Richard J. Harris

A Primer of Multivariate Statistics

Second Edition

A Primer of Multivariate Statistics

Second Edition

Richard J. Harris

University of New Mexico

ACADEMIC PRESS, INC.
(Harcourt Brace Jovanovich, Publishers)
Orlando San Diego San Francisco New York
London Toronto Montreal Sydney Tokyo São Paulo

Academic Press, Inc.
Orlando, Florida 32887

United Kingdom Edition published by
Academic Press, Inc. (London) Ltd.
24/28 Oval Road, London NW1 7DX

ISBN: 0-12-327252-6
Library of Congress Catalog Card Number: 84-70636

PRINTED IN THE UNITED STATES OF AMERICA

Preface

Since publication of the first edition of *A Primer of Multivariate Statistics*, I have had the benefit of comments from colleagues and other users, a great deal of analytic and Monte Carlo work by multivariate statisticians, and, above all, hindsight. I have tried to put all this new information to effective use in this second edition.

Many users have suggested that *primer* was not as descriptive of the first edition as I had intended. In this second edition I have added many more examples from a wider variety of areas. I have moved almost all derivations to a separate appendix, so as not to interfere with the flow of the discussion on a first reading. I have also expanded greatly the description of the multivariate approach to repeated-measures designs.

This edition puts even greater emphasis than did the first on interpreting the linear combinations (regression variates, discriminant functions, canonical variates, principal components, and factors) produced by multivariate analysis. This change resulted from the realization that Bonferroni-adjusted univariate tests are fully adequate for controlling the experimentwise error rate if the tester is only interested in examining each predictor or each dependent variable by itself. If, therefore, the tester is not going to interpret the optimal linear combination(s) uncovered by a multivariate technique, using such a technique only costs power, relative to univariate tests with Bonferroni-adjusted critical values.

One of the major reasons for feeling a second edition was necessary was my admiration for the work by Boik (1981) and by Bird and Hadzi-Pavlovic (1983) extending the issues of power and robustness of multivariate criteria from their performance as tests of omnibus null hypotheses to their use as bases for post hoc specific comparisons. This extension strongly reverses what had begun to emerge as consensus positions on how to handle repeated-measures designs (via adjustment of degrees of freedom within an otherwise standard Anova approach) and on which test criterion to use in Manova and in canonical analysis (a multiple-root test such as Pillai's trace). Boik's results strongly support testing each repeated-measures contrast against its own variance, rather than against a pooled error term. Bird and Hadzi-Pavlovic show that multiple-root tests have abysmal power for testing specific contrasts on specific linear combinations of measures in Manova, even in situations where their power as tests of the overall null hypothesis is much higher than that of the greatest-characteristic-root (gcr) test. Moreover, the sometimes appalling lack of robustness of the gcr test is greatly improved if the researcher is willing to consider the overall null hypothesis rejected only if a *substantively interpretable* contrast on an interpretable linear combination of the dependent variables has a univariate *F* that exceeds the gcr-based post hoc critical value. These findings have very important implications that I felt should be brought before a general audience and at as early a stage as possible in the process of learning to use multivariate statistics.

There are a number of other ways in which this second edition has been updated: for example, it provides more explicit treatment of the use of multiple regression (MRA) to handle unequal-*n* Anova designs and of alternatives to least-squares regression; it delineates more clearly the relationship between MRA and two-group discriminant analysis; it provides "progress reports" on the development of finite intersection tests in Manova and of redundancy analysis and repeated-battery tests in canonical analysis; and it offers what I hope will be a more compelling argument for and more consistent application of the use of factor-score coefficients, rather than factor loadings, in "naming" factors.

Throughout I have attempted to couch all of this new material in terms that anyone with a solid understanding of univariate statistics can follow. I have also sought terms that give priority to the implications for research behavior over the more traditional statistical criteria that have too often led to mathematically elegant answers to the wrong questions. It is my fond hope that this second edition has thereby earned the right to be considered a true *primer* of multivariate statistics.

If so, much of the credit must go to the feedback generously provided by users of the first edition. I would like to thank Hebert H. Blumberg of Goldsmiths' College, London, Joe Rodgers of the University of Oklahoma, and Robert M. Pruzek of the State University of New York at Albany for their reviews. I have found especially valuable the opportunity to discuss multivariate issues with Kevin O'Grady (often while "on the run" through the hills of Cedar Crest) while he was a fellow faculty member at UNM and with Kevin Bird during my sabbatical year at the University of New South Wales (Sydney). Indeed, production of this second edition might still be years down the road if I hadn't had the challenge of producing a rough draft for use by Kevin Bird's multivariate class and the opportunity of seeing a master teacher then put that draft to use in his course.

Dick Harris
November 1984

Contents

1 Introduction
The Forest before the Trees

2 Multiple Regression:
Predicting One Variable from Many

3 Hotelling's T^2: *Tests on One or Two Mean Vectors*

4 Multivariate Analysis of Variance: *Differences among Several Groups on Several Measures*

5 Canonical Correlation: *Relationships between Two Sets of Variables*

6 Principal Component Analysis: *Relationships within a Single Set of Variables*

7 Factor Analysis: *The Search for Structure*

8 The Forest Revisited

Digression 1 *Finding Maxima and Minima of Polynomials*

Digression 2
Matrix Algebra

Digression 3

Digression 4
Uses of Canned Computer Programs

Appendix A
Statistical Tables

Appendix B
Computer Program Listings

Appendix C
Derivations

1

The Forest before the Trees

1.0 WHY STATISTICS?

This text and its author subscribe to the importance of sensitivity to data and of the wedding of humanitarian impulse to scientific rigor. Therefore, it seems appropriate to discuss the present author's conception of the role of statistics in the overall research process. This section assumes familiarity with the general principles of research methodology. It also assumes some acquaintance with the use of statistics, especially significance tests, in research. If this latter is a poor assumption, the reader is urged to delay reading this section until after having read Section 1.2.

Statistics is a form of social control over the professional behavior of researchers. The ultimate justification for any statistical procedure lies in the kinds of research behavior it encourages or discourages.

In their descriptive applications, statistical procedures provide a set of tools for efficiently summarizing the researcher's empirical findings in a form that is more readily assimilated by the intended audience than would be a simple listing of the raw data. The availability and apparent utility of these procedures generate pressure on researchers to employ them in reporting their results, rather than relying on a more discursive approach. On the other hand, most statistics summarize only certain aspects of the data; consequently, automatic (e.g., computerized) computation of standard (cookbook?) statistics without the intermediate step of "living with" the data in all of its concrete detail may lead to overlooking important features of these data. A number of authors (see especially Anscombe, 1973, and Tukey, 1977) have offered suggestions for preliminary screening of the data so as to ensure that the summary statistics finally selected are truly relevant to the data at hand.

The inferential applications of statistics provide protection against the universal tendency to confuse aspects of the data that are unique to the particular sample of subjects, stimuli, and conditions involved in a study with the general properties of the population from which these subjects, stimuli, and conditions were sampled. For instance, it often proves difficult to convince a subject who has just been through a binary prediction experiment involving, say, predicting which of two lights will be turned on in each of several trials that the experimenter used a random-number table in selecting the sequence of events. Among researchers, this tendency expresses itself as a proneness to generate complex post hoc explanations of their results that must be constantly revised since they are based in part on aspects of the data that are highly unstable from one replication of the study to the next. Social control is obtained over this tendency, and the "garbage rate" for published studies is reduced, by requiring that experimenters first demonstrate that their results cannot be plausibly explained by the null hypothesis of no true relationship in the population between their independent and dependent variables. Only after this has been established are experimenters permitted to foist upon their colleagues more complex explanations. The scientific community generally accepts this control over their behavior because

(a) bitter experience with reliance on investigators' informal assessment of the generalizability of their results has shown that some formal system of "screening" data is needed.

(b) the particular procedure just (crudely) described, which we may label the *null-hypothesis-testing procedure,* has the backing of a highly developed mathematical model. If certain plausible assumptions are met, this model provides rather good quantitative estimates of the relative frequency with which we will falsely reject (Type I error) or mistakenly fail to reject (Type II error) the null hypothesis. Assuming again that the assumptions have been met, this model also provides clear rules concerning how to adjust both our criteria for rejection and the conditions of our experiment (such as number of subjects) so as to set these two "error rates" at prespecified levels.

(c) the null-hypothesis-testing procedure is usually not a particularly irksome one, thanks to the ready availability of formulas, tables, and computer programs to aid in carrying out the testing procedure for a broad class of research situations.

However, acceptance is not uniform. Bayesian statisticians, for instance, point out that the mathematical model underlying the null-hypothesis-testing procedure fits the behavior and beliefs of researchers quite poorly. No one, for example, seriously entertains the null hypothesis, since almost any treatment or background variable will have *some* systematic (though possibly miniscule) effect. Similarly, no scientist accepts or rejects a conceptual hypothesis on the basis of a single study. Instead, the scientist withholds final judgment until a given phenomenon has been replicated on a variety of studies. Bayesian approaches to statistics thus picture the researcher as beginning each study with some degree of confidence in a particular hypothesis and then revising this confidence in (the subjective probability of) the hypothesis up or down, depending on the outcome of the study. This is almost certainly a more realistic description of research behavior than that provided by the null-hypothesis-testing model. However, the superiority of the Bayesian approach as a descriptive theory of research behavior does not necessarily make it a better prescriptive (normative) theory than the null-hypothesis-testing model. Bayesian approaches are not nearly so well developed as are null-hypothesis-testing procedures, and they demand more from the user in terms of mathematical sophistication. They also demand more in terms of ability to specify the nature of the researcher's subjective beliefs concerning the hypotheses about which the study is designed to provide evidence. Further, this dependence of the result of Bayesian analyses on the investigator's subjective beliefs means that Bayesian "conclusions" may vary among different investigators examining precisely the same data. Consequently, the mathematical and computational effort expended by the researcher in performing a Bayesian analysis may be relatively useless to those of his or her readers who hold different prior subjective beliefs about the phenomenon. (The "may's" in the preceding sentence derive from the fact that many Bayesian procedures are robust across a wide range of prior beliefs.) For these reasons, Bayesian approaches are not employed in the *Primer*. Press (1972) has incorporated Bayesian approaches wherever possible.

An increasingly "popular" objection to null hypothesis testing centers around the contention that these procedures have become *too* readily available, thereby seducing researchers and journal editors into allowing the tail (the inferential aspect of statistics) to wag the dog (the research process considered as a whole). Many statisticians have appealed for one or more of the following reforms in the null-hypothesis-testing procedure:

(a) Heavier emphasis should be placed on the *descriptive* aspects of statistics, including, as a minimum, the careful examination of the individual data points before, after, during, or possibly instead of "cookbook" statistical procedures to them.

(b) The research question should dictate the appropriate statistical analysis, rather than letting the ready availability of a statistical technique generate a search for research paradigms that fit the assumptions of that technique.

(c) Statistical procedures that are less dependent on distributional and sampling assumptions, such as randomization tests (which compute the probability that a completely random reassignment of observations to groups would produce as large an apparent discrepancy from the null hypothesis as would sorting scores on the basis of the treatment or classification actually received by the subject) or jackknifing tests (which are based on the stability of the results under random deletion of portions of the data), should be developed. These procedures have only recently become viable as high-speed computers have become readily available.

(d) Our training of behavioral scientists (and our own practice) should place more emphasis on the hypothesis-*generating* phase of research, including the use of post hoc examination of the data gathered while testing one hypothesis as a stimulus to theory revision or origination.

Kendall (1968), Mosteller and Tukey (1968), Anscombe (1973), and McGuire (1973) can serve to introduce the reader to this growing "protest literature."

The ultimate answer to all of these problems with traditional statistics is probably what Skellum (1969) refers to as the "mathematization of science" as opposed to the (cosmetic?) application of mathematics to science in the form of very broad statistical models. Mathematization of the behavioral sciences involves the development of mathematically stated theories leading to quantitative predictions of behavior and to derivation from the axioms of the theory of a multitude of empirically testable predictions. An excellent discussion of the advantages of this approach to theory construction vis à vis the more typical verbal–intuitive approach is provided by Estes (1957), and illustrations of its fruitfulness are provided by Atkinson, Bower, and Crothers (1965), Cohen (1963), and Rosenberg (1968). The impact on statistics of the adoption of mathematical approaches to theory construction is at least twofold:

(a) Since an adequate mathematical model must account for variability as well as regularities in behavior, the appropriate statistical model is often implied by the axioms of the model itself, rather than being an ad hoc addition to what in a verbal model are usually overtly deterministic predictions.

(b) Because of the necessity of amassing large amounts of data in order to test the quantitative details of the model's predictions, ability to reject the overall null hypothesis is almost never in doubt. Thus, attention turns instead to measures of goodness of fit relative to other models and to less formal criteria such as the range of phenomena handled by the model and its ability to generate counterintuitive (but subsequently confirmed) predictions.

As an example of the way in which testing global null hypotheses becomes an exercise in belaboring the obvious when a math model is used, Harris' (1969) study of the relationship between rating scale responses and pairwise preference probabilities in personality impression formation found, for the most adequate model, a correlation of .9994 between predicted and observed preference frequency for the pooled "psychometric function" (which was a plot of the probability of stating a preference for the higher rated of two stimuli as a function of the absolute value of the difference in their mean ratings). On the other hand, the null hypothesis of a perfect relationship between predicted and observed choice frequencies could be rejected at beyond the 10^{-30} level of significance, thanks primarily to the fact that the pooled psychometric function was based on 3,500 judgments from each of 19 subjects.

Nevertheless, the behavioral sciences are too young and researchers in these sciences are as yet too unschooled in mathematical approaches to hold out much hope of mathematizing all research; nor would such complete conversion to mathematically stated theories be desirable. Applying math models to massive amounts of data can be uneconomical if done prematurely. In the case of some phenomena, a number of at least partly exploratory studies need to be conducted first in order to narrow somewhat the range of plausible theories and point to the most profitable research paradigms in the area. Skellum (1969), who was cited earlier as favoring mathematization of science, also argues for a very broad view of what constitute acceptable models in the early stages of theory construction and testing. (See also Bentler and Bonnett's 1980 discussion of goodness-of-fit criteria and of the general logic of model testing.) Null hypothesis testing can be expected to continue for some decades as a quite serviceable and necessary method of social control for most research efforts in the behavioral sciences. Null hypotheses may well be merely convenient fictions, but no more disgrace need be attached to their fictional status than to the ancient logical technique of reductio ad absurdum, which null hypothesis testing extends to probabilistic, inductive reasoning.

As will become obvious in the remaining sections of this chapter, the *Primer* attempts in part to plug a "loophole" in the current social control exercised over researchers' tendencies to read too much into their data. (It also attempts to add a collection of rather powerful techniques to the descriptive tools available to behavioral researchers. Van de Geer (1971) has in fact written a textbook on multivariate statistics that deliberately omits any mention of their inferential applications.) It is hoped that the above discussion has convinced the reader—including those who, like the author, favor the eventual mathematization of the behavioral sciences—that this will lead to an increment in the quality of research, rather than merely prolonging unnecessarily the professional lives of researchers who, also like the author, find it necessary to carry out exploratory research on verbally stated theories with quantities of data small enough to make the null hypothesis an embarrassingly plausible explanation of our results.

1.1 WHY MULTIVARIATE STATISTICS?

As the name implies, *multivariate statistics* refers to an assortment of descriptive and inferential techniques that have been developed to handle situations in which sets of variables are involved either as predictors or as measures of performance. If researchers were sufficiently narrowminded or theories and measurement techniques so well developed or nature so simple as to dictate a single independent variable and a single outcome measure as appropriate in each study, there would be no need for multivariate statistical techniques. In the classic scientific experiment involving a single outcome measure and a single manipulated variable (all other variables being eliminated as possible causal factors through either explicit experimental control or the statistical control provided by randomization), questions of *patterns* or *optimal combinations* of variables scarcely arise. Similarly, the problems of *multiple comparisons* do not becloud the interpretations of any t test or correlation coefficient used to assess the relation between the independent (or predictor) variable and the dependent (or outcome) variable. However, for very excellent reasons, researchers in all of the sciences—behavioral, biological, or physical—have long since abandoned sole reliance on the classic univariate design. It has become abundantly clear that a given experimental manipulation (for example, positively reinforcing a class of responses on each of N trials) will affect many somewhat different but partially correlated aspects (for example, speed, strength, consistency, and "correctness") of the organism's behavior. Similarly, many different pieces of information about an applicant (for example, high school grades in math, English, and journalism; attitude toward authority; and the socioeconomic status of his or her parents) may be of value in predicting his or her grade point average in college, and it is necessary to consider how to combine all of these pieces of information into a single "best" prediction of college performance. (It is widely known—and will be demonstrated in our discussion of multiple regression—that the predictors having the highest correlations with the criterion variable when considered singly might contribute very little to that combination of the predictor variables that correlates most highly with the criterion.)

As is implicit in the discussion of the preceding paragraph, multivariate statistical techniques accomplish two general kinds of things for us, these two functions corresponding roughly to the distinction between *descriptive* and *inferential* statistics. On the descriptive side, they provide rules for combining the variables in an optimal way. What is meant by "optimal" varies from one technique to the next, as will be made explicit in the next section. On the inferential side, they provide a solution to the *multiple comparison* problem. Almost any situation in which multivariate techniques are applied could be analyzed through a series of univariate significance tests (for example, t tests), using one such univariate test for each possible combination of one of the predictor variables with one of the outcome variables. However, since *each* of the univariate tests is designed to produce a significant result $\alpha \times 100\%$ of the time (where α is the "significance level" of the test) when the null hypothesis is correct, the probability of having at least one of the tests produce a significant result when in fact nothing but chance variation is going on increases rapidly as the number of tests increases. It is thus highly desirable to have a means of explicitly controlling the experimentwise error rate. Multivariate statistical techniques provide this control.[1]

---

1. A possible counterexample is provided by the application of analyses of variance (whether univariate or multivariate) to studies employing factorial designs. In the univariate case, for instance, a k-way design (k experimental manipulation combined factorially) produces a summary table involving 2^{k-1} terms—k main effects and $\binom{k}{2}$ two-way interactions, and so on—each of which typically yields a test having a Type I error rate of .05. The experimentwise error rate is thus $1 - (95)^{2k-1}$. The usual multivariate extension of this analysis suffers from exactly the same degree of compounding of Type I error rate, since a separate test of the null hypothesis of no difference among the groups on any of the outcome measures (dependent variables) is conducted for each of the components of the among-group variation corresponding to a term in the univariate summary table. In the author's view, this simply reflects the fact that the usual univariate analysis is appropriate only if each of the terms in the summary table represents a truly a priori (and, perhaps more important, theoretically relevant) comparison among the groups, so that treating each comparison independently makes sense. Otherwise, the analysis should be treated (statistically at least) as a one-way Anova followed by specific contrasts employing Scheffé's post hoc significance criterion (Winer, 1971, p. 198), which holds the experimentwise error rate (that is, the probability that any one or more of the potentially infinite number of comparisons among the means yields rejection of H_0 when it is in fact true) to at most α. Scheffé's procedure is in fact a multivariate technique designed to take into account a multiplicity of *independent* variables. Further, Anova can be shown to be a special case of multiple