SHACHTER, R. D. (1986). Evaluating influence diagrams. Technical Report, Dept. Engineering-Economic Systems, Stanford Univ.

SHEPARD, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **27** 125–140.

SHEPARD, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* **27** 219–246.

SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H., NAYLOR, J. C. and DRANSFIELD, M. (1984). The implementation of the Bayesian paradigm. *Comm. Statist. A—Theory Methods* **14** 1079–1102.

SPIEGELHALTER, D. J. (1986). A statistical view of uncertainty in expert systems. In *Artificial Intelligence and Statistics* (W. Gale, ed.) 17–55. Addison-Wesley, Reading, Mass.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press.

SUBRAMANIAM, K. and SUBRAMANIAM, K. (1973). *Multivariate Analysis: A Selected and Abstracted Bibliography, 1957–1972.* Dekker, New York.

TORGERSON, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* **17** 401–419.

TUFTE, E. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Conn.

TUKEY, P. A. and TUKEY, J. W. (1981a). Preparation; prechosen sequences of views. In *Interpreting Multivariate Data* (V. Barnett, ed.) 189–213. Wiley, New York.

TUKEY, P. A. and TUKEY, J. W. (1981b). Data-driven view selection; agglomeration and sharpening. In *Interpreting Multivariate Data* (V. Barnett, ed.) 215–243. Wiley, New York.

TUKEY, P. A. and TUKEY, J. W. (1981c). Summarization; smoothing; supplemented views. In *Interpreting Multivariate Data* (V. Barnett, ed.) 245–275. Wiley, New York.

WIJSMAN, R. A. (1984). Two books on multivariate analysis. *Ann. Statist.* **12** 1145–1150.

Comment

T. W. Anderson

1. OBJECTIVES

I am pleased that *Statistical Science* has furnished the opportunity to review my second edition and stimulate a discussion of the development and future of multivariate statistical analysis. The reviewer makes clear that his paper is "a thoroughly biased and narrow look"; I look forward to an unbiased, broad and comprehensive view in the future.

This article contrasts two books on multivariate statistical analysis that are very different in content and objectives. I shall hold my discussion to Schervish's remarks concerning my book. Let me first elucidate my criteria for inclusion of material. Writing a book on multivariate statistical analysis originated as an idea some forty years ago. It was accomplished over a period of years in connection with teaching courses in the Department of Mathematical Statistics at Columbia University. I wanted to write about statistical analysis that I thought has a sound foundation, about methods that were widely accepted. When the first edition was published in 1958, I had no thought that a quarter of a century would pass before the second edition would appear. When I finally came to revise the book, I found that most of the contents had stood the test of time; there was little that I wanted to change or delete, although there was a good deal that could be added. It has been a great satisfaction to me that the book has stood up so well; the initial selection of material has been justified. The objectives and organization of the first edition have been retained. In fact, the headings of the chapters and of most of the sections have been kept.

Although the book includes a considerable amount of mathematics, the primary objective is to provide and explain the methods and their properties. I think that the purpose of statistical theory is to initiate, develop, clarify and evaluate statistical methods. One criterion for inclusion of a topic is that it contributes to understanding useful procedures. Accordingly, there is not much theory in the book for its own sake, but I will admit that the relevance of some material is a matter of personal taste and some theory is to satisfy intellectual curiosity.

A second criterion, as the reviewer has surmised, is that a topic has a mathematical backing. For a confident and thorough understanding, the mathematical theory is necessary. This implies a rigorous treatment.

Thirdly, I wanted to organize the contents coherently. This desire is partly for the sake of clarity and efficiency of exposition and partly for personal satisfaction—aesthetics, if you will.

An outcome of following these criteria was that the inference treated here is based on normal distributions as models. There was not a place for ad hoc methods, valuable though they may be. Normal distributions serve as suitable models for generating many sets of data, but, of course, not for all sets.

Because the book is aimed at statistical practice, I included a number of examples, perhaps not enough. Beside the twelve examples mentioned by

T. W. Anderson is Professor of Statistics and Economics, Stanford University, Stanford, California 94305.

the reviewer, there are some additional data sets in problems. To assist the practitioner some tables of significance points have been added.

Although the book is now 675 pages instead of the 378 pages of the first edition, I, as author, was the first to realize that not everything of value in multivariate analysis could be included in a single volume. (Some people have told me they like that first edition better because they do not have so much to learn.) However, I thought that no purpose was served by listing in the book what had been excluded. A book should be judged on the basis of the author's objectives: Are they worthwhile? Is he successful in achieving them?

Before I discuss some omissions relevant to the review, let me point out that I assume some statistical knowledge and sophistication on the part of my readers. They know, presumably, that the power of a test is important and that the significance level is to be balanced against power. Thus, the appropriate significance level should be adjusted to sample size. A null hypothesis is only an approximation to the question that the investigation really wants to ask; $\mu = 0$ is to be interpreted as μ close to 0. This degree of approximation is related to sample size, the choice of significance level and the resulting effect on power. Achieved significance levels may serve as a measure of departure from the null hypothesis. I can recommend a good text for these ideas (Anderson and Sclove, 1986).

In this book under review I have not gone into detail about the applicability of the methods and how to carry them out; there are important practical matters that I did not have space for. Some suggestions about computation were made, but the development of efficient computational methods requires different expertise and depends to some extent on the size of the data set and the equipment available. Computational programs and packages were not mentioned. (Will they last 26 years?) Although I think graphical techniques are essential, a comprehensive exposition of them did not fit in because a rigorous foundation for them is not available.

2. TESTING HYPOTHESES AND INVARIANCE

Something like one-quarter of my book is devoted to tests of hypotheses and their properties. I think that this emphasis is not out of line with current practice. In terms of statistical methods with a firm theoretical basis it is certainly not. Indeed, I think a greater proportion of research reported in the journals of mathematical statistics concerns such issues. I suppose I am partly responsible because my first edition made a number of problems apparent; a lot of doctoral dissertation advisers owe me thanks. Of our knowledge of statistical procedures with a firm mathematical foundation, tests of hypotheses form a fairly large part and my treatment reflects this share.

The power of a test is an important property; its knowledge is desirable in choosing a significance level; and selection of a particular test among alternative tests may depend on a comparison of their powers. The powers of some tests are easily ascertained and described, for example, the T^2 test, but for most multivariate tests the powers cannot be explained compactly. An important class of multivariate tests is the class of invariant tests of the general linear hypothesis. Comparison of the powers of some competing tests is made in Section 8.6.5 in terms of asymptotic expansions; the leading terms depend on only two functions of the invariants of the parameters. To tabulate the power function of such a test, however, involves (1) the dimensionality, (2) the number of linear restrictions (or columns specified), (3) the number of degrees of freedom in the estimate of the covariance matrix, (4) the significance level, (5) the sum of the roots of a certain population determinantal equation and (6) the sum of squares of those roots. The exact power functions which could be calculated from exact noncentral distributions (in Muirhead, 1982, for example) might involve more parameters.

It should be pointed out further that in principle a family of tests may provide a family of confidence regions. In many cases these are feasible regions; see Sections 5.3 and 8.7, for example.

This review and ensuing discussion is not the place for a full scale debate on hypothesis testing. (See Lehmann, 1986, for example). It surely has a place in statistical inference—where sampling variability is evaluated and its effects controlled. My feeling about statistical theory is that it yields guideposts on the basis of definite statements. A map does not give every topographical detail; yet it can be used to locate unstated positions by their relation to marked positions.

The property of invariance is used in much of the book and is in part a unifying thread. In the simplest case of testing H: $\mu = 0$ in $N(\mu, \Sigma)$ the invariant of the sufficient statistics $\bar{x}$ and S with respect to linear transformations $X^* = CX$ is $T^2 = N\bar{x}'S^{-1}\bar{x}$ (or a function of it) and of the parameters μ and Σ is $N\mu'\Sigma^{-1}\mu$. One desirable property of any test based on the sample invariant is that it has a fixed significance level; that is, the probability of rejecting H when it is true does not depend on Σ. Another desirable effect is that the T^2 test (rejecting H when T^2 is larger than a constant) has power against any alternative $\mu \neq 0$ that is greater than the significance level and the power increases along any ray $\mu = \tau\gamma$ (that is, τ increasing and the vector γ fixed). As a matter of fact, it is further proved in Section 5.6 that the T^2 test is admissible in the class of all tests, not just invariant tests. In this and many other cases an optimal invariant test cannot be improved on by any test—invariant or not.

In the univariate case, invariance with respect to positive scalar C implies use of the t statistic; the requirement is that inference is independent of the scale of measurement, a reasonable condition. Invariance with respect to all scalar C implies the two-sided t test, which has power greater than the significance level against all alternatives. (It is seldom, if ever, that one sees a two-sided t test made with different probabilities assigned to the two tails.) In the multivariate case, invariance with respect to all nonsingular C requires that the scales of measurements of the components do not affect the outcome, but also imposes invariance with respect to oblique transformation of coordinates.

Obviously, there are problems where the parameter invariants do not describe suitably the alternatives the investigator considers relevant. In testing $\mu = 0$, one might be interested in alternatives $\mu \geq 0$, $\mu \neq 0$ that is, each component being nonnegative and at least one component being positive. Perlman (1969) found the likelihood ratio criterion for testing the null hypothesis against this set of alternatives. (Further references are given by Perlman.) He showed that the distribution of the criterion under the null hypothesis depends on the nuisance parameter Σ and obtained upper and lower bounds for the probability of rejection under the null hypothesis. Such a test will give greater power against some alternatives $\mu \geq 0$ than the T^2 test but not against all such alternatives. How suitable such a test is depends on which alternatives in the set $\mu \geq 0$ are considered most important. Much more study needs to be done and is being done.

The reviewer (Section 11) has quoted my example of test of the hypothesis that the means of four populations are equal and questions whether the asserted difference in means is meaningful in the sense of classification. I carried out the test at the 1% significance level, but it would have been better to take a smaller percentage because the sample sizes are large; the minimum sample size was 70. Nevertheless the null hypothesis would be rejected (at even unreasonable significance levels, such as 10^{-4}).

A test is often, if not usually, conducted before using the results for other procedures. In this instance Schervish suggests using the statistics for classification into one of the four populations. I agree that in this case a reasonable purpose for the significance test is to see if classification on such a basis would be effective. He estimates the probabilities of correct classification as given in Table 4. He does not give any measure of sampling variability of these estimates, but it would appear that those probabilities are better than the probability of .25 attained by guessing. Who decides the value of the improvement?

3. BAYESIAN INFERENCE AND PREDICTION

3.1 Bayesian Inference

The reviewer has used this opportunity to express himself on many issues in inference. Because of limited space here, I am not discussing his general views, but am concentrating on what is specific to multivariate analysis and my objectives in writing this book. As noted, I developed the posterior distribution of the mean μ and covariance matrix Σ by using the conjugate prior distributions. However, I find it difficult to see how an investigator determines an inverted Wishart distribution as a prior for Σ, presumably representing his prior knowledge about Σ. To make use of prior knowledge beyond that of determining a suitable model for the data seems to me desirable when it is possible. However, to develop that approach systematically is beyond the scope of my book. Press (1982) has applied the Bayesian approach systematically.

3.2 Prediction

I agree that prediction can be an important aspect of a statistical investigation, not only under the conditions of the collection of the data, but also under altered conditions. I wish that I had developed these aspects more. And I would add control as another objective. How might one modify the parameters (as estimated) to obtain a desired output? However, the Bayesian methodology is not the only approach to these problems. For example, within the classical framework one can obtain "confidence" regions for future observations.

4. MORE SPECIFIC COMMENTS

4.1 The Multivariate Behrens-Fisher Problem: "What's New?"

It seems strange to me that the reviewer finds Section 5.5 "The Two-Sample Problem with Unequal Covariance Matrices" an *interesting addition* in the second edition because the first contained Section 5.6 "The Multivariate Behrens–Fisher Problem." The only differences between these sections in the two editions is the addition or modification of six sentences, the change of title, and the interchanging of section numbers. The purpose of these changes was not to confuse reviewers.

A more important misunderstanding by Schervish concerns the purported "discarding of observations." For ease of exposition here we consider the two-sample problem of testing $H\colon \mu^{(1)} - \mu^{(2)} = 0$. The T^2 statistic in question is $T^2 = N_1 \bar{y}' S^{-1} \bar{y}$, where $\bar{y} = \bar{x}^{(1)} - \bar{x}^{(2)}$, $\bar{x}^{(i)}$ is the mean of the ith sample, $i = 1, 2$, $N_1 \leq N_2$, and S is based on $x_\alpha^{(1)} - \sqrt{N_1/N_2}\, x_\alpha^{(2)}$, $\alpha = 1, \ldots, N_1$.

(I assumed that my reader would know that he/she should randomize the ordering.) If $N_1 < N_2$, not all of the observations are used in estimating the covariance matrix of $\bar{y}$, but all of the observations are used in $\bar{y}$, the natural estimate of $\mu^{(1)} - \mu^{(2)}$. My claim is that the power of the test would not be increased much if somehow all of the data were used in estimating the covariance matrix of $\bar{y}$.

The T^2 statistic multiplied by a constant has the F distribution with p and $N_1 - p$ degrees of freedom and noncentrality parameter

$$\tau^2 = (\mu^{(1)} - \mu^{(2)})'\left(\frac{1}{N_1}\Sigma_1 + \frac{1}{N_2}\Sigma_2\right)^{-1}(\mu^{(1)} - \mu^{(2)});$$

the matrix in this quadratic form is the covariance matrix of $\bar{y} = \bar{x}^{(1)} - \bar{x}^{(2)}$. The number of observations not used in calculating S can affect only the "denominator" number of degrees of freedom of F. (Suppose we were given $N_2 - N_1$ more observations on $X^{(1)}$.) A possible question is the effect on the power of an F test of increasing the "denominator" number of degrees of freedom while keeping the noncentrality parameter fixed. I found it easy to take a few numbers from the table of Tang (1938). In this case, $p = 4$ as in several examples and the other number of degrees of freedom is of the order of sample size in some examples. The power is not increased very much by increasing the number of degrees of freedom in the estimate of the covariance matrix as long as that number is reasonably large to begin with.

Power for $p = 4$ and Significance Level .05

D.F.\τ^2	5	20	45
30	.140	.775	.996
60	.163	.835	.999
∞	.190	.885	1.000

This is not the place for a comprehensive study of the multivariate Behrens–Fisher problem. If the two covariance matrices are approximately equal, an investigator might pool the sample to increase the number of degrees of freedom substantially, but in the resulting T^2 statistic the inverse of the matrix of the quadratic form in $\bar{x}^{(1)} - \bar{x}^{(2)}$ estimates $(1/N_2)\Sigma_1 + (1/N_1)\Sigma_2$ instead of $(1/N_1)\Sigma_1 + (1/N_2)\Sigma_2$. A huge distortion of significance level may be a consequence.

The reviewer seems to be confused by my statement that "another problem amenable to this kind of treatment" is the problem of testing $\mu^{(1)} = \mu^{(2)}$, where $\mu^{(1)}$ and $\mu^{(2)}$ are two subvectors of $\mu = (\mu^{(1)'}, \mu^{(2)'})'$ in $N(\mu, \Sigma)$. The procedure is exactly the one discussed above: Use the T^2 statistic calculated on the basis of $y_\alpha = x_\alpha^{(1)} - x_\alpha^{(2)}$ to test $\mathscr{E}y_\alpha = 0$. Of necessity $N_1 = N_2$. It is the reviewer's remarks that are not amenable.

4.2 Unbiased Estimates

I do not understand why the reviewer makes the point that it is unwise to use the unique unbiased estimator of $\bar{R}^2$ based on R^2 because I make that point perfectly clear in Section 4.4.3. I gave the expression (47) to satisfy the reader's curiosity inasmuch as I had given the unique unbiased estimate of ρ based on r. The authors of this work (Olkin and Pratt, 1958) had already pointed out that the unbiased estimator of $\bar{R}^2$ was unacceptable. This is one of the few theoretical results reported that does not have some practical implication.

4.3 Discriminant Analysis

Chapter 6, "Classification of Observations," has been (in the reviewer's words) "expanded somewhat," actually from 28 pages to 49 pages. Because the term "discriminant analysis" has been used to cover many different methods and approaches, I have separated out the clearly defined topic of classification. Another aspect of discriminant analysis of more than two samples is to provide a minimum number of classification functions. In the context of means of normal distributions with a common covariance matrix this involves the question of the dimensionality of the space spanned by the means. This question can be subsumed under canonical correlation analysis of regression matrices. Although I have done a considerable amount of research in this area, I included little in the book because I felt that a thorough and rigorous treatment of the necessary asymptotic theory was beyond its scope.

It is true that I stated that "to determine the number of nonzero and zero population canonical correlations one can test" a sequence of hypotheses, but I warned that the tests in the sequence "are not statistically independent, even asymptotically." The implication was that such a procedure does not control the probabilities of errors. (This is an example of material without a full mathematical foundation, and, hence, is debatable.) In the case of nested hypotheses with a suitable structure of complete sufficient statistics, a sequence of such significance tests are independent and may be optimal. (See Anderson, 1962.) However, how to determine the number of nonzero canonical correlations and the number of factors (in factor analysis) has not been settled definitively.

4.4 Principal Components

If it is desired to predict a vector X by means of a linear combination of X itself, the mean squared error of the residuals is minimized by taking as the linear combination the principal component with the largest variance. The reviewer states this result when the covariance matrix is the correlation matrix, but it holds without that condition (proved in Schervish, 1986). This interesting interpretation was what I had in mind when I stated the result as part (b) of Problem 3 of Chapter 11 (Problem 4 in the first edition). Another interpretation is suggested by part (a) of the problem. Suppose $X = S + E$, where the systematic part S and the error E are uncorrelated and $\mathscr{E}E = 0$, $\mathscr{E}EE' = \sigma^2 I$; the last assumption is reasonable if the noise is due to the error in a measuring device that is applied independently to the components. The first principal component is proportional to the linear combination of X that maximizes the variance of the linear combination of the systematic part relative to the variance of the linear combination of error. The last is a multiple of the square of the norm of the vector of coefficients; in this setting it is natural to set the norm to 1.

4.5 Factor Analysis

It is amusing that Schervish writes "there are *traditionally* two modes in which one can perform factor analysis." The so-called *exploratory* mode dates back to Spearman (1904). With regard to the *confirmatory* mode, Anderson and Rubin (1956) developed estimates of the structure when coefficients are specified in advance and stated some asymptotic theory of inference. However, not much attention was paid to this mode until Jöreskog (1969) gave it the name confirmatory, and it was not put into practice until Jöreskog wrote a program for it later.

Factor analysis and latent structure analysis (Section 12 of the review) were put into a general framework based on conditional or local independence in Anderson (1959). A more highly developed model has been studied by Bartholomew (1984).

4.6 Path Analysis and LISREL

The model (2) in the review is a simultaneous equations model as treated in my Section 12.7, but the model for the observed data involves measurement error added to the unobservable variables in (2). The limited information *maximum likelihood* estimator for a single equation in a simultaneous equation system presented in Section 12.7.4 is hardly ad hoc; it was exposited as minimizing a variance ratio (as in the F statistic) to avoid considerable algebra.

5. CONCLUSION

The usefulness of multivariate statistical analysis is growing rapidly—due to greater data collection, increased computer power and expanding knowledge of statistical techniques. At the present time great progress is being made in all aspects: statistical theory and methodology, data analysis, graphical capabilities and computing facilities. There is good reason to expect this progress to continue and expand in other directions.

ACKNOWLEDGMENT

This research was supported by United States Army Research Office Contract DAAG29-K-0239 at Stanford University.

ADDITIONAL REFERENCES

ANDERSON, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In *Probability and Statistics: The Harald Cramér Volume* (Ulf Grenander, ed.) 9–38. Almqvist and Wiksell, Stockholm.

ANDERSON, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **33** 255–265.

ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **5** 111–150. Univ. California Press.

ANDERSON, T. W. and SCLOVE, S. L. (1986). *The Statistical Analysis of Data*, 2nd ed. Scientific Press, Redwood City, Calif.

BARTHOLOMEW, D. J. (1984). The foundations of factor analysis. *Biometrika* **71** 221–232.

JÖRESKOG, K. C. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34** 183–202.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

OLKIN, I. and PRATT, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Ann. Math. Statist.* **29** 201–211.

PERLMAN, M. D. (1969). One-sided testing problems in multivariate analysis. *Ann. Math. Statist.* **40** 549–567.

SPEARMAN, C. (1904). "General-intelligence," objectively determined and measured. *Amer. J. Psychol.* **15** 201–293.

TANG, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statist. Res. Memoirs* **2** 126–157.