METHODS & DESIGNS

Parametric and bootstrap approaches to repeated measures designs

JEFFREY LEE RASMUSSEN

Indiana University-Purdue University at Indianapolis, Indianapolis, Indiana

Recently, a nonparametric technique called bootstrapping has been recommended over the more well-known analysis of variance (ANOVA) for analyzing repeated measures data. Advocates cite as bootstrap's advantages over ANOVA the fact that the former uses distributional information and is free of normal theory assumptions. The present study used a computer simulation to compare the two techniques calculated using data sampled from normal and nonnormal distributions. The parametric test had adequate control of Type I error rates; the nonparametric test had overly liberal Type I error rates and therefore is not recommended.

Many experiments in the behavioral sciences employ a repeated measures design in which each subject gives multiple scores on the dependent variable. Frequently the independent variable of interest is some trials factor, and the experimental question is whether subjects demonstrate change on the dependent variable across the trials. The classic technique for analyzing data from this type of experiment is the parametric repeated measures analysis of variance (ANOVA). Recently, a nonparametric approach called bootstrapping has been advocated for use with repeated measures designs. This paper briefly discusses the additional assumption necessary for repeated measures ANOVA and two approaches for dealing with violation of the assumption. It also discusses the method and purported advantages of the bootstrap technique and point out a number of unanswered questions about it. Finally, a Monte Carlo simulation is employed to answer these questions and evaluate the adequacy of bootstrapping relative to ANOVA.

In addition to assumptions of normality and homoscedasticity, repeated measures ANOVA has an assumption concerning the form of the variance-covariance matrix. The variance-covariance matrix contains the variance of the levels of the within-subject variable on the major diagonal, and the covariance of the levels off the major diagonal. When the variances are equal and the covariances are equal, the matrix is said to have compound symmetry, which is a sufficient condition for a repeated measures F test to be distributed as an F distribution under the null hypothesis. A less restrictive condition than compound symmetry, referred to as circularity, also results in the F test being distributed as an F distribution (Huynh & Feldt, 1970; Kirk, 1982; Rouanet & Lépine,

1970). There are two approaches to dealing with violation of the variance-covariance assumption. The first is to correct for assumption violation using a degrees of freedom adjustment. The second approach is to circumvent the problem by using only two repeated measures for the main analysis or for subsequent tests.

The first approach is often referred to as an e -adjusted F test. Box (1954a, 1954b) demonstrated that if the trials and residual degrees of freedom are corrected by a factor e , then the F test would be distributed approximately as an F distribution under the null hypothesis. The value e is calculated from elements of the population variance-covariance matrix. When the variance-covariance homogeneity assumption is met, then e equals 1, and no correction of the degrees of freedom results; when it is not met, e is less than 1 and the degrees of freedom are reduced accordingly. A number of estimates of e , as well as strategies for its use, are discussed by Keselman, Rogan, Mendoza, and Breen (1980) and by Kirk (1982).

The second approach is to use only repeated measures designs in which the variance-covariance homogeneity assumption is guaranteed to be met (Boik, 1975, 1981; Rogan, Keselman, & Mendoza, 1979). In the case of a two-trial repeated measures design, the degrees of freedom for trials equal 1, and the assumption of equal variance-covariance is trivially satisfied. Along the same lines, employing a subsequent test, such as an orthogonal decomposition of sum of squares or polynomial trend analysis with the degrees of freedom for each contrast equal to 1, also satisfies the assumption. Very often the subsequent analyses are of more interest than is the main analysis; thus, this is a very useful approach to analysis of repeated measures data.

Lunneborg and Tousignant (1985) recommended the bootstrap approach as an alternative to repeated measures ANOVA. The bootstrap is a general data-analytic technique created by Efron (Diaconis & Efron, 1983; Efron,

---

The author's mailing address is: Department of Psychology, Purdue School of Science, Indiana University-Purdue University at Indianapolis, Box 647, Indianapolis, IN 46223.

1979a, 1979b, 1980; Efron & Gong, 1983) in which the sampling distribution of a statistic is generated by repeated sampling from a data set. Although the theory behind bootstrapping is straightforward, the large number of computations necessitate using a computer to carry out the bootstrap.

The bootstrap has been applied to a wide range of data analysis procedures. Bootstrapping a repeated measures design is carried out as follows. First, a repeated measures ANOVA is carried out on the original data using ordinary least squares multiple regression (see Edwards, 1979). The dependent variable scores are predicted and the errors in prediction are obtained (i.e., error = actual – predicted). A large number, say 1,000, of bootstrap replications are carried out. For each replication, a bootstrap sample is generated by adding randomly selected error values to the predicted values. Least squares regression estimators for the independent variable are calculated for each bootstrap sample. Confidence intervals are then established on the 1,000 rank ordered estimators. For example, the 2.5 and 97.5 percentiles—in the present case, the 25th and the 976th rank-order values—would represent the 95% confidence interval. The null hypothesis can be tested by noting whether the confidence interval brackets a value of 0. If it does not, the null hypothesis is rejected.

Lunneborg and Tousignant (1985) cited two advantages of the bootstrap over parametric analysis. First, the bootstrap is free of the normal theory and equal variance-covariance matrix assumptions of the repeated measures ANOVA. Significance levels and confidence intervals are generated directly from the bootstrapped sampling distribution; no critical values derived from normal theory are necessary. The bootstrap thus avoids the controversy concerning the robustness of parametric tests to assumption violations (see, e.g., Bradley, 1978; Games, 1983; Levine & Dunlap, 1982, 1983). It also circumvents the lack of consensus about appropriate correction of violation of the variance-covariance assumptions (Lunneborg & Tousignant, 1985).

Second, the bootstrap approach retains information about the form of the original sample. As such, it does not throw away distributional information, as do most nonparametric techniques, which convert data to ranks (see, e.g., Conover & Iman, 1981). According to Lunneborg and Tousignant (1985), this should make the bootstrap approach particularly appropriate for nonnormal data distributions; the shape of the distribution is used in the calculation of the statistic.

They applied the bootstrap to a set of data obtained from 50 male military personnel confined to quarters following incidence involving alcohol use. During confinement they were not allowed access to alcohol. Five repeated measures were taken on a battery of cognitive tests. The main experimental hypothesis concerned whether days since drink contributed linearly or quadratically to task performance. Also of interest was whether the log of session numbers affected task performance. The hypotheses were tested, analogously to the testing procedure for the

correlation coefficient, by noting whether the 95% and 99% confidence intervals for the regression parameter associated with each effect was greater than zero. The analyses indicated that all effects contributed to task performance.

Although Lunneborg and Tousignant's (1985) paper presents a novel analysis of data, there are a number of questions that, until answered, cast serious doubt upon the usefulness of bootstrapping. First, Lunneborg and Tousignant pointed out that parametric repeated measures ANOVA usually must be corrected for variance-covariance inequality. However, the bootstrap that they developed is for the one case in which repeated measures ANOVA would *not* need to be corrected: the case of 1 degree of freedom for the effect. That is, their bootstrap is a nonparametric counterpart to subsequent tests on 1 degree of freedom, where the variance-covariance assumption is trivially satisfied (Boik, 1975, 1981; Rogan et al., 1979). In short, their bootstrap has no advantage over ANOVA concerning this assumption.

Second, although Lunneborg and Tousignant (1985) claimed that bootstrapping is superior to ANOVA, they gave no evidence to support this claim. That is, they did not compare the two techniques. Most researchers are familiar with ANOVA techniques, but not with bootstrapping. In order to justify bootstrapping as a useful alternative to the more well-known ANOVA, it is necessary to demonstrate that it is superior in terms of flexibility, simplicity of calculation, sensitivity, power, and so forth. A recent paper by Rasmussen (1987) indicated that the parametric approach was superior to the bootstrap approach for estimating Type I error rates and confidence intervals of the correlation coefficient.

Finally, Lunneborg and Tousignant (1985) claimed that bootstrapping may be superior to parametric tests on nonnormal data, yet there is no evidence that this is so. The authors did not assess whether the cognitive battery yields nonnormally distributed data, or whether bootstrapping is more appropriate for data sets from nonnormal populations.

The present study employs two computer simulations to compare the Type I error rate of the nonparametric bootstrap with the parametric ANOVA on repeated measures data. In the first simulation, data are sampled from a normal distribution, and in the other, from a nonnormal distribution. The results will indicate the appropriateness of using bootstrapping as an alternative to parametric repeated measures ANOVA when the assumptions of normality are met and when they are not met.

METHOD

Two FORTRAN 77 programs were written to compare normal theory and bootstrap approaches to repeated measures data with two trials. The programs calculated Type I error rates at the .01 and .05 significance levels for number of subjects ($n = 10, 15, 20, 30$). The first program compared the approaches on an underlying normal popu-

lation, and the second on an underlying nonnormal population.

The first program generated n pairs of data from a normal population using Ahren and Dieter's normal deviate pseudorandom number algorithm (in Lehman, 1977). The data for the two trials for the n subjects was generated as follows. Three independent pseudorandom normal deviates, x , y , and z , were generated. The score for the first trial was made equal to $x + y$, and the score for the second trial was made equal to $x + z$. This method assures that the data for the trials are correlated, as is typical of repeated measures designs.

The FORTRAN program generated 1,000 sets of data for each of the four sample sizes. For the parametric test, the Type I error rate was estimated by noting the proportion of the 1,000 F ratios that exceeded the critical values for the appropriate degrees of freedom and significance levels. For the bootstrap approach, the Type I error rate for the .05 significance level was computed by noting the proportion of the 1,000 data sets in which the 95% confidence interval on the regression parameter did not include zero. The Type I error rate for the .01 significance level was computed analogously by noting the proportion of the 99% confidence intervals that did not include zero.

The second program was identical to the first, except that it generated data from a skewed lognormal distribution. The lognormal deviates were obtained by taking the exponent of the $x + y$ and the $x + z$ deviates for the two trials. Descriptions of the lognormal distribution are given by Aitchison and Brown (1957, p. 8) and by Hastings and Peacock (1975, pp. 84-89).

RESULTS AND DISCUSSION

Table 1 compares ANOVA with bootstrap when the assumption of normality is met. The Type I error rate for the .01 and the .05 significance levels is shown for the four sample sizes. The actual Type I error rate for the bootstrap approach is liberal in comparison with its nominal rate. For example, with the $n = 10$ case, the actual Type I error rates are .036 and .095 for the nominal .01 and .05 significance levels, respectively. That is, using

Table 1
Type I Error Rates for ANOVA and Bootstrap

Number of Subjects	Type of Test	
	ANOVA	Bootstrap
.01 Significance Level		
10	.012	.036
15	.011	.026
20	.012	.022
30	.007	.018
.05 Significance Level		
10	.048	.095
15	.046	.081
20	.054	.075
30	.046	.068

Note—Statistics calculated on data drawn from normal population.

Table 2
Type I Error Rates for ANOVA and Bootstrap

Number of Subjects	Type of Test	
	ANOVA	Bootstrap
.01 Significance Level		
10	.003	.053
15	.002	.040
20	.001	.040
30	.002	.026
.05 Significance Level		
10	.018	.120
15	.021	.107
20	.031	.100
30	.025	.097

Note—Statistics calculated on data drawn from nonnormal population.

the bootstrap approach, one would incorrectly reject the null hypothesis approximately four times too often using the .01 significance level and two times too often using the .05 significance level.

Table 2 compares the two tests calculated on samples drawn from a nonnormal distribution. Population nonnormality causes the parametric test to become conservative; it has the opposite effect on the bootstrapping technique.

Clearly, the bootstrapping technique cannot be recommended as an alternative to parametric ANOVA for analysis of repeated measures data. ANOVA is a well-known data analysis technique; bootstrapping is not. The computer time needed to calculate an ANOVA is a small fraction of the time necessary to calculate the bootstrap analysis—an important consideration for researchers with limited funds for computer time. Finally, and most importantly, bootstrapping shows inferior control of Type I error rate. The fact that nonnormality results in a conservative Type I error rate for parametric tests is well known and usually is not seen as a matter for concern (Glass, Peckham, & Sanders, 1972). For the bootstrap approach, however, there is a serious inflation of the Type I error rate both on normal and nonnormal data distributions.

REFERENCES

- AITCHISON, J., & BROWN, J. A. C. (1957). *The lognormal distribution*. London: Cambridge University Press.
- BOIK, R. J. (1975). *Interactions in the analysis of variance: A procedure for interpretation and a Monte Carlo comparison of univariate and multivariate methods for repeated measures designs*. Unpublished doctoral dissertation, Baylor University, Waco, TX.
- BOIK, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, *46*, 241-255.
- BOX, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290-302.
- BOX, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, *25*, 484-498.
- BRADLEY, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology*, *31*, 144-152.

- CONOVER, W. J., & IMAN, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, **35**, 124-129.
- DIACONIS, P., & EFRON, B. (1983, May). Computer-intensive methods in statistics. *Scientific American*, pp. 116-130.
- EDWARDS, A. L. (1979). *Multiple regression and the analysis of variance and covariance*. San Francisco: W. H. Freeman.
- EFRON, B. (1979a). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.
- EFRON, B. (1979b). Computers and the theory of statistics: Thinking the unthinkable. *Society for Industrial & Applied Mathematics Review*, **21**, 460-480.
- EFRON, B. (1980). Censored data and the bootstrap. *Journal of the American Statistical Association*, **76**, 312-319.
- EFRON, B., & GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **37**, 36-48.
- GAMES, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin*, **93**, 382-387.
- GLASS, G. V., PECKHAM, P. D., & SANDERS, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, **42**, 237-288.
- HASTINGS, N. A. J., & PEACOCK, J. B. (1975). *Statistical distributions*. New York: Wiley.
- HUYNH, H., & FELDT, L. S. (1970). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, **1**, 69-82.
- KESELMAN, H. J., ROGAN, J. C., MENDOZA, J. L., & BREEN, L. (1980). Testing the validity conditions of repeated measures. *Psychological Bulletin*, **87**, 479-481.
- KIRK, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks-Cole.
- LEHMAN, R. S. (1977). *Computer simulation and modeling: An introduction*. Hillsdale, NJ: Erlbaum.
- LEVINE, D. W., & DUNLAP, W. P. (1982). Power of the *F* test with skewed data: Should one transform or not? *Psychological Bulletin*, **92**, 272-280.
- LEVINE, D. W., & DUNLAP, W. P. (1983). Data transformation, power and skew: A rejoinder to Games. *Psychological Bulletin*, **93**, 596-599.
- LUNNEBORG, C. E., & TOUSIGNANT, J. P. (1985). Efron's bootstrap with application to the repeated measures design. *Multivariate Behavioral Research*, **20**, 161-178.
- RASMUSSEN, J. L. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin*, **101**, 136-139.
- ROGAN, J. C., KESELMAN, H. J., & MENDOZA, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical & Statistical Psychology*, **32**, 269-286.
- ROUANET, H., & LÉPINE, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical & Statistical Psychology*, **23**, 147-163.

(Manuscript received March 12, 1987;
revision accepted for publication April 6, 1987.)