
  - Fixed effects
    - One-way ANOVA
    - Simple regression
    - Multiple classifications
      - ANOVAs
      - Mixed ANOVA and regression
  - Random effects
    - One-way ANOVA
    - Intra-class correlation

[1] As will be seen in Chapter 7, this technique is not limited to the case where a scientific hypothesis prescribes a linear function but can be generalized to any specified function.

Mixed models, including repeated measures
Two-way ANOVAs
Two-way ANOVA/Regression
Higher order and nested designs
Multiple regression and analysis of covariance
Unbalanced Designs

For most purposes, we can take "balanced design" to denote one in which all cells contain equal numbers of scores, although there are a few infrequently occurring exceptional cases that qualify (see Howell, 1987, p. 392). Imbalance raises no special problems for one-way designs, but for two-way and higher-order classifications, the distinction between balanced and unbalanced designs is critical. The familiar algebraic breakdown of a total sum of squares into components associated with main effects, interactions, and error; the property of ANOVA tables that the component sums of squares add up to the total sum of squares; the obvious correspondence of F tests to simply stateable hypotheses about population parameters—all hold only for balanced designs. Therefore, I defer the problem of dealing with unbalanced designs for special treatment at the end of the tour.

My treatment of methods and designs stops short of multivariate analysis of variance (MANOVA). One reason is that in order to concentrate on mathematical reasoning rather than derivations, I limit the mathematics used in the book to simple algebra, eschewing even matrix operations. A second reason is that there are many presentations of MANOVA available (e.g., Dunteman, 1984; Finn, 1974; Harris, 1985) that are fully adequate for courses on multivariate methods and for the needs of investigators who work with predominantly correlational data. With this limitation, it has been possible to give a compact but reasonably complete presentation of the basic statistical models needed by investigators whose research is primarily experimental in character and is often oriented toward the testing of quantitative theories.

2 Statistics, Probability, and Decision

GENERAL CONSIDERATIONS

I think it would be hard to find an informed person contesting the proposition that probability theory is basic to statistics. But why do people believe this proposition to be true? An answer suggested by many textbook presentations and by the unending controversies between adherents of different approaches to probability and statistics (Bayesian, Fisherian, etc.) is that the probabilities computed in the course of statistical applications can be regarded as properties of events that occur in the research situations. When the applications are to gambling situations involving, for example, fair coins, dice, or roulette wheels, it is known that probabilities derived from statistical theory do closely describe actual long-term experience and can be ignored by the gambler only at the cost of certain ruin. When the applications are to scientific research situations, however, whether in physical science, agriculture, or social science, I know of no evidence to suggest that the same is true.

Why, then, do scientists in all fields depend heavily on statistical methods based on probability theory? I suggest that the answer is much the same as for other uses of formal models in science. A statistical model for an experiment is a deliberate idealization of the actual empirical situation, and probabilities derived by means of the model hold strictly only for the idealized, not for the actual, situation. In a few instances, such as gambling and perhaps some applications in physical science, the idealized situation of the model may be so close to the

empirical one that probabilities derived are literally interpretable as properties of events. In biology, psychology, and social science, however, such correspondences are rare, and there must be other reasons for using statistics and taking derived probabilities to be empirically significant. The only defensible answer, I would say, is that just as with any methods in science, the use of statistical methods is justified by long-term experience. Thus, although there is no reason to think that the probability value derived from application of a statistical test to a psychological experiment tells us how often an observed result would occur in actual research if a hypothesized effect were absent, a large body of experience assures us that, other things equal, a result significant at the .01 level will more often prove replicable than a result significant at the .05 level. In general, relative values of derived probabilities are often highly informative, although absolute values may not be.

The reader should not, however, leap to the conclusion that I am one of those who depreciate the use of conventional significance levels. Quite to the contrary, I think the use of conventional significance levels is useful and entirely defensible, not because the absolute probability values are empirically meaningful, but because the use of conventional significance levels can be conducive to clear thinking. To see why this is so, we need to review a few basic concepts of probability and decision.

OUTCOME TREES AND DECISION CRITERIA

When considering application of statistics to a research situation, the investigator needs routinely to attend to two preliminaries. The first is to consider whether there is reason to believe that the research design includes an element of randomization that would make application of a probability model appropriate. The second, if the answer to the first is affirmative, is to lay out an outcome tree for the experiment and to specify decision criteria.

We say that a set of events is random if the long-term relative frequencies of occurrence of the events over replications of the situation settle down to stable values, and if manipulation of these values according to the laws of probability yields empirically confirmable predictions. We know from experience that these criteria are satisfied by tosses of coins or dice and by the outputs of well-constructed computer programs for the generation of random numbers. Thus, we can and do use these devices to lift ourselves by our boot straps, so to speak, in research situations and introduce the randomness required for appropriate application of statistical models by making random assignments of