DOCUMENT RESUME

ED 269 480 TM 860 340

AUTHOR Hummel, Thomas J.; Johnston, Charles B.

TITLE An Empirical Comparison of Size and Power of Seven Methods for Analyzing Multivariate Data in the Two-Sample Case.

PUB DATE Apr 86

NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Analysis of Variance; *Comparative Analysis; Correlation; Differences; Effect Size; Error of Measurement; Hypothesis Testing; Mathematical Models; Measurement Techniques; *Monte Carlo Methods; *Multivariate Analysis; Research Methodology; *Sample Size; Statistical Studies

IDENTIFIERS Bonferroni Procedure; *Empirical Analysis; Error Analysis (Statistics); *Power (Statistics); Type 1 Errors; Univariate Analysis

ABSTRACT

This study investigated seven methods for analyzing multivariate group differences. Bonferroni t statistics, multivariate analysis of variance (MANOVA) followed by analysis of variance (ANOVA), and five other methods were studied using Monte Carlo methods. Methods were compared with respect to (1) experimentwise error rate; (2) power; (3) number of Type 1 errors in experiments with at least one error; and (4) for experiments with at least one false univariate hypothesis, the probability of rejecting at least one of the true hypotheses. One method emerged as having the best all around performance. This method used repeated T-squared statistics and removed the variable with maximum significant F statistic, providing a good balance between power and Type 1 errors. It consisted of the following steps: (1) MANOVA on p variables followed by ANOVAs; (2) reject the hypothesis for the variable with the largest significant F statistic and remove that variable; (3) MANOVA on p-1 variables; (4) repeat Step 2 with p-1 variables; (5) MANOVA on p-2 variables...and so on until no MANOVAs are significant, no ANOVAs are significant, or there are no variables left. (Author/PN)

An Empirical Comparison of Size and Power of Seven Methods for Analyzing Multivariate Data in the Two-Sample Case

by

Thomas J. Hummel
University of Minnesota

and

Charles B. Johnston
Control Data Corporation

A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1986.

ABSTRACT

This study investigated seven methods for analyzing multivariate group differences. Bonferroni t statistics, MANOVA followed by ANOVAs, and five other methods were studied using Monte Carlo methods. Methods were compared with respect to 1) experimentwise error rate, 2) power, 3) number of Type I errors in experiments with at least one error, and 4) for experiments with at least one false univariate hypothesis, the probability of rejecting at least one of the true hypotheses. One method emerged as having the best all around performance and consisted of the following steps: 1) MANOVA on p variables followed by ANOVAs, 2) reject the hypothesis for the variable with the largest significant F statistic and remove that variable, 3) MANOVA on p-1 variables, 3) repeat Step 2 with p-1 variables, 4) MANOVA on p-2 variables ... and so on until no MANOVAs are significant, no ANOVAs are significant, or there are no variables left.

An Empirical Comparison of Size and Power of Seven Methods for Analyzing Multivariate Data in the Two-Sample Case[1]

MANOVA and Bonferroni procedures can be used to control familywise error rates (hereafter experimentwise error rate) when analyzing multivariate group differences, and with the availability of computer software to perform the complex computations, these procedures have become easy to apply. However, in a review of approaches to analyzing multivariate data, Bray and Maxwell (1982) indicated that there are a number of areas of controversy on how to use these methods. One area of concern is how to perform further analysis on the dependent variables once a significant overall effect has been found. A variety of methods have been presented for analyzing and interpreting data after obtaining a significant overall MANOVA statistic.

Hummel and Sligo's Monte Carlo study (1971) compared univariate and multivariate analysis of variance procedures for analyzing multivariate data. They compared three

---

1 This paper is a summary of the second author's doctoral dissertation carried out under the supervision of the first author in the Department of Educational Psychology, University of Minnesota, Minneapolis.

procedures using univariate and multivariate procedures used both singly and together. The first procedure tested each univariate null hypothesis, H_0: $u_{j1} = u_{j2}$ for j = 1 to p (where p is the number of dependent variables), using a univariate analysis of variance (ANOVA). The second procedure, suggested by Cramer and Bock (1966), used an overall multivariate test of the hypothesis H_0: $\mathbf{u}_1 = \mathbf{u}_2$. If this hypothesis was rejected, ANOVAs were run separately on each of the variables. The third procedure, proposed by Morrison (1967), began with an overall multivariate test of H_0: $\mathbf{u}_1 = \mathbf{u}_2$. Following rejection of this hypothesis, a simultaneous F test derived from the simultaneous confidence interval procedure of Roy and Bose (1953) was employed to test the univariate null hypothesis for each dependent variable.

In another Monte Carlo study, Ramsey (1982) compared five procedures. The first procedure was the one suggested by Cramer and Bock (1966) mentioned earlier. In the second procedure, if the overall hypothesis was rejected, then multiple T^2 tests were performed on all subsets of the p variables that include variable j. If T^2 was significant for each of the subsets at the specified alpha level, then the hypothesis H_0: $u_{j1} = u_{j2}$ was rejected. The third procedure was the simultaneous F test procedure from Hummel and Sligo (1971). The fourth and fifth procedures were modifications of the Bonferroni procedure proposed by Bird (1975). In the fourth procedure, univariate F tests were performed for each variable at $alpha_p = 1 - (1 - alpha)^{1/p}$. The fifth procedure began by performing univariate F tests at $alpha_p$ and rejected the univariate hypothesis H_0: $u_{j1} = u_{j2}$ for each variable j where the F test was significant. If m (>0) variables were found to have signi 'cant differences, then the remaining p-m variables were tested at $alpha_{p-m}$. This continued until no significance was found for all variables tested at $alpha_{p-m}$ or until the final variable was tested. Each of these five procedures was tested for several different numbers of dependent variables, various effect sizes, a variety of correlation values, and several sample sizes. The number of variables with true differences was also varied.

The methods studied by Hummel and Sligo (1971) and Ramsey (1982) need further investigation. Neither study examined the methods with respect to the rate of incorrectly rejecting at least one of the true hypotheses when some hypotheses are false. Ramsey studied the methods only with small sample sizes (n = 5, 7, 9, 15, and 17). Also, one of the methods proposed by Ramsey is impractical in its application, requiring T^2 tests to be performed on all subsets of p variables that include variable j. To reject a single univariate hypothesis, 2^{p-1} T^2 tests would need to be performed. For nine variables, to reject a single univariate null hypothesis would require 256 T^2 tests.

The present study examined procedures not only with respect to power, but also looked at the probability of incorrect rejections when true differences do not exist on some of the variables. The Bonferroni method, four methods investigated by both Ramsey (1982) and Hummel and Sligo (1971), and two new methods were compared.

Monte Carlo methods similar to those used by Hummel and Sligo (1971) and Ramsey (1982) were employed to study seven methods for analyzing multivariate data in the two sample case. The methods were compared across a variety of sample sizes (n = 10, 30, and 50), numbers of dependent variables (p = 3, 6, and 9), proportions of variance in common among the variables (rho^2 = 0.1, 0.3, 0.5, and 0.7, where the off-diagonal elements equal *rho* and the diagonal elements equal 1.0), and effect sizes ($theta_1$ = 0.0, 0.2, 0.5, and 0.8, where $theta_1$ is the noncentrality parameter for variable 1). The goal was to find a method that had an acceptable experimentwise error rate and, when one of the univariate hypotheses was false, had adequate power and an acceptable rate of incorrect rejections for the remaining true hypotheses. This method should also be simple in its application given current computing hardware and software.

The seven methods compared in this study were the following:

- Univariate analyses of variance--Univariate F tests are used to test separately the hypothesis for each of the p dependent variables.

- Multivariate analysis of variance followed by simultaneous F tests--The T^2 statistic is used to test the overall hypothesis. If the statistic is significant then simultaneous F tests are performed separately on each of the p dependent variables.This simultaneous F test is equivalent to testing singly each of the p dependent variables using Roy and Bose's (1953) simultaneous confidence interval. The method of performing a MANOVA followed by simultaneous confidence intervals was suggested by Morrison (1967).

- Combination of univariate and multivariate analyses of variance--The T^2 statistic is used to test the overall hypothesis. If this hypothesis is rejected, then univariate F tests are conducted on each of the dependent variables.

- Bonferroni--Univariate F tests are used to test the hypothesis for each of the dependent variables at *alpha*/p.

- Multiple Bonferroni--Univariate tests are used to first test the hypothesis for each of the p dependent variables at $alpha_p = 1 - (1 - alpha)^{1/p}$. If hypotheses are rejected for m (>0) variables, then the tests are carried out for the remaining p-m variables at $alpha_{p-m}$. This is repeated until there are no rejections or until the final variable is rejected.

- Method 6--The T^2 statistic is used to test the overall hypothesis. If this statistic is significant, then the hypothesis for the variable with the maximum F statistic is rejected and the variable is removed. The T^2 statistic is computed for the remaining p-1 variables. If it is significant, then the hypothesis for the next highest F statistic is rejected and the variable is removed. This is repeated until the T^2 for the remaining variables is no longer significant or until no variables remain.

- Method 7--The same process is followed as Method 6, conducting repeated T^2 tests, except that for a univariate hypothesis to be rejected the highest remaining F statistic must also be significant.

The results of our study show that Method 7, which used repeated T^2 statistics and removed the variable with maximum significant F statistic, provides a good balance between power and Type I errors. Other methods provided either better power or better protection from Type I errors, but overall Method 7 achieved good results for power while still maintaining acceptable control over the Type I error rates. The application of Method 7 should be relatively simple with currently available statistical software (e.g., Statistical Package for the Social Sciences).

For the other six methods, various performance characterisics make them either unacceptable or less acceptable than Method 7.

As Hummel and Sligo (1971) and Ramsey (1982) have previously found, carrying out univariate ANOVAs without the protection of a prior MANOVA should be discouraged because of the lack of control over experimentwise error rate. Carrying out an overall MANOVA and following it up with univariate ANOVAs provides adequate protection against inflated experimentwise error rate, but when one of the univariate hypotheses is false, it does not perform well. As can be anticipated from Miller's (1966) concern regarding Fisher's LSD method, when a single dependent variable is responsible for the rejection of a multivariate hypothesis, the F tests on the remaining p-1 variables are not protected. This leads to an inflated probability of rejecting the hypothesis for at least one of the p-1 variables for which the hypothesis is true. Table 1 shows that only unprotected univariate tests have poorer performance in this respect, and that the problem worsens as $theta_1$ increases and rho^2 decreases.

As was originally anticipated by Hummel and Sligo (1971), the conservative nature of the simultaneous F test leads to its having low power. Tables 2 and 3 show that, overall,

Table 1
Average Rate of Incorrectly Rejecting at Least One True Null Hypothesis for $theta_1 > 0$ and
Across All Values of n and p

	rho^2			
Method	0.1	0.3	0.5	0.7
Univariate	0.204	0.170	0.146	0.112
Bonferroni	0.038	0.033	0.031	0.023
Multiple Bonferroni	0.042	0.036	0.034	0.026
Multivariate-Univariate	0.100	0.093	0.090	0.078
Multivariate	0.003	0.004	0.003	0.002
Method 6	0.063	0.080	0.100	0.125
Method 7	0.053	0.055	0.055	0.047

Table 2
Average Power Across All Values of *rho*^{2}, n, and p

	theta$_1$		
Method	0.2	0.5	0.8
Univariate	0.119	0.454	0.745
Bonferroni	0.034	0.246	0.575
Multiple Bonferroni	0.036	0.249	0.577
Multivariate-Univariate	0.047	0.339	0.657
Multivariate	0.007	0.082	0.313
Method 6	0.070	0.434	0.712
Method 7	0.043	0.336	0.655

Table 3
Average Power for $theta_1>0$ and Across All Values of n and p

Method	rho^2 0.1	0.3	0.5	0.7
Univariate	0.441	0.436	0.440	0.439
Bonferroni	0.288	0.283	0.284	0.285
Multiple Bonferroni	0.290	0.285	0.286	0.288
Multivariate-Univariate	0.293	0.332	0.386	0.398
Multivariate	0.135	0.132	0.134	0.134
Method 6	0.294	0.361	0.437	0.530
Method 7	0.286	0.329	0.366	0.397

its power is less than half that of other procedures. In Table 7, where $theta_1$ = 0.5 and *rho* = 0.7, n = 30, and p = 9, the simultaneous F test procedure has a power of about 0.007, while by comparison Method 7 has a power of 0.450. Morrison's recommendation to use Roy and Bose simultaneous confidence intervals for one-variable-at-a-time comparisons should not be followed, especially as p increases and $theta_1$ decreases.

Method 6 performed well with respect to power and experimentwise error rate. Table 1 shows, however, that the probability is elevated for rejecting at least one true hypothesis when one hypothesis is false. While these values are higher than one might wish, the cross tabulation in this Table 1 masks extremes which can occur within this method. Figure 1 gives an example for which the probability is almost 0.25 that at least one true hypothesis will be rejected. Because of its potential for high error rates of this kind, Method 6 is not recommended.

To this point, only Bonferroni and multiple Bonferroni methods have yet to be discussed. The performance of these two methods is very similar, with multiple Bonferroni being slightly more powerful. For this reason, only multiple Bonferroni will be discussed further.

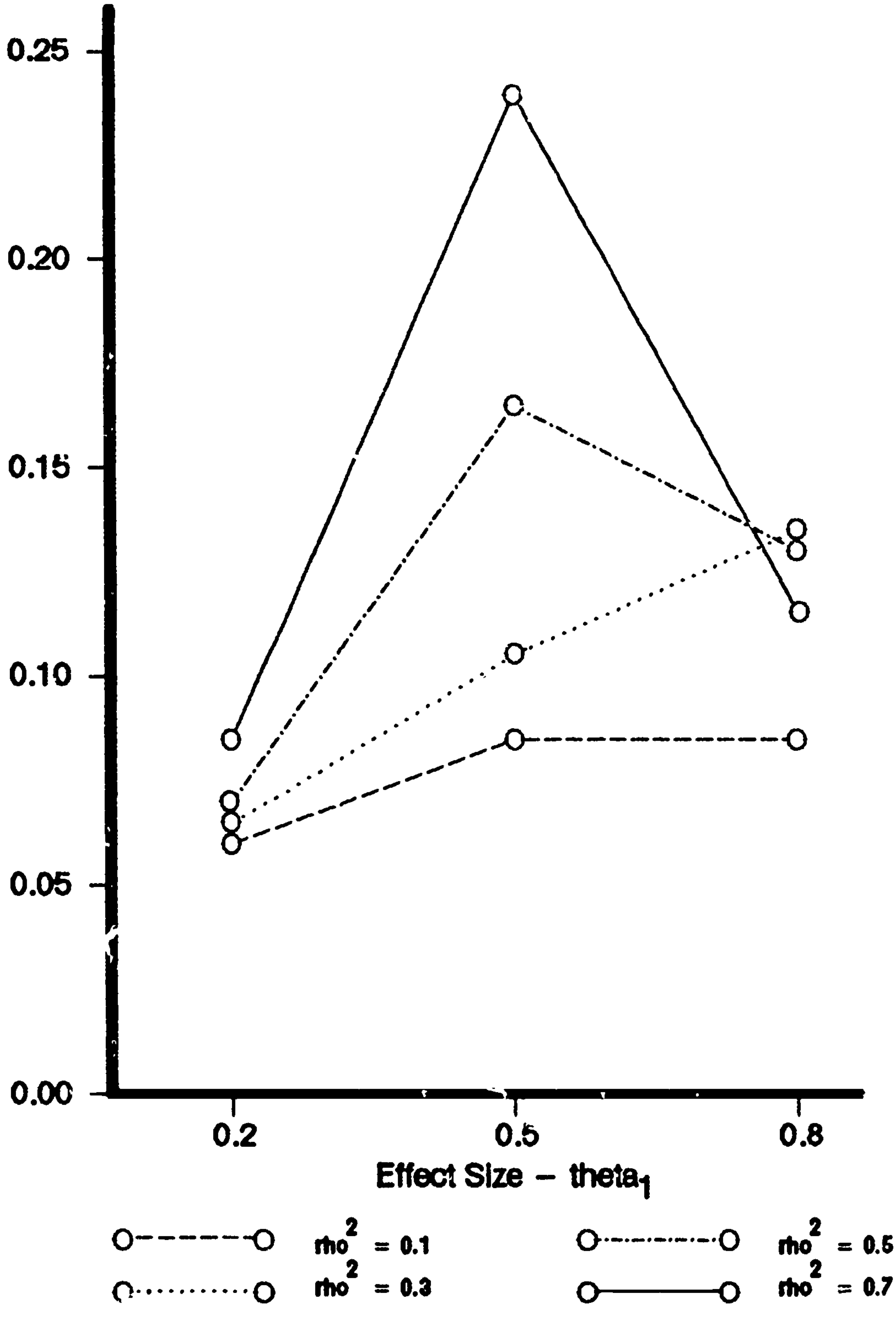

Figure 1 – Rate of Incorrectly Rejecting at Least One True Hypothesis with Method 6 when n = 30 and p = 9

A direct comparison of multiple Bonferroni and Method 7 leads us to conclude that while Method 7 is not uniformly better than multiple Bonferroni, it is, on balance, to be preferred. Tables 4 and 5 indicate that both methods provide slightly conservative experimentwise error rates. Table 1 shows that the probability of rejecting at least one true hypothesis when one hypothesis is false is similar for the two methods, with multiple Bonferroni being slightly more conservative. There is, however, a tendency for Method 7 to produce somewhat higher probabilities of error under the same conditions that cause Method 6 to have inflated error rates. In contrast to Method 6, though, Method 7 rates stay relatively close to the nominal value of 0.05. One of the worst cases is found in Table 8, where the rate increases to 0.086. Being 0.036 over the nominal value is more than offset, in our opinion, by the increase in power when using Method 7.

Tables 2 and 3 show that as $theta_1$ and rho^2 increase, the power of Method 7 improves relative to multiple Bonferroni, with the largest difference of 0.11 in Table 3 occurring when $rho^2 = 0.70$. In our study, 108 sets of conditions are used to compare the power of the methods. Across these, Method 7 has higher power than multiple Bonferroni in 81 cases (75%).

Table 4
Method 3: Multiple Bonferroni
Average Experimentwise Error Rate
for $theta_1=0$ and All Sample Sizes

Number of dependent variables	rho^2 0.1	0.3	0.5	0.7
3	0.045	0.040	0.042	0.030
6	0.047	0.043	0.038	0.028
9	0.047	0.039	0.030	0.024

Table 5

Method 7: Repeated T^2 Statistics Removing Variable with Maximum Significant F Statistic Average Experimentwise Error Rate for $theta_1$=0 and All Sample Sizes

Number of dependent variables	rho^2 0.1	0.3	0.5	0.7
3	0.041	0.035	0.038	0.026
6	0.042	0.037	0.029	0.025
9	0.041	0.037	0.028	0.025

Most of the differences in favor of multiple Bonferroni are small. Further, if one attends only to differences in power of at least 0.03, 50 of these differences favor Method 7, while only three favor multiple Bonferroni. The three differences favoring multiple Bonferroni were 0.031, 0.031 and 0.032, while the three largest differences favoring Method 7 were 0.309, 0.267 and 0.267. The conclusion is that Method 7 is the more powerful procedure, and in those situations where multiple Bonferroni is slightly better, the differences are so small as to have no practical importance.

Hummel and Sligo (1971) also compared methods with respect to the number of errors in experiments having at least one error, a measure of the degree to which errors "clump" together. As is shown in Table 6, Method 7 compares favorably with other methods.

In summary, then, the performance of Method 7 is well balanced with respect to experimentwise error rate, average numbers errors, power, and probability of at least one Type I error when one hypothesis is false. With an acceptable risk of error, one obtains better power with Method 7 than with other methods which provide a similar level of protection against Type I errors.

Table 6
Average Number of Type I Errors in Those Experiments Having at Least One Error
When rho^2=0.7, n=50, and p=9

	$theta_1$			
Method	0.0	0.2	0.5	0.8
Multivariate-Univariate	3.333	2.534	3.007	2.577
Multiple Bonferroni	2.765	2.526	3.071	3.300
Method 7	1.519	1.675	2.779	2.611

In addition to the computer runs described to this point, several additional runs were made to explore the generalizability of the findings. Tables 7 and 8 show results when two variables had non-zero values of *theta*. As can be seen, the results for the one false hypothesis and two false hypotheses cases are quite similar.

Two values of $theta_1$ (0.05 and 2.00) outside the range studied were used to see if these values had any unusual results. They did not.

Last, covariance matrices were used where the off-diagonal elements were not all equal. In the main, the study followed Hummel and Sligo (1971), using equal off-diagonal elements, a practice criticized by Wilkinson (1975). However, results on the multivariate normal distribution obtained by Gupta (1966) would lead one to believe that matrices with equal off-diagonal elements present no particular limitation. For example, Gupta's results indicate that one matrix with equal off-diagonal elements of, say, 0.5477, and another matrix with equal off-diagonal elements of, say, 0.8367, provide boundaries for experimentwise error rates such that all matrices with unequal off-diagonal elements bounded by 0.5477 and 0.8367 will have experimentwise error rates bounded by the

Table 7
Power for One False Hypothesis ($theta_1$ = 0.5)
Compared with Power for Two False Hypotheses ($theta_1$ and
$theta_2$ = 0.5) When rho^2 = 0.7, n = 30, and p = 9

	One false hypothesis	Two false hypotheses	
Method	$theta_1$=0.5	$theta_1$=0.5	$theta_2$=0.5
Univariate	0.472	0.474	0.487
Bonferroni	0.184	0.183	0.193
Multiple Bonferroni	0.184	0.198	0.200
Multivariate-Univariate	0.438	0.472	0.485
Multivariate	0.006	0.008	0.007
Method 6	0.737	0.757	0.761
Method 7	0.438	0.449	0.461

Table 8
Rate of Incorrectly Rejecting at Least One True Null Hypotheses for One False Hypothesis ($theta_1$ = 0.5) Compared with the Rate for Two False Hypotheses ($theta_1$ and $theta_2$ = 0.5) When rho^2 = 0.7, n = 30, and p = 9

Method	One false hypothesis $theta_1$=0.5	Two false hypotheses $theta_1$=0.5 $theta_2$=0.5
Univariate	0.146	0.126
Bonferroni	0.021	0.025
Multiple Bonferroni	0.022	0.027
Multivariate-Univariate	0.138	0.126
Multivariate	0.000	0.000
Method 6	0.238	0.271
Method 7	0.086	0.073

experimentwise error rates for the matrices with equal off-diagonal elements of 0.5477 and 0.8367 (the test statistics would be p univariate z statistics). Table 9 demonstrates that Gupta's results generalize to the kind of multivariate t distributions investigated in this study. Further, a detailed study of Table 9 supports an emergent generalization that experimentwise error is best predicted by the determinant of the correlation matrix for the p dependent variables, regardless of pattern in the correlations and including the presence of negative correlations.

Table 9
Comparison of the Experimentwise Error Rates for Univariate Method for the Homogeneous Matrices (rho = 0.5467, 0.7071, and .8367) and Heterogeneous Matrices When p=9

Matrix	Sample Size	Lowest Element	Highest Element	Determinant	Experimentwise Error Rate
1	10	0.5477	0.5477	9.422×10^{-3}	0.275
2	10	0.7071	0.7071	3.605×10^{-4}	0.209
3	10	0.8367	0.8367	3.898×10^{-6}	0.163
4	10	0.5500	0.7753	3.605×10^{-4}	0.215
5	10	-0.6500	0.7753	3.605×10^{-4}	0.227
1	30	0.5477	0.5477	9.422×10^{-3}	0.255
2	30	0.7071	0.7071	3.605×10^{-4}	0.210
3	30	0.8367	0.8367	3.898×10^{-6}	0.167
4	30	0.5500	0.7753	3.605×10^{-4}	0.229
5	30	-0.6500	0.7753	3.605×10^{-4}	0.206
1	50	0.5477	0.5477	9.422×10^{-3}	0.252
2	50	0.7071	0.7071	3.605×10^{-4}	0.210
3	50	0.8367	0.8367	3.898×10^{-6}	0.149
4	50	0.5500	0.7753	3.605×10^{-4}	0.211
5	50	-0.6500	0.7753	3.605×10^{-4}	0.200

REFERENCES

Bird, K. D. (1975). Simultaneous contrast testing procedures for multivariate experiments. *Multivariate Behavioral Research, 10*, 343-351.

Bray, J. H. and Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. *Review of Educational Research, 53*, 340-367.

Cramer, E. M. and Bock, R. D. (1966). Multivariate analysis. *Review of Educational Research, 36*, 604-617.

Gupta, S. S. (1963). Multivariate t. *Annals of Mathematical Statistics, 34*, 792-838.

Hummel, T. J. and Sligo, J. R. (1971). Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin, 76*, 49-57.

Johnston, C. B. (1986). *An Empirical Comparison of Size and Power for Seven Methods of Analyzing Multivariate Data in the Two-Sample Case.* Unpublished doctoral dissertation. University of Minnesota, Minneapolis.

Miller, R. G. (1966). *Simultaneous Statistical Inference.* New York: McGraw-Hill.

Morrison, D. F. (1967). *Multivariate Statistical Methods.* New York: McGraw-Hill.

Ramsey, P. H. (1982). Empirical power of procedures for comparing two groups on p variables. *Journal of Educational Statistics, 7*, 139-156.

Roy, S. N. and Bose, R. C. (1953). Simultaneous confidence interval estimation. *Annals of Mathematical Statistics, 24*, 513-536.

Wilkinson, L. (1975). Response variable hypotheses in the multivariate analysis of variance. *Psychological Bulletin, 82*, 408-412.