DOCUMENT RESUME

ED 308 240 TM 013 633

AUTHOR	Belli, Gabriella M.
TITLE	Parameters in MANOVA Robustness Studies: Comparability to Empirical Research.
PUB DATE	Mar 89
NOTE	19p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE	Information Analyses (070) -- Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Analysis of Covariance; *Analysis of Variance; Comparative Analysis; Literature Reviews; *Multivariate Analysis; Research Methodology; *Research Reports; Statistical Analysis
IDENTIFIERS	*Empirical Research; Journal Articles; Parameter Identification; *Robustness

ABSTRACT

An examination of articles in the last 5 years of the "American Educational Research Journal" (AERJ) was conducted to determine the extent to which choices of levels of parameters in a multivariate analysis were comparable to empirical research that uses multivariate analysis of variance (MANOVA). Originally, a main question was to determine the likelihood of homogeneity of variance and covariance in these studies. A secondary question related to the extent to which empirical researchers using MANOVA techniques reported checks on the tenability of homogeneity before reporting their multivariate results. Of the 212 articles considered, 18 used MANOVA. Six articles that used one-way fixed effects designs included 25 MANOVAs and 5 multivariate analyses of covariance (MANCOVA). Two-way designs were used in 8 articles, with 12 MANOVAs and 9 MANCOVAs. Five articles used higher-order designs in four MANOVAs and three MANCOVAs. In general, for most realistic situations, similar parameters were found in the robustness literature. When the exact design configuration was not available, one could extrapolate between the parameters that were used. Only two articles addressed the issue of robustness. Results indicate that it is imperative that researchers test for assumptions before using multivariate analyses. Three tables give the design features of studies using MANOVA and MANCOVA. The AERJ years, volumes, and pages of the 18 articles assessed are appended. (SLD)

PARAMETERS IN MANOVA ROBUSTNESS STUDIES: COMPARABILITY TO EMPIRICAL RESEARCH

Gabriella M. Belli

Virginia Polytechnic Institute and State University

Paper presented at the annual meetings of the American Educational Research Association, San Francisco, CA, March 27-31, 1989.

Session 13.20 - Properties of Multivariate Statistical Methods

The effects on multivariate test statistics when the underlying assumptions of normality and homoscedasticity are violated have been considered in diverse situations with a variety of parameters being manipulated in Monte Carlo studies. On one extreme are studies that considered only Hotelling's T^2 criterion (the multivariate analog to the univariate t-test) in comparisons of mean vector differences between two groups (e.g., Holloway and Dunn, 1967). At the other extreme, a study by Olson (1974), who considered six different multivariate test criteria using several combinations of equal size groups.

Researchers in all cases have focused on a one-way fixed effects classification for the independent factor. General consensus is that non-normality does not have serious effects on either the significance level or the power in most cases (Ito, 1980). In contrast, even mild levels of heterogeneity could be a serious problem.

Partly because of the differences in parameters used in the robustness studies, conclusions about the effects of heterogeneity tend to be fairly general in scope. No single study can readily vary all relevant parameters at all possible levels. These general conclusions are that, for two equal samples, the significance level of Hotelling's T^2 is not seriously affected by heterogeneity, but that this is not necessarily true for unequal n's (Hopkins and Clay, 1963). Actual level of significance increases when any of the following occur: (1) the number of dependent variables increases, (2) total sample size with equal groups decreases, (3) the ratio of sample sizes in unequal groups departs from one, or (4) heterogeneity increases (Holloway and Dunn, 1967; Ito and Schull, 1964).

For more than two groups of large equal samples, robustness may be achieved only with moderate departures from homogeneity, but even this produces large effects on significance levels when samples are unequal. For several small or moderately large groups ($N < 50$), even maintaining equal samples do not protect against departure from nominal significance levels, with test criteria tending to be liberal (Olson, 1974).

Recognizing the complexity of a multivariate robustness study with respect to all the possible levels of parameters, researchers typically select relatively few levels of each one. If there is any mention of how the selection was made, it is to stipulate that these are levels typically found in practice. To what extent are these choices comparable to empirical research that uses MANOVA? In an attempt to begin answering this question, an examination of articles in the last five years of the American Educational Research Journal (AERJ) was conducted.

Originally, a main question of concern was to determine the likelihood of homogeneity of variance and covariance in these studies. None of the articles considered, however, reported variance-covariance matrices and some did not even report variances or standard deviations for their group means. A few articles that did report a table of intercorrelations among the dependent measures did so over the entire sample and not for each cell in the design. Although actual variance-covariance matrices could not be compared, ratios of largest to smallest standard deviations for a given variable across groups were calculated, when possible, to provide at least some information about likely heterogeneity. Comparability to other design features in empirical research

articles and the parameters used in MANOVA robustness studies, however, was possible.

Findings from the robustness studies, although general, are fairly consistent about the potential problems with Type I error rates and power under heterogeneity conditions unless samples are equal and fairly large. Also, much of the literature reporting these findings is not of recent vintage. Hence, a secondary question relates to the extent to which empirical researchers using MANOVA techniques utilize this literature and report checks on the tenability of homogeneity before reporting their multivariate results.

Parameters in MANOVA Robustness Studies

Along with type and amount of heterogeneity, parameters that can be varied in MANOVA robustness studies include the number of dependent measures (p), the number of groups (k), total sample size (N) with both equal and unequal n's, and the nominal Type I error rate. Some studies used only a few levels of these parameters while others spanned a relatively large range.

For example, Hopkins and Clay (1963) considered two samples of sizes 5, 10, or 20 (where $N = n_1 + n_2$ ranged from 10 to 40 and sample ratios were 1:1 or 2:1) with only two dependent variables having the same variance within groups and varying between groups in ratios of 1.6 and 3.2. Holloway and Dunn (1967) used two equal samples of sizes 5, 10, 25, 50, or 100 and combinations of unequal samples where $n_1 + n_2 = 50$ and one group was 1.5 or 2.3 times larger than the other, p = 1, 2, 3, 5, 7, or 10, and where either all or

half the variances in one group were 1.5, 3, 10, or 100 times larger than in the other group. Hakstian, Roed, and Lind (1979) considered two equal and unequal groups (where total N = 12 or 40 and sample size ratios were 1:1, 2:1 or 5:1), p = 2, 6, or 10, and either all or half the variances in one group being 1.44 or 2.25 times larger than those in the other group.

Similarly, differences can be seen across three selected studies with more than two groups. Korin (1972) used k = 2, 3, or 6, equal n's = 5, 7, or 10, and p = 2 or 4. Olson (1974) considered a large span of levels for his parameters, with both k and p = 2, 3, 6, or 10, and equal n's = 5, 10, or 50. Cervourst (1980) used k = 2, 3, or 6, p = 2 or 3, and varied sample sizes so that degrees of freedom error were 18, 60, or 180. The first two of these studies used variance-covariance matrices in canonical form with various combinations of all variances or only one variance in a group being larger than in other groups to varying degrees, by factors of 1.5 or 10 in the former case and factors of 4, 9, or 36 in the latter. The third study varied both variances and covariances in various patterns, using several combinations of three variances (1, 4, or 9) and three correlations (.2, .5, or .8).

Design Features in Empirical Studies Using MANOVA

Of the 212 articles considered, 18 employed multivariate analysis of variance (MANOVA) (see Appendix for volumes and page references). This was the sole or primary analysis in some cases and, in others, only a minor analysis. Where MANOVA was the main analysis, some articles included only

one design and associated analysis; others consisted of several MANOVA analyses on either different designs or different subsets of dependent variables.

Excluded from consideration for this study were articles where the MANOVA included one or more within-group factors for a multivariate analysis of repeated measures. Cases where a covariate was included in the analyses, however, were included.

Six articles that used one-way fixed effects designs included 25 MANOVAs and five MANCOVAs. Two-way designs were used in eight articles, with 12 MANOVAs and nine MANCOVAs. In addition, five articles used higher-order designs in four MANOVAs and three MANCOVAs.

For comparisons to parameters used in MANOVA robustness studies, the following were recorded for one-way designs: the number of dependent measures (p), the number of groups (k), total number of subjects (N), and group sample sizes (n_j). In addition, obtained p-values and, whenever possible, the ratios of largest to smallest sample sizes (r(n)) and standard deviations (r(s)) were determined. The number of groups was replaced by the number of levels of each factor for factorial designs.

One-Way Designs. Table 1 gives a summary of relevant parameters from the eight published articles that used one-way fixed-effects designs. Where several analyses were performed using the same design and the same number of dependent variables, this is listed only once. Multiple listings for an article are given if the analyses involved different designs or sample sizes.

In only two analyses (article #11 and second analysis in #18), were groups of equal size, and only one set of analyses (article #4) used extremely large total N's.

TABLE 1: Design Features of Studies Using One-way MANOVA

article #	p	k	# of analyses	N	n_j	r(n)	p-value	r(s)
(9)	2	2	3 MANCOVAs	81	22,59	2.7	.0005	
				87	28,59	2.1	.009	-
				109	50,59	1.2	.0002	-
(2)	3	2	3 MANOVAs	<70	-	-	.173	1.1-1.3
					-	-	.005	
					-	-	.001	
(18)	3	2	2 MANCOVAs	53	20,33	1.7	.005	1.1-1.5
				65	32,33	1	.001	1.1-1.3
(14)	4	4	2 MANOVAs	117	8,10,34,65	8.1	ns	1.1-1.5
							.001	1.2-1.7
(4)	4	2	3 MANOVAs Hotelling	1777	-	-	.01	
					-	-	.03	-
					-	-	.28	-
		4	1 MANOVA	1777	-	-	.01	-
		5	1 MANOVA	1777	-	-	.01	-
		5	5 MANOVAs	206-496	-	-	all .01	-
		4	7 MANOVAs	206-520	-	-	5 - .01	-
							1 - .28	-
							1 - .56	-
		3	2 MANOVAs	168	-	-	.13	-
				718	-	-	.01	-
(11)	6	3	1 MANOVA Wilks	120	40,40,40	1	.001	1.1-1.8 (for 5) 3.5 (for 1)

article # - refer to Appendix for corresponding volume and page number
p = number of dependent measures
N = total sample size
r(n) = max(n)/min(n)
k = number of groups
n_j = group sample sizes
r(s) = max(sd)/min(sd)

Article (2) - Unclear if k=2 for all three analyses, or if k=2, 4, and 5, respectively. Exact N also unclear for each analysis; from dfe appears to have been 47, 51, and 69.

Article (14) - No s.d. were given, but stated that variance of one of the four dependent measures was larger than that of the other items.

Article (4) - First five one-way MANOVAs addressed one research question; rest were follow-up analyses to significant interactions in three two-way MANOVAs for second research question.

Although, in the latter case, it was impossible to tell how diverse the 2, 3, 4, or 5 groups in each analysis were with respect to either sample size or variance. The remaining 10 analyses in four articles all had small to moderate samples and different sample sizes, article #14 being the worst case.

The two-group design is the most prevalent in robustness studies and was also the most common one in this group of empirical studies (11 of 30 analyses, or 37%). The second most common empirical design consisted of four groups (33%). Robustness studies bracket this design by using k = 3 and 6 as the second most frequently seen designs. The largest number of groups compared in this set of empirical studies was five, whereas the robustness studies have investigated heterogeneity in cases with as many as 10 groups.

With respect to number of variables, 70% of the 30 analyses used p = 4 (however, these came from only two studies). In contrast, p = 2 or 3 is the most prevalent in robustness studies.

Two-Way Designs. Table 2 gives a summary of the design features from empirical two-way designs. Although these are not investigated in the robustness literature, the impact of heterogeneity would certainly affect the tests of main effects and, most likely, would also affect the interactions.

One of the factors had two levels in all but one study, with the second level being two (in four studies), three (in three studies), or four (in three studies). The only other designs, both in one article, were 4x5 and 5x5. All values of p from two to six were used, with three dependent measures being the most prevalent in this set of two-way designs.

TABLE 2: Design Features of Studies Using Two-way MANOVA

article #	p	axb	# of analyses	N	n_j	r(n)	p-value	r(s)
(3)	2	2x4	1 MANCOVA	72	9	1	a: .00001 b: .048 ab: ns	a: 1.2-1.9 b: 1.2-1.3 ab: 2.5-4.1
(7)	3	2x2	1 MANOVA Wilks	(?)	@ =	-	a: .008 b: .001 ab: .233	1.1-2.9(?)
		2x3	1 MANOVA				ns	
(1)	3	2x4	6 MANOVAs Wilks	7535 to 9118	841-1259	1.0-1.3	12 <.0001 4 <.01 2 ns	1.0-1.2
(12)	4	2x2	1 MANCOVA	68	-	-	a: .001 b: .005 ab: ns	-
(15)	4	2x3	1 MANCOVA Wilks	39	-	-	a: .02 b: .001 ab: ns	1.2-1.5
(4)	4	4x5	2 MANOVAs	1750	-	-	-	-
		5x5	1 MANOVA Hotelling	1750	-	-	-	-
(5)	5	2x2	1 MANOVA Wilks	84	21(?)	1(?)	a: .001 b: ns	1.0-1.4
(16)	6	2x2	4 MANCOVAs	240	49-65	1.1-1.5	a: .01 b: ns ab: .03 3 - all ns	-
		2x3	1 MANCOVA	234	16-61	1.4-3.8	a: .01 b: ns ab: .05	
		2x4	1 MANCOVA	-	-	-	ns	

article # - refer to Appendix for corresponding volume and page number
p = number of dependent measures
N = total sample size
r(n) = max(n)/min(n) across cells
r(s) = max(sd)/min(sd) across cells, except for (3).
axb = number of levels in factors A and B
n_j = group sample sizes
p-value given for A, B, & AB effects

Article (7) - N - difficult to tell from discussion. Numbers stated for the 2x2 analysis implied N=188, with an almost balance design. Based on dfe, N=46 for first analysis and 32 for second.
- The s.d. for one dependent variable was 17 times larger than a corresponding one in another cell, but this is likely an error because both mean and s.d. for that variable was given as 69.4.

Article (5) - Given procedures described, n=42 in each level of one factor. Likely that this was a balanced design, but this was not explicitly stated.

In two studies (#1 and #4) total N was extremely large, as were individual sample sizes. Only one study used equal sample sizes (#3). Another study that indicated a nearly balanced design (#7) did not provide enough clarity to determine exact sample sizes, or even total N for the two MANOVAs performed.

Higher Factorial Designs. Four studies used a three-way design and one a four-way (see Table 3). The most prevalent number of levels was two or three, with number of cells being 12, 16, or 18. Unlike the simpler designs, these studies tended to include larger number of dependent variables (5, 6, or 8) and only one used two measures.

Relatively speaking, a greater proportion of these studies maintained equal n's (2 of 5) than those with simpler designs. In both cases (#10 and #17), sample sizes for main effect comparisons were reasonable (48 or 96, respectively) with respect to overcoming potential moderate heterogeneity problems. Equal cell sizes, however, were extremely small (6 and 12, respectively) in these examples. It is interesting that both situations produced two significant main effects and one significant interaction. Given the extremely small p-values for these tests, it would certainly appear that there was enough power to produce significant results.

One concern would be the range in variability across cells for article #10. All standard deviations were less than .25 and half were less than .1. Although small, a dependent measure in one cell had a standard deviation 16 times larger than in another cell. Without a test of the viability of homogeneity, it is impossible to speculate whether results are reasonable.

TABLE 3: Design Features of Studies Using Higher Level Factorial MANOVA

article #	p	design	# of analyses	N	n_j	r(n)	p-value	r(s)
(13)	2	2x3x3	1 MANCOVA	450	-	-	a: .001 b: .01 ab: ns	-
(6)	5	2x2x3	1 MANCOVA	77	-	-	ab: .05 rest: ns	
(17)	5	2x2x4	1 MANOVA Wilks	192	12	1	a: .001 c: .0001 ac: .002 4 - ns	
(8)	6	2x2x3	3 MANOVAs Wilks	539	-	-	2 < .01 3 < .001 16 - ns	
(10)	8	2x2x2x2	1 MANCOVA	96	6	1	a: .0003 b: .0002 ac: .04	1.8-16.0

article # - refer to Appendix for corresponding volume and page number
p = number of dependent measures
N = total sample size
r(n) = max(n)/min(n) across cells
r(s) = max(sd)/min(sd) across cells, except for (3).
design = number of levels in each factor
n_j = group sample sizes
p-value given for A, B, & AB effects

Two of the remaining three studies (#8 and #13) had substantial Ns and, unless group sizes were seriously discrepant, probably had reasonable ns as well. The last study (#6) used 77 subjects spread over 12 cells. Even if cells were close to equal, this would imply no more that six or seven subjects per cell. Only one of the seven multivariate tests produced significance in that analysis.

Test Statistics. For two-group comparisons, Hotelling T^2 is the preferred statistic. For more than two groups, no statistic is uniformly most powerful. Four that are most commonly used are Roy's largest root (R), Hotelling-Lawley trace (H), Wilks likelihood ratio (W), and Pillai-Bartlett trace (V).

The R statistic tends to depart more from nominal alpha than other tests (Korin, 1972). For large samples, T, W, and V are asymptotically equivalent. Olson (1974) suggested as a rule of thumb that they may be so considered whenever df error are at least 10p times larger than df hypothesis. This would hold true in at least 10 of the articles reviewed. In the remaining cases, the three statistics should not behave in similar fashion.

For small to moderately large samples, the T, W, and V tests tend to be robust against relatively mild heterogeneity, but in general, T and W do not fare as well. If heterogeneity is concentrated (i.e., occurring in only one canonical variate of the more variable group), T, V, and W perform well with respect to Type I error rates. If heterogeneity is dispersed among several dimensions in the more variable group, the V is generally closer to nominal alpha.

When population means differ, the W and T are more powerful in detecting differences if they are heavily concentrated in one canonical

dimension, particularly if p = 5 or 6. In contrast, V is more powerful if population means differ in several canonical dimensions (Olson, 1974; Schatzoff, 1966).

Regarding choice of an appropriate multivariate statistic, one cannot base a decision on the nature of unknown population means. But, regardless of the noncentrality structure, the choice of V seems most appropriate because it typically has reasonable power and either fares as well as W or T, or does better, with respect to Type I error rates (Ito, 1980).

Only eight of the 18 articles cited a specific multivariate statistic. One used Hotelling and seven used Wilks. Given the somewhat better performance of V for protection against heterogeneity, it is interesting to note that W is the primary one used. It is probably more curious that the majority of the studies using multivariate statistics do not even state which test is being applied. For those that do so, none mention which F-approximation is being used.

Comparison of Parameters

Much of the MANOVA robustness work concentrates on equal sample cases. In contrast, in most of the empirical research groups vary in size. However, the robustness work done with unequal samples can apply to most practical situations. The largest sample size ratios typically examined in robustness studies is 5:1, with most being less discrepant in size. In all but one case (#14), the largest group was less than four times bigger than the smallest group in this set of AERJ articles. Frequently, the largest group was

less than three, and often less than two times larger. In only the one case, where four groups varied dramatically, was the largest discrepancy between groups a factor of eight.

Although some robustness studies model a few extremely large variance differences across groups, the majority of the cases involve variance ratios of less than or equal to four. Of the 10 articles that provided standard deviations for cell means, variances for variables in one group were rarely greater than four times variances in other groups. In one analysis (#11), one variance was as much as 12 times larger than a corresponding one. Another case discussed previously (#10), the highly discrepant variance ratios are a function of dividing very small decimal numbers.

Problems with robustness are emphasized when nominal significance levels are small. Although several levels of alpha are often used in robustness studies, .05 is the most common practical significance level. Only one article (#4) used a .01 level, and another (#18) used both .05 and .10 for two different analyses. The remaining analyses, whether explicit or implicit, were performed at the .05 level.

In general, it can be concluded that for most realistic situations, one can find similar parameters in the robustness literature. If the exact design configuration is not available, one can generally extrapolate between the parameters that are used.

Testing for Assumption Violation

Nine of the 18 articles reviewed used covariates in the multivariate analyses. Five of these articles mentioned having tested for homogeneity of regression coefficients or linearity, as appropriate when covariates are used. Two of these five articles additionally stated having tested for homogeneity of variance. Generally, these were one-sentence statements concluding that the assumption(s) was tenable, with no indication of what analyses were performed. None of the articles using MANOVA mentioned having tested for any of the multivariate assumptions.

Given the fairly clear results about the potential problem of heterogeneity in multivariate analyses, it is remarkable that researchers who use these techniques are not testing for the likelihood of having met the assumptions. Or, if they are, they are not so reporting. One article (#14) even used a reference to justify the use of MANOVA with unequal and disproportionate samples, which in this case were dramatically unequal (8, 10, 34, and 65). In addition, variances for different variables across the four groups varied from being close to equal to almost three times larger.

Perhaps the well known fact that the univariate F-test is fairly robust to assumption violations carries over to use of multivariate statistics, and so the assumptions are ignored. Or perhaps we are not doing a good enough job in our statistics and research courses to inform future researchers of the problems associated with assumption violation in MANOVA and to provide them with the tools to test these assumptions. That over half the articles using MANCOVAs

tested for the additional assumptions needed when a covariate is involved may be due to the fact that this is also necessary in univariate ANCOVA, a much more familiar analysis encountered in basic statistics courses.

Conclusion

Whether any of the research articles reviewed for this study are in fact in violation of a multivariate assumption is an empirical question. One that could be addressed by submitting their data to appropriate tests. Assuming heterogeneity is present, some of the articles have design features comparable to those found to be robust. Others do not.

Of greater concern, however, is the fact that only two of 18 articles addressed this issue, even in passing. A minor concern is the fact that all too often pertinent information about basic design features (e.g., sample sizes and standard deviations) were missing. Thus not allowing the reader to assess if the situation is likely to be problematic and thereby use caution in interpreting the results.

We can accept the fact that robustness results are certainly not concrete enough with respect to defining the exact conditions that would create problems or those necessary to ensure adequate robustness and power in all cases. However, that literature does provide adequate guidelines about the pitfalls of using small and discrepant samples, particularly when either the number of variables or the number of groups is moderately large. The parameters most commonly used in these studies are fairly comparable to those in empirical

research. It is imperative that researchers test for assumptions before employing multivariate analyses.

APPENDIX

Year, Volume, and Page Numbers of AERJ Articles Using MANOVA

article #	year	volume	pages
(1)	1987	24(3)	437-461
(2)		24(4)	541-555
(3)	1986	23(1)	87-94
(4)		23(1)	117-128
(5)		23(2)	303-313
(6)		23(3)	348-356
(7)		23(3)	507-516
(8)	1985	22(1)	135-148
(9)		22(3)	369-388
(10)		22(4)	527-547
(11)		22(4)	549-560
(12)	1984	21(1)	145-162
(13)		21(1)	227-236
(14)		21(2)	245-259
(15)		21(2)	449-460
(16)		21(3)	565-578
(17)		21(3)	691-701
(18)	1983	20(4)	563-580

References

Ceurvorst, R.W. (1980) Robustness of MANOVA under heterogeneity of variance and correlation. (Doctoral dissertation, Arizona State University). Dissertation Abstracts International, 41/02B.

Hakstian, A., Roed, J.C., and Lind, J.C. (1979) Two-sample T^2 procedure and the assumption of homogeneous covariance matrices. Psychological Bulletin, 86, 1255-1263.

Holloway, L.N. and Dunn, O.J. (1967) The robustness of Hotelling's T^2. Journal of the American Statistical Association, 62, 124-136.

Hopkins, J.W. and Clay, P.P.F (1963) Some empirical distributions of bivariate T^2 and homoscedasticity criterion M under unequal variance and leptokurtosis. Journal of the American Statistical Association, 58, 1048-1053.

Ito, P.K. (1980) Robustness of ANOVA and MANOVA procedures. In P.R. Krishnaiah (ed.) Handbook of Statistics, Vol I. North-Holland Publishing Company, 199-236.

Ito, K. and Schull, W.J. (1964) On the robustness of the T_0^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. Biometrika, 51, 71-82.

Korin, B.P. (1972) Some comments on the homoscedasticity criterion M and the multivariate analysis of variance tests T^2, W and R. Biometrika, 59, 215-216.

Olson, C.L. (1974) Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 64, 894-908.

Schatzoff, M. (1966) Sensitivity comparisons among tests of the general linear hypothesis. Journal of the American Statistical Association, 18, 22-41.