THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A COMPARISON OF SCHEFFÉ'S S-METHOD AND TUKEY'S T-METHOD FOR VARIOUS NUMBERS OF ALL POSSIBLE CONTRASTS UNDER VIOLATION OF ASSUMPTIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

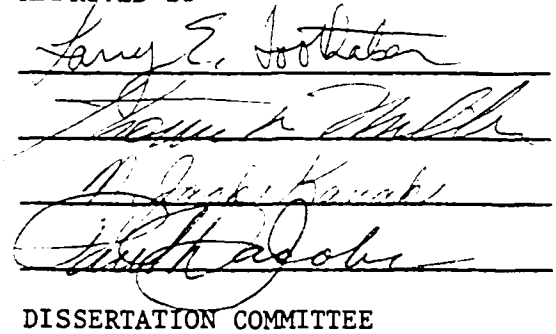
H. J. KESELMAN

Norman, Oklahoma

1971

A COMPARISON OF SCHEFFÉ'S S-METHOD AND TUKEY'S T-METHOD FOR VARIOUS NUMBERS OF ALL POSSIBLE CONTRASTS UNDER VIOLATION OF ASSUMPTIONS

APPROVED BY



DISSERTATION COMMITTEE

ACKNOWLEDGMENTS

Gratitude takes three forms: a feeling in the heart, an expression in words and a giving in return. (Author unknown).

. . . my parents

. . . my wife

. . . Larry Toothaker, friend and Chairman of my dissertation committee

. . . Jack Kanak, friend and committee member

. . . Tom Miller, friend and committee member

. . . Paul Jacobs, friend and committee member

TABLE OF CONTENTS

Manuscript to be submitted for publication.

Second manuscript to be submitted for publication.

A COMPARISON OF SCHEFFÉ'S S-METHOD AND TUKEY'S T-METHOD FOR VARIOUS NUMBERS OF ALL POSSIBLE CONTRASTS UNDER VIOLATION OF ASSUMPTIONS

INTRODUCTION

The researcher in the behavioral sciences intending to explore multiple treatment-effects, where there are two or more levels of the treatment variable, has available to him a most versatile statistical tool to aid him in evaluating his data. Indeed, the analysis of variance (ANOVA) and its various theoretical-mathematical models have at last become a primary statistical tool to aid behavioral researchers evaluate their data.

A one-way fixed effects ANOVA is one of the models typically employed by behavioral scientists. The hypothesis of interest usually is that $\mu_1 = \mu_2 = \ldots = \mu$. That is, the null hypothesis subjected to a statistical test is that the population means for the various treatment levels are equal; hence the observations within each treatment level have been randomly sampled from one population with mean μ. Having rejected the null hypothesis for the one-way fixed effects ANOVA, the researcher can conclude in probabilistic terms that the means differ statistically. If the experimenter was interested in determining whether any treatment effects existed, then the one-way ANOVA is indeed a most convenient and versatile statistical tool to

detect such differences and this hypothetical experimenter could consider that his statistical question had been answered. On the other hand, the above example is indeed hypothetical in that it is a rare occasion when the experimenter is content in just being able to state that the treatment levels were different. Of course, the experimenter's interest then is for further exploration of these different means. The one-way fixed effects ANOVA merely reflects whether at least any two of the treatment levels differ. Did treatment level one differ from level two or level three or level four or perhaps the combination of treatment levels one and two differed from the combination of three and four, etc., etc.? These are the type of questions that are generally of interest. How many researchers are content in merely being able to say that there were differences, without being able to specify exactly where the differences lie?

Because behavioral scientists are usually interested in digging deeper into their data, probing techniques were developed to be used following the rejection of the analysis of variance null hypothesis.

Tukey (1953, unpublished, privately circulated manuscript) is credited by Scheffé for devising a method to simultaneously estimate all contrasts (Scheffé, 1959). Tukey's technique, the T-method, utilizes the Studentized Range distribution to investigate differences among means, following the rejection of the ANOVA null hypothesis. Scheffé (1959) states that for Tukey's T-method the probability is $1-\alpha$ that the relationship in (1) holds for all pairwise contrasts.

$$(\bar{X}_k - \bar{X}_{k'}) - q_{\nu_1 \nu_2}\left(MS_e/n\right)^{\frac{1}{2}} \leq (\mu_k - \mu_{k'}) \leq (\bar{X}_k - \bar{X}_{k'}) + q_{\nu_1 \nu_2}\left(MS_e/n\right)^{\frac{1}{2}} \quad (1)$$

In repeated experiments therefore, the probability is 1-α that all pairwise intervals simultaneously cover their true value of the population contrast. In its original formulation, according to Scheffé (1959), Tukey's method was designed to set limits around pairwise contrasts, e.g., $\hat{\psi} = c_1\bar{X}_{.k} + c_2\bar{X}_{.k'}$. Scheffé (1959), Winer (1962), and Kirk (1968) present ammended procedures for Tukey's T-method, sometimes called the Honestly Significant Difference technique, that are appropriate for contrasts other than pairwise contrasts, and also when the number of observations per treatment level are not equal.

To circumvent the limited applicability of Tukey's T-method, Scheffé (1953, 1959) formulated his S-method which is a generalized version of Tukey's method but uses the sampling distribution of F. For all possible contrasts the probability is 1-α that all contrasts simultaneously satisfy the relationship in (2).

$$\hat{\psi} - \left[(k-1)\, F_{\nu_1\nu_2}\right]^{\frac{1}{2}}\left[MS_e\, \Sigma\left(c_k^2/n_k\right)\right]^{\frac{1}{2}} \leq \psi \leq \hat{\psi} + \left[(k-1)F_{\nu_1\nu_2}\right]^{\frac{1}{2}}\left[MS_e\, \Sigma\left(c_k^2/n_k\right)\right]^{\frac{1}{2}} \quad (2)$$

The probability is 1-α that the confidence intervals for all contrasts will simultaneously cover their true psi values. For example, for four treatment levels there are twenty-five possible contrasts. In 1000 experiments the probability should be (1-α)% that all contrasts simultaneously bracket their true psi values. If the 95% confidence limit was chosen, 950 experiments would have all twenty-five intervals bracketing their true psi values; 50 experiments will have at least one interval (experimentwise error rate) not bracketing its true psi value.

Scheffé's S-method is not dependent upon equal variances nor consequently equal sample sizes, for its validity. Also, Scheffé's technique is applicable to any form of contrast and not merely to pairwise contrasts.

In addition to setting limits around a contrast, Tukey's and Scheffé's techniques can be used to test the hypothesis that the contrast equals zero, e.g., $\Psi = 0$. Scheffé (1953, 1959) states that the hypothesis is tested according to whether his interval inclusively includes or excludes the value of $\Psi = 0$. According to Scheffé (1953, 1959) and Miller (1966), the null hypothesis for the ANOVA is equivalent to the statement that all the contrasts are zero.

There have been few comparisons between Tukey's T-method and Scheffé's S-method. Scheffé (1953, 1959) compared the relative efficiencies of the two methods for a one-way fixed effects analysis of variance for four and six treatment levels. Scheffé's comparison was restricted to conditions of equal variances for the contrasts and equal observations per treatment level (the restriction under which Tukey's T-method was derived). When there were four levels of the treatment factor, Tukey's method was more efficient than the Scheffé method for the pairwise contrasts, e.g., $\hat{\psi} = (+1)\overline{X}._k + (-1)\overline{X}._{k'}$. For six treatment levels, the T-method was more efficient for not only the pairwise contrasts but also when comparing one mean with the average of two other means, e.g., $\hat{\psi} = (+1)\overline{X}._k + (-\frac{1}{2})\overline{X}._{k'} + (-\frac{1}{2})\overline{X}._{k''}$.

For the pairwise contrasts Tukey's method is preferable, while for the more complicated contrasts Scheffé's method is more efficient and gives shorter intervals.

Petrinovich and Hardyck (1969), while not specifically focusing their attention on just the Tukey and Scheffé methods, nevertheless provide data on the two techniques, enabling us to compare them. Under the null hypothesis conditions, both techniques control the Type I error as they were designed to do, but the empirical probability of a Type I error for Scheffé's method was consistently less than theoretical alpha, .05; it appears to over protect. Consistent with this pattern of overprotection for the first type of error, the empirical probability of a Type II error for the S-method was larger than the probability found for the T-method and was therefore generally less powerful. Since Petrinovich and Hardyck limited their study to pairwise contrasts, their findings are consistent with Scheffé's analytical results. Specifically then, for pairwise contrasts, Tukey's method sets shorter intervals and is more powerful in detecting differences for this type of contrast. However for both the Scheffé and Tukey methods the empirical probability of a Type I error was oftentimes conservative.

Ryan (1959, 1962), concerned with controlling the number of false statements, in particular Type I error statements, argues that the probability of a Type I error can be adequately controlled by adopting the experimentwise error rate. The Tukey and Scheffé methods are techniques which control the probability of a Type I error experimentwise. Consider the probability statement that is made for the Scheffé

method: the probability is 1-α that all contrasts simultaneously cover their true psi values. The probability statement for the Tukey method was originally intended for pairwise contrasts (Scheffé, 1959, p. 76).

Petrinovich and Hardyck (1969) investigated only pairwise contrasts in their research. They found that when a subset of all possible contrasts are performed, i.e., pairwise contrasts, the empirical probability of a Type I error for the Scheffé method was generally less than theoretical alpha. If the probability of a Type I error is related to the number of contrasts investigated, then, the results reported by Petrinovich and Hardyck are biased.

Generally, it is desirable to control the probability of a Type I error and set alpha at a conservative experimentwise level of significance. It would be beneficial to the researcher, however, if he knew whether the probability of a Type I error for his statistical test would fluctuate from theoretical alpha because of the number of contrasts that would be computed. The first objective for this research was, therefore, to determine for the Scheffé and Tukey methods whether the empirical probability of a Type I error for all possible contrasts, deviates from the probability when only 75 per cent, 50 per cent, 25 per cent, and the pairwise contrasts are computed.

The second phase of this research will investigate the empirical probability of a Type II error for the S and T methods. When a researcher uses the techniques correctly, that is after a significant ANOVA F test, then his concern should be for detecting the differences in the means which the overall ANOVA F test indicates are present.

That is, he does not want to commit a Type II error. Thus, the researcher should carefully control the probability of committing a Type II error rather than placing his emphasis on controlling too stringently the probability of a Type I error. He should be concerned about the power of the test as well as the probability of a Type I error (Scheffé, 1959, p. 361).

Procedure

Pseudo-random numbers were selected, using a pseudo random number generator.[1] Depending upon the assumption violation, the numbers were selected from either a normal or skewed distribution with $\mu = 0$ and $\sigma = 1$. The random numbers were distributed to the four treatment levels that comprise a one-way fixed effects analysis of variance. Mean differences (differences between adjoining means expressed in standard deviation units) were set at 2.6 σ-unit differences. Therefore, $\mu_1 = 0$, $\mu_2 = 2.6$, $\mu_3 = 5.2$, and $\mu_4 = 7.8$.

The observations from the normal distribution were generated by means of GAUSS (IBM, 1967), which generates pseudo-random normal deviates with $\mu = 0$ and $\sigma = 1$. The skewed population was derived from a chi-square distribution with three degrees of freedom, having mean three, variance six, third moment 24, fourth moment 252, skewness measure $\gamma_1 = \left(\mu_3^2 / \mu_2^3\right)^{1/2} = 1.663$ and kurtosis measure $\gamma_2 = \mu_4 / \mu_2^2 - 3 = 4$ (Kendall and Stuart, 1969).

---

[1] I would like to thank Jesse May for extending RANDU (IBM, 1967), so that the contingencies of this particular research problem could be handled by an IBM 1130 machine. The recycling of a random number generator is determined by the word length, which is a function of the machine size. Jesse May altered RANDU by increasing the word length, via an assembler subroutine, thereby substantially increasing the cycling pattern to that of an IBM 360/50 machine.

Pseudo-random chi-square variables with three degrees of freedom were generated by summing the squares of three N(0, 1) variables. The numbers were then scaled so that the mean and variance of the skewed population would be the same as the mean and variance of the normal population, first by subtracting three from each score and then multiplying by $1/\sigma$, where $\sigma^2 = 6$. The resulting skewed population has mean zero, variance one, skewness measure $\gamma_1 = 1.663$ and kurtosis measure $\gamma_2 = 4$, as γ_1 and γ_2 are invariant under additive and multiplicative transformations.

The choice of sample size was guided by two dictates: (1) conformability to a tabled value of the Studentized Range distribution and (2) large enough to show differences, if any, in the power of the multiple comparison tests. Since population mean values were prespecified, determining $\sum_k^K \alpha^2$ was straightforward and consequently the sample size could be calculated such that the power was at least .90 for detecting 2.6 σ-unit differences with the ANOVA $\underline{F}$ test.

For any one sampling from the random number generator, a one-way fixed effects ANOVA was calculated. If the $\underline{F}$ value equaled or exceeded the critical $\underline{F}$ value, the multiple comparison procedures were initiated. When the obtained $\underline{F}$ value failed to reach significance the program returned to the random number generator and the sampling procedure and an ANOVA $\underline{F}$ test was once again performed.

To achieve mean differences of 2.6 σ-units, the random numbers Y were transformed, X = Y + MD (mean difference) where MD was incremented by 2.6 σ-units, into X variates, thereby creating the differences for the means of the four treatment levels. In the name

of efficiency, the contrasts were not explored under the null hypothesis condition until the transformed data, X, yielded a significant F value. Once significance was obtained, the flow of the program passed to the stage of calculating the linear contrasts. At that time, contrasts were calculated on the X data (to check for Type II errors) and also on the stored Y data (to check for Type I errors). By adopting the above procedure, both types of errors could be counted with just a single pass through the random number generator.

All possible contrasts, 75 per cent, 50 per cent, 25 per cent, and all pairwise contrasts were computed. To randomly select 75, 50, and 25 per cent of all possible contrasts, the random number generator was again utilized. The random numbers were rescaled to inclusively contain the numbers one through twenty-five and were then used as subscripts to randomly select the contrasts. After selecting 18 (75%) unique random numbers, these distinct numbers were associated via the subscripts with a particular contrast. The same procedure was used for randomly selecting 12 (50%) and six (25%) of the contrasts. Since the pairwise contrasts were stored as the first six contrasts, locating and recalling these contrasts was straightforward. For each set of contrasts, all possible, 75 per cent, 50 per cent, 25 per cent, and all pairwise contrasts, Scheffé's S-method and Tukey's T-method were calculated to determine the number of contrasts which did or did not bracket zero. The procedure of generating random samples (K=4) with n_k observations per cell, and thereafter, if the F test was significant, calculating Scheffé's

and Tukey's multiple comparison procedures constituted one single experiment; the procedure was repeated for 1000 experiments.

Unequal variances and unequal sample sizes were combined when sampling from a normal distribution to explore the two types of error of Scheffé's and Tukey's procedures under conditions of assumption violations. Therefore, the five combinations examined when sampling from a normal distribution were: (1) equal observations per treatment level - equal variances, (2) equal observations per treatment level - unequal variances, (3) unequal observations per treatment level - equal variances, (4) unequal observations per treatment level - unequal variances (proportionately paired) and (5) unequal observations per treatment level - unequal variances (inversely proportionately paired). These five conditions were also investigated for the non-normal skewed population.

The first criterion for selecting the unequal sample sizes was to get ϕ (non centrality parameter for the noncentral F distribution) as close to the value of ϕ for the equal sample case. The second criterion was to have the sample size divergent enough to be interesting as an assumption violation. Both criteria, it is believed, were adequately satisfied.

The Tukey method was derived under the restriction that the variances of the contrasts are equal. In order to satisfy this restriction there must be an equal number of observations per cell. The harmonic mean, one of many suggested procedures that can be employed with the Tukey method when there are an unequal number of observations per cell, was used in the present investigation (Kirk, 1968; Smith, 1971).

For comparisons involving unequal variances, the variances were specified to be in the ratio of 1:2:3:4. A further qualifier placed upon the choice of values for the variances was that the average of the variances should equal one, thereby not differentially affecting the original calculations of sample size for a desired power which was calculated for the ANOVA F test.

Results

Normal Distribution, Equal n's (7), Equal σ^2's (1):

The empirical probability of a Type I error for .67 σ-unit differences varies with the different number of contrasts that are computed. For all possible contrasts, the probability of a Type I error when setting the error rate experimentwise, coincides with theoretical alpha, within the bounds of sampling variability. The probability of a Type I error generally decreases when the subsets of all possible contrasts are sampled. That is, the empirical probability of a Type I error when 75 per cent of the contrasts (18 contrasts) are sampled is smaller than the probability when all possible contrasts (25 contrasts) are computed. Similarly, the probability of a Type I error decreases for each succeeding subset. The probability is larger when 50 per cent (12 contrasts) of the contrasts are sampled than when 25 per cent (six contrasts) of the contrasts are computed. The probability of a Type I error for the six pairwise contrasts is generally larger than the probability for 25 per cent of the contrasts, even though both subsets sample six contrasts. Evidently, the likelihood of committing a Type I error when six pairwise contrasts are computed is greater than if just six contrasts are chosen randomly from among all the possible contrasts.

Table 1 contains the empirical probabilities of a Type I error for the Scheffé and Tukey methods for the experimentwise, per comparison and per experiment error rates (Ryan, 1959).

For all possible contrasts, when the error rate is set experimentwise, the empirical probability of a Type I error for Scheffé's method should be close to theoretical alpha since the S-method was designed to control the experimentwise error rate at α. The empirical value of .052 is indeed consistent with theoretical alpha. For pairwise contrasts, Tukey's method is designed to control the Type I error experimentwise, and we would expect an empirical value close to $\alpha = .05$. The value of .059 for pairwise contrasts for the T-method indicates agreement within sampling variability. However, the probability of a Type I error for the S-method, designed for all possible contrasts (Scheffé, 1958, p. 76), is less than theoretical alpha for the pairwise contrasts, as was found by Petrinovich and Hardyck (1969). In general, the Scheffé and Tukey methods do control the probability of a Type I error for the experimentwise error rate, but, the latitude of protection that each method provides varies with the number of contrasts computed.

The empirical probability of a Type I error when counting with a per comparison rule, is extremely conservative regardless of the number of contrasts that are computed. The tabled probabilities are indicative of the probability of an error for any one t test if the error rate is set at .05 experimentwise for the set of all possible multiple t tests (Aitkin, 1969, p. 195). The empirical values in Table 1 are within sampling variability of the theoretical value of

Table 1. Type I Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, .67 MD, Normal Distribution, Equal n's (7), Equal σ^2's (1).

Error Rates	Contrasts	Empirical Estimates Scheffé	Tukey
Experimentwise	All Possible	.052	.059
	75 per cent	.048	.048
	50 per cent	.042	.044
	25 per cent	.034	.031
	Pairwise	.039	.059
Comparison	All Possible	.007	.006
	75 per cent	.007	.006
	50 per cent	.007	.006
	25 per cent	.008	.007
	Pairwise	.008	.012
Experiment	All Possible	.179	.152
	75 per cent	.119	.100
	50 per cent	.088	.074
	25 per cent	.047	.040
	Pairwise	.046	.074

$.05\sigma_P = .007$

.0065.

The number of Type I errors for the per experiment error rate is similar to the pattern found with the experimentwise rate. That is, the long run average of a Type I error fluctuates with the number of contrasts sampled. One can anticipate committing more errors if all of the possible contrasts are computed and fewer errors when working with subsets of all the possible contrasts.

Table 2 contains the empirical probabilities of a Type II error for the three error rates with .67 mean differences. The probabilities are excessive regardless of how many contrasts are sampled. For the experimentwise error rate the probabilities indicate a one hundred per cent likelihood of committing a Type I error; therefore, in all 1000 experiments at least one Type II error was committed. Even for the per comparison rule of counting errors, the probabilities are unreasonably in excess of acceptable standards.

For .67 σ-unit differences with seven observations per cell the ANOVA $\underline{F}$ test would detect these differences approximately ninety per cent of the time. The probabilities of a Type II error from Table 2, would indicate that the power that was "built into" the ANOVA $\underline{F}$ test does not carry over to the Scheffé and Tukey methods. This important finding should be reiterated: the Scheffé and Tukey methods do not retain the same degree of power for detecting mean differences that had been "built into" the ANOVA $\underline{F}$ test.

From the data enumerated in Table 2, it would appear that .67 σ-unit differences and seven observations per cell are not of sufficient magnitude to bring to light any possible differences in

Table 2. Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, .67 MD, Normal Distribution, Equal n's (7), Equal σ^2's (1).

Error Rates	Contrasts	Empirical Estimates Scheffé	Tukey
Experimentwise	All Possible	1.000	1.000
	75 per cent	1.000	1.000
	50 per cent	1.000	1.000
	25 per cent	.998	.999
	Pairwise	1.000	1.000
Comparison	All Possible	.677	.718
	75 per cent	.678	.718
	50 per cent	.679	.718
	25 per cent	.682	.721
	Pairwise	.691	.639
Experiment	All Possible	16.933	17.950
	75 per cent	12.197	12.929
	50 per cent	8.145	8.620
	25 per cent	4.091	4.328
	Pairwise	4.148	3.835

$.05\sigma_p = .007$

the empirical probabilities of a Type II error for the different number of contrasts computed, when counting the errors with the experimentwise error rate. Therefore, if the empirical probability of a Type II error does fluctuate with the number of contrasts sampled, the σ-unit differences between the means and/or the sample size per cell had to be increased so that, such differences if present, would be most obvious.

Tables 3 and 4 contain the empirical probability of a Type I and Type II error, respectively, when there were seven observations per cell but the mean differences had been increased to 2.6 σ-units. The probability of a Type I error is not affected by mean differences; what can be gleaned from Table 3 is that the relationship between the probability of a Type I error and the number of contrasts computed is again evident. The probability of a Type II error should be and is affected by the increase in the mean differences. The empirical probabilities begin to show some additional divergence due to the number of contrasts sampled as can be seen from Table 4. Since the empirical probabilities for the experimentwise error rate were still excessive, except for the pairwise contrasts, and therefore could not show meaningful differences as a function of the number of contrasts sampled, the next change was an increase in sample size to 11 observations per cell.

Sampling from a normal distribution with variance one, for 2.6 σ-unit differences between the means, and with 11 observations per cell, Table 5 contains the probabilities of a Type I and Type II error. The probability of a Type II error for the experimentwise

Table 3. Type I Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Normal Distribution, Equal n's (7) Equal σ^2's (1).

Error Rates	Contrasts	Empirical Estimates Scheffé	Tukey
Experimentwise	All Possible	.037	.047
	75 per cent	.034	.043
	50 per cent	.031	.029
	25 per cent	.020	.019
	Pairwise	.022	.047
Comparison	All Possible	.005	.004
	75 per cent	.006	.005
	50 per cent	.006	.004
	25 per cent	.005	.004
	Pairwise	.004	.010
Experiment	All Possible	.136	.111
	75 per cent	.103	.088
	50 per cent	.073	.055
	25 per cent	.029	.027
	Pairwise	.027	.057

$.05\sigma_P = .007$

Table 4. Type II Error Rates for Scheffé's S-Method and Tukey's T-Method:

$\alpha = .05$, 2.6 MD, Normal Distribution, Equal n's (7) Equal σ^2's (1).

Error Rates	Contrasts	Empirical Estimates Scheffé	Tukey
Experimentwise	All Possible	1.000	1.000
	75 per cent	.994	.995
	50 per cent	.956	.965
	25 per cent	.713	.729
	Pairwise	.103	.065
Comparison	All Possible	.181	.193
	75 per cent	.181	.193
	50 per cent	.182	.194
	25 per cent	.177	.185
	Pairwise	.018	.012
Experiment	All Possible	4.530	4.828
	75 per cent	3.251	3.477
	50 per cent	2.179	2.331
	25 per cent	1.060	1.108
	Pairwise	.109	.069

$.05\sigma_P = .007$

Table 5. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Normal Distribution, Equal n's (11), Equal σ^2's (1).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.038	.049	1.000	1.000
	75 per cent	.036	.044	.990	.994
	50 per cent	.031	.032	.905	.919
	25 per cent	.018	.015	.644	.656
	Pairwise	.028	.049	.004	.003
Comparison	All Possible	.006	.005	.144	.152
	75 per cent	.006	.005	.145	.155
	50 per cent	.006	.005	.146	.155
	25 per cent	.004	.004	.140	.148
	Pairwise	.006	.011	.001	.001
Experiment	All Possible	.141	.118	3.598	3.811
	75 per cent	.102	.086	2.606	2.763
	50 per cent	.069	.057	1.754	1.864
	25 per cent	.026	.021	.840	.890
	Pairwise	.033	.064	.004	.003

$.05\sigma_P = .007$

error rate is still excessive except for the pairwise contrasts. The mean differences and sample size were evidently still not large enough to show meaingful differences between the probability for all possible contrasts and the probabilities when 75 and 50 per cent of the contrasts were computed. Nonetheless, it would have been unrealistic to again increase the mean differences and also costly, in terms of computer time, to increase the sample size. Though Petrinovich and Hardyck (1969) showed that increasing the mean differences is one procedure that will reduce the probability of a Type II error, the procedure cannot be used by the behavioral scientist, for he can only work with the mean differences that have been brought about by his manipulation of the independent variable(s). Therefore, the remaining conditions dealing with assumption violations were examined for 2.6 σ-unit differences and 11 observations per cell. The effect of the assumption violation in each condition may be readily assessed by remembering that Scheffé's method, for all possible contrasts, and Tukey's method, for pairwise contrasts, should yield empirical alpha values close to an alpha of .05.

Normal Distribution, Equal n's (11), Unequal σ^2's (.4, .8, 1.2, 1.6):

The empirical probability of a Type I and Type II error for Scheffé's S-method and Tukey's T-method are contained within Table 6. When sampling from populations with unequal variances, the probability of a Type I error for the S-method is greater than theoretical alpha for the experimentwise error rate with all possible contrasts computed. The probability decreases until 25 per cent of the contrasts

Table 6. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Normal Distribution, Equal n's (11), Unequal σ^2's (.4, .8, 1.2, 1.6).

Error Rates	Contrasts	Empirical Estimates: Type I errors		Empirical Estimates: Type II errors	
		Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.063	.075	1.000	1.000
	75 per cent	.061	.064	.990	.993
	50 per cent	.053	.051	.925	.934
	25 per cent	.038	.033	.665	.685
	Pairwise	.045	.075	.028	.015
Comparison	All Possible	.010	.008	.156	.164
	75 per cent	.010	.008	.155	.163
	50 per cent	.009	.006	.155	.163
	25 per cent	.010	.008	.159	.167
	Pairwise	.009	.016	.005	.002
Experiment	All Possible	.247	.192	3.902	4.112
	75 per cent	.176	.137	2.788	2.932
	50 per cent	.105	.078	1.856	1.959
	25 per cent	.060	.046	.956	1.002
	Pairwise	.056	.093	.028	.015

$.05\sigma_P = .007$

are sampled and increases for the pairwise contrasts. Thus, the empirical probabilities do differ from theoretical alpha when computing the different number of contrasts. Also, for the pairwise contrasts, the empirical probability for the Tukey method is larger than alpha, but by a larger amount than that for the S-method for all possible contrasts. Compared to what has been found when investigating assumption violations for the ANOVA F test (Box, 1954a, b; Box and Anderson, 1955; Horsnell, 1953), variance heterogeneity, for the S and T multiple comparison procedures, may or may not cause the empirical probability of a Type I error to substantially differ from theoretical alpha; the correspondence between the empirical probabilities and theoretical alpha is conditional upon both the number of contrasts computed and the variance heterogeneity.

The empirical probability of a Type I error for the per comparison error rate and the long run average found for the per experiment error rate also reflect the effect of variance heterogeneity and the effect of the number of contrasts computed.

The probability of a Type II error for all possible and pairwise contrasts, is generally greater than the probabilities when sampling is from populations with equal variances.

Normal Distribution, Unequal n's (8, 9, 11, 16), Equal σ^2's(1):

Table 7 contains the probability of a Type I and Type II error for the three error rates when the number of observations per treatment level are unequal. Again the empirical probabilities do vary with the number of contrasts sampled. For the experimentwise Type I

Table 7. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Normal Distribution, Unequal n's (8, 9, 11, 16), Equal σ^2's (1).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.039	.050	1.000	1.000
	75 per cent	.036	.041	.992	.993
	50 per cent	.034	.041	.914	.920
	25 per cent	.023	.019	.621	.642
	Pairwise	.027	.050	.015	.003
Comparison	All Possible	.005	.004	.149	.157
	75 per cent	.005	.004	.146	.154
	50 per cent	.006	.005	.152	.160
	25 per cent	.005	.004	.139	.149
	Pairwise	.005	.010	.002	.000
Experiment	All Possible	.132	.109	3.727	3.936
	75 per cent	.097	.082	2.626	2.769
	50 per cent	.071	.061	1.823	1.925
	25 per cent	.029	.025	.836	.895
	Pairwise	.032	.063	.015	.003

$.05\sigma_P = .007$

error rate the Tukey empirical probabilities for all possible and pairwise contrasts are equal to theoretical alpha. The probability of a Type I error for the Scheffé method is slightly conservative for all possible contrasts, and conservative for the remaining sets of contrasts.

The empirical probabilities of a Type II error for the three error rates generally, do not substantially differ from the data for the equal n's case.

Normal Distribution, Unequal n's (8, 9, 11, 16), Unequal σ^2's (.4, .8, 1.2, 1.6):

Sampling for this assumption violation was from populations with different variances. The unequal sample sizes were proportional to the variances. That is, the smaller sample was taken from the population with the smaller variance, while the largest sample size was paired with the population with the largest variance. Consistent with previous findings (Box, 1954a, b; Box and Anderson, 1955; Horsnell, 1953) the empirical probability of a Type I error is less than theoretical alpha, when proportionately pairing unequal variances and unequal sample sizes. Table 8 contains the two types of errors for the three error rates. For the experimentwise rate for all possible contrasts, the empirical probability of a Type I error is conservative for the Scheffé method. For the pairwise contrasts, the probability of a Type I error for Tukey's method is very conservative, in fact, it is 5 σ_p less than $\alpha = .05$.

Since decreasing alpha will, when all other factors are held constant, increase beta, the probability of a Type II error is somewhat

Table 8. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: α = .05, 2.6 MD, Normal Distribution, Unequal n's (8, 9, 11, 16), Unequal σ^2's (.4, .8, 1.2, 1.6).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.023	.015	1.000	1.000
	75 per cent	.021	.012	.991	.993
	50 per cent	.016	.007	.937	.950
	25 per cent	.015	.008	.722	.744
	Pairwise	.013	.015	.046	.027
Comparison	All Possible	.003	.001	.177	.190
	75 per cent	.003	.002	.175	.187
	50 per cent	.002	.001	.176	.189
	25 per cent	.004	.002	.177	.189
	Pairwise	.003	.003	.008	.004
Experiment	All Possible	.071	.035	4.427	4.739
	75 per cent	.058	.030	3.145	3.370
	50 per cent	.030	.014	2.106	2.264
	25 per cent	.021	.011	1.063	1.136
	Pairwise	.017	.019	.048	.027

$.05\sigma_p$ = .007

higher than when there are no assumption violations. Consequently, when there are unequal variances and unequal sample sizes and the sample sizes and variances are proportional, there is an increase in the probability of a Type II error and accordingly, the statistical tests lose some of their power for detecting mean differences.

Normal Distribution, Unequal n's (8, 9, 11, 16), Unequal σ^2's (1.6, 1.2, .8, .4):

For the inversely proportional pairings of unequal variances and unequal sample sizes, the empirical probabilities for the experimentwise Type I error far exceeds theoretical alpha, as expected from what had been found with the ANOVA F test (Box, 1954a, b). Table 9 contains the two types of errors for the S and T methods. Even for the pairwise contrasts, the probability of a Type I experimentwise error is larger than can be attributed to sampling variability. However, the Scheffé method appears to be more robust to this type of violation than does Tukey's method, even for pairwise contrasts.

The empirical probabilities of a Type II error do not substantially differ from the data reported when there are no assumption violations, although the probabilities are slightly larger.

Skewed Distribution:

The empirical probabilities of a Type I and Type II error when sampling from the skewed distribution are enumerated in Tables 10-14 for the same five conditions that were investigated when sampling from a normal distribution.

The probabilities are generally consistent with the probabilities found when sampling from a normal distribution. However, for some

Table 9. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Normal Distribution, Unequal n's (8, 9, 11, 16), Unequal σ^2's (1.6, 1.2, .8, .4).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.147	.186	1.000	1.000
	75 per cent	.144	.168	.994	.993
	50 per cent	.129	.149	.915	.921
	25 per cent	.104	.111	.650	.673
	Pairwise	.113	.186	.064	.028
Comparison	All Possible	.029	.029	.149	.155
	75 per cent	.030	.029	.152	.160
	50 per cent	.029	.028	.150	.156
	25 per cent	.029	.030	.149	.155
	Pairwise	.029	.048	.011	.005
Experiment	All Possible	.733	.731	3.725	3.869
	75 per cent	.533	.525	2.728	2.825
	50 per cent	.350	.338	1.796	1.869
	25 per cent	.176	.180	.892	.931
	Pairwise	.172	.290	.064	.028

$.05\sigma_P = .007$

of the conditions investigated, the empirical probabilities are quite different from those found in the same conditions with a normal distribution, and will be the only conditions discussed for the skewed distribution.

When the variances are unequal and the observations are sampled from a skewed distribution (Table 11), the empirical experimentwise Type I probabilities for all possible and pairwise contrasts differ from the probabilities when sampling is from a normal distribution. For the skewed distribution, the value for the Scheffé method for all possible contrasts is less than alpha, where for the normal distribution, it is considerably larger than theoretical alpha (Table 6). The value for the Tukey method for the pairwise contrasts is also closer to theoretical alpha.

The Type II probabilities differ when sampling is from populations with different variances and the sample sizes are unequal and proportional to the variances (Table 13). The empirical probabilities are generally smaller, for the three error rates, when the observations are sampled from the skewed distribution. Consequently, the power for the Scheffé and Tukey tests is increased when the variances and sample sizes are unequal and proportionately paired and the observations are sampled from a skewed distribution.

The most notable change in the empirical probabilities for a Type I error occur when sampling is from the skewed distribution and the variances and sample sizes are unequal and inversely proportional to one another (Table 14). The empirical probability of a Type I error is considerably closer to theoretical alpha when sampling is

Table 10. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Skewed Distribution, Equal n's (11), Equal σ^2's (1).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.036	.037	1.000	1.000
	75 per cent	.032	.032	.988	.988
	50 per cent	.029	.025	.914	.920
	25 per cent	.011	.013	.656	.668
	Pairwise	.020	.037	.015	.010
Comparison	All Possible	.004	.003	.141	.147
	75 per cent	.004	.003	.140	.145
	50 per cent	.004	.003	.145	.151
	25 per cent	.002	.002	.148	.155
	Pairwise	.004	.008	.002	.002
Experiment	All Possible	.105	.084	3.535	3.674
	75 per cent	.075	.061	2.513	2.603
	50 per cent	.051	.040	1.740	1.813
	25 per cent	.014	.015	.890	.929
	Pairwise	.025	.046	.015	.010

$.05\sigma_P = .007$

Table 11. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Skewed Distribution, Equal n's (11), Unequal σ^2's (.4, .8, 1.2, 1.6).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.046	.056	1.000	1.000
	75 per cent	.041	.047	.995	.996
	50 per cent	.036	.039	.911	.919
	25 per cent	.030	.028	.660	.678
	Pairwise	.031	.056	.010	.005
Comparison	All Possible	.007	.006	.142	.149
	75 per cent	.007	.006	.142	.149
	50 per cent	.007	.006	.144	.150
	25 per cent	.008	.006	.147	.153
	Pairwise	.006	.012	.002	.001
Experiment	All Possible	.168	.147	3.557	3.721
	75 per cent	.120	.111	2.555	2.680
	50 per cent	.084	.077	1.724	1.797
	25 per cent	.047	.036	.881	.918
	Pairwise	.038	.073	.010	.005

$.05\sigma_P = .007$

Table 12. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Skewed Distribution, Unequal n's (8, 9, 11, 16), Equal σ^2's (1).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.035	.046	1.000	1.000
	75 per cent	.034	.039	.992	.994
	50 per cent	.031	.031	.913	.918
	25 per cent	.023	.022	.642	.661
	Pairwise	.026	.046	.021	.006
Comparison	All Possible	.006	.005	.145	.150
	75 per cent	.006	.005	.145	.151
	50 per cent	.006	.006	.147	.152
	25 per cent	.006	.006	.141	.147
	Pairwise	.006	.011	.004	.001
Experiment	All Possible	.142	.130	3.616	3.757
	75 per cent	.106	.093	2.615	2.718
	50 per cent	.073	.069	1.767	1.825
	25 per cent	.034	.033	.847	.881
	Pairwise	.037	.065	.022	.007

$.05\sigma_P = .007$

Table 13. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: $\alpha = .05$, 2.6 MD, Skewed Distribution, Unequal n's (8, 9, 11, 16), Unequal σ^2's (.4, .8, 1.2, 1.6).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Experimentwise	All Possible	.020	.022	1.000	1.000
	75 per cent	.018	.022	.994	.996
	50 per cent	.016	.016	.919	.925
	25 per cent	.010	.009	.694	.709
	Pairwise	.013	.022	.018	.012
Comparison	All Possible	.003	.002	.150	.158
	75 per cent	.003	.002	.152	.159
	50 per cent	.003	.002	.148	.154
	25 per cent	.003	.002	.157	.164
	Pairwise	.003	.005	.003	.002
Experiment	All Possible	.068	.055	3.755	3.938
	75 per cent	.050	.042	2.727	2.862
	50 per cent	.029	.027	1.777	1.855
	25 per cent	.017	.014	.944	.981
	Pairwise	.019	.032	.020	.012

$.05\sigma_P = .007$

Table 14. Type I and Type II Error Rates for Scheffé's S-Method and Tukey's T-Method: α = .05, 2.6 MD, Skewed Distribution, Unequal n's (8, 9, 11, 16), Unequal σ^2's (1.6, 1.2, .8, .4).

		Empirical Estimates			
		Type I errors		Type II errors	
Error Rates	Contrasts	Scheffé	Tukey	Scheffé	Tukey
Eexperimentwise	All Possible	.074	.112	1.000	1.000
	75 per cent	.067	.101	.990	.991
	50 per cent	.063	.082	.895	.903
	25 per cent	.042	.051	.610	.631
	Pairwise	.051	.112	.036	.016
Comparison	All Possible	.012	.013	.142	.147
	75 per cent	.012	.013	.143	.149
	50 per cent	.013	.013	.141	.147
	25 per cent	.011	.013	.138	.145
	Pairwise	.012	.026	.006	.003
Experiment	All Possible	.308	.327	3.548	3.683
	75 per cent	.226	.239	2.579	2.685
	50 per cent	.158	.161	1.692	1.760
	25 per cent	.067	.077	.829	.872
	Pairwise	.071	.158	.036	.016

$.05\sigma_P$ = .007

from the skewed distribution, and Scheffé's method remains relatively more robust than Tukey's method. Only the Type I estimates are affected by the shape of the population; the Type II probabilities are similar for both the normal and skewed data.

Discussion

Petrinovich and Hardyck (1969) found the Scheffé S-method generally, and the Tukey T-method occasionally, to be conservative for the probability of a Type I error. For reasonable differences between the means, these authors also found that the empirical probability of a Type II error for the Scheffé and Tukey methods were excessively large and consequently the S and T methods lacked any substantial power for detecting reasonable mean differences.

The probability statement associated with the Scheffé multiple comparison procedure is couched in terms of all possible contrasts, while the probability statement for the Tukey method was originally intended for only pairwise contrasts (Scheffé, 1959, p. 76). The comparisons between the S and T methods and the conclusions drawn by Petrinovich and Hardyck (1969) were therefore biased due to the fact that these authors worked with only a subset of all possible contrasts, e.g., pairwise contrasts. The probability of a Type I and Type II error for the Scheffé and Tukey methods might vary with the number of contrasts computed.

The present empirical investigation therefore examined the empirical probability of a Type I and Type II error for the Scheffé and Tukey methods for all possible contrasts, 75, 50, and 25 per cent of

the contrasts and for the pairwise contrasts.

The empirical probabilities of a Type I error when setting the error rate experimentwise, varies with the number of contrasts computed. The probability of a Type I error for all possible contrasts for the Scheffé method and pairwise contrasts for the Tukey method was closer to theoretical alpha than when sampling 75, 50, and 25 per cent of all the possible contrasts. The probability of a Type I error generally decreased when sampling fewer than all of the possible contrasts but, the pattern changed when just the pairwise contrasts were sampled. For example, though Scheffé's method was generally conservative for the pairwise contrasts, the probability was nonetheless larger than the probability when only 25 per cent of the contrasts were sampled. It appears that when six pairwise contrasts are computed the likelihood of committing a Type I error is greater than when six of the all possible contrasts are randomly selected, however, both are conservative. Not only were the probabilities for the pairwise contrasts larger than the probabilities when 25 per cent of the contrasts were sampled, but the empirical probabilities found for the Tukey method were generally larger and closer to theoretical alpha than the probabilities for the S-method. Since Tukey's method was originally designed for pairwise contrasts the empirical probabilities of a Type I error that have been found in this investigation were consistent with the derivation of the T-method and with the findings of previous investigators (Petrinovich and Hardyck, 1969; Scheffé, 1953, 1959).

The Scheffé and Tukey methods share the robustness or lack of

robustness of the ANOVA F test except that the degree of departure of the empirical probabilities from the theoretical probability was most certainly also a function of the number of contrasts computed. The nonproportional normal distribution case exemplifies this point. The empirical probability of a Type I error for the S-method for all possible contrasts was found to be .147 while for the pairwise contrasts it was .113. In each case the empirical probability differed from theoretical alpha of .05 but, the magnitude of the difference was a function of the number of contrasts computed.

Of much greater importance than the noted relationship between the probability of a Type I error and the number of contrasts sampled was the large probabilities found for the Type II error. Only by increasing the mean differences and sample size was the probability of a Type II error reduced. The mean differences were increased so that if there were any differences in the probability of the experimentwise Type II error, as a function of the number of contrasts computed, such differences would be more visible. Increasing the mean differences to reduce the probability of a Type II error for just the pairwise contrasts (Petrinovich and Hardyck, 1969) does not offer the researcher a practical means for overcoming this problem. The researcher is not in the same privileged position as the investigator of statistical techniques. That is, how can the behavioral scientist manipulate the differences between his means? The behavioral scientist can only work with the mean differences that have been brought about by the manipulation of the independent variable(s).

For reasonable mean differences of .67 σ-units, a difference of 2.01 σ-units between the largest and smallest means, seven observations per cell does assure the researcher that he will reject the ANOVA null hypothesis 90 per cent of the time. Empirically, the power for the ANOVA F test in this investigation was indeed found to be approximately .90. Following a significant F value, Scheffé's and Tukey's multiple comparison procedures were calculated on the data. Referring back to Table 2, the reader can once again note the large probabilities of a Type II error for all three of the error rates. The crucial finding is that the Scheffé and Tukey methods do not retain the power for detecting reasonable mean differences as was "built into" the ANOVA F test.

Scheffé (1959, p. 71) intimates while Aitkin (1969, p. 193), states that Scheffé's and Tukey's multiple comparison procedures lack sensitivity for detecting differences because of the dependence that researchers have for the conventional five and one per cent significance levels. Also, the vagaries of psychological experimentation can cause the Scheffé and Tukey techniques to be even less sensitive in detecting mean differences. That is, psychological research is most often characterized by small to moderate mean differences, small to moderate sample sizes, large within-subjects variability, and inexact measurements of the dependent variable. Given all of the above, the psychological investigator should not be surprised to find no significant mean differences when using the Scheffé and Tukey methods even though his ANOVA F test was significant. With small mean differences, at the

traditional .05 significance level, the Scheffé and Tukey methods will commit many more Type II errors than the ANOVA F test and therefore will not be as powerful a statistic for detecting mean differences. Therefore, Scheffé's S-method and Tukey's T-method should be investigated under conditions of varied levels of significance (α ranging from .05 to .25) and/or varied conditions of sample size to determine the optimum alpha level and sample size that would be sufficient to increase the power of these methods to the power that had been originally "built into" the ANOVA F test for detecting reasonable differences in the means.

References

Aitkin, M. A. Multiple comparisons in psychological experiments. The British Journal of Mathematical and Statistical Psychology, 1969, 22, 193-198.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. I, Effect of inequality of variances in the one-way classification. Annals of Mathematical Statistics, 1954a, 25, 290-302.

Box, G. E. P. 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, 1954b, 25, 484-498.

Box, G. E. P., and Anderson, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, Series B, 1955, 17, 1-26.

Horsnell, G. The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 1953, 40, 128-136.

IBM, 1130, Scientific Subroutine Package (1130-CM-02X) Programmer's Manual, H20-0252-1, International Business Machines Corporation, 1967.

Kendall, M. A., and Stuart, Alan. The Advanced Theory of Statistics, Vol. 1. New York: Hafner Publishing Company, 1969.

Kirk, Roger, E. Experimental Design: Procedures for the Behavioral Sciences. California: Brooks/Cole Publishing Company, 1968.

Miller, Rupert, G., Jr. Simultaneous Statistical Inference. New York: McGraw-Hill Book Company, 1966.

Petrinovich, Lewis, F., and Hardyck, Curtis, D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

Ryan, Thomas, A. Multiple comparisons in psychological research. Psychological Bulletin, 1959, 56, 26-47.

Ryan, Thomas, A. The experiment as the unit for computing rates of error. Psychological Bulletin, 1962, 59, 301-305.

Scheffé, Henry. A method for judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheffé, Henry. The Analysis of Variance. New York: John Wiley & Sons, Inc., 1959.

Smith, Robert, A. The effect of unequal group size on Tukey's HSD procedure. Psychometrika, 1971, 36, 31-34.

Tukey, J. W. The problem of multiple comparisons. Unpublished manuscript, Princeton University, 1953.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1962.

APPENDICES

Appendix I

SUGGESTIONS FOR INCREASING THE POWER OF THE SCHEFFÉ STATISTIC

INTRODUCTION

Petrinovich and Hardyck (1969) investigated many multiple comparison procedures, one of which was Scheffé's multiple comparison method (S-method). Under their null hypothesis conditions, Petrinovich and Hardyck found that the Scheffé method does generally control the probability of a Type I error as it is designed to do, but the S-method seems to overprotect. That is, the empirical probabilities were less than theoretical alpha. Keselman and Toothaker (1971) have found that the conservative probabilities of a Type I error found for the Scheffé method can be attributed to the fact that Petrinovich and Hardyck limited their investigation to just pairwise contrasts, e.g., $\hat{\psi} = (+1)\ \overline{X}._k + (-1)\ \overline{X}._{k'}$. According to the empirical estimates found by Keselman and Toothaker (1971), the empirical probability of a Type I error appears to be related to the number of contrasts computed. The latter authors found that the empirical probability of a Type I error was closest to theoretical alpha when all of the possible contrasts were computed. The empirical probabilities decreased when only 75, 50, and 25 per cent of the all possible contrasts were computed. Theugh the Type I estimates for the pairwise contrasts were generally conservative, the values were always larger than the probabilities when 25 per cent of the

contrasts were computed, even though both subsets contained six contrasts.

Of greater concern to Keselman and Toothaker (1971) was that the empirical probability of a Type II error for the Scheffé method was extremely large. These authors point out that the S-method lacks substantial power for detecting reasonable mean differences. For reasonable .67 σ-unit differences, a difference of 2.01 σ-units between the largest and smallest means in their investigation, Keselman and Toothaker (1971) discovered that the Scheffé method was not nearly as powerful as was the ANOVA F test which preceded the Scheffé multiple comparison method.

That is, with seven observations per cell, for a one-way fixed effects analysis of variance (ANOVA) with four treatment levels, the ANOVA null hypothesis should be rejected ninety per cent of the time with the above mean differences. Following a significant F value, the Scheffé method was then computed for all possible of the contrasts, 75, 50, and 25 per cent of the contrasts and for the pairwise contrasts. Keselman and Toothaker (1971) found that, for the Scheffé method, the empirical probabilities of a Type II error were excessively large for the three error rates they investigated; therefore, the S-method lacked any substantial power for detecting mean differences regardless of whether the errors were counted with an experimentwise, per comparison, or per experiment error rate (Ryan, 1959, 1962). For example, Keselman and Toothaker calculated that for seven observations per cell the ANOVA F test should detect .67 σ-unit differences approximately 90 per cent of the time

the ANOVA F statistic is computed; consequently, the probability of a Type II error would be approximately .10. Following a significant ANOVA, the power of Scheffé's method for detecting the mean differences was approximately zero for the experimentwise error rate and approximately .33 with the per comparison rate. Therefore, the empirical probability of a Type II experimentwise error was one and approximately .67 when the errors were counted with a per comparison rule. Miller (1966, p. 32), emphasizes this point by stating that multiple comparison procedures were designed to control the probability of a Type I error and not the probability of a Type II error.

Two reasonable approaches to increase the power of the Scheffé method would be, to increase the level of significance for the Scheffé test and/or increase the sample size per cell. Many readers may feel that increasing alpha and increasing the number of observations per cell are not reasonable approaches. We are cognizant that for increases in sample size the power for the ANOVA F test could be inflated to the point that perhaps meaningless differences would be detected. On the other hand, sample sizes larger than those that would assure a power of $\simeq .90$ for the ANOVA F test would be essential in order to detect true differences in the means with multiple comparison procedures (Keselman and Toothaker, 1971; Petrinovich and Hardyck, 1969). Since the ANOVA F test is only a barometer of overall treatment effects, while multiple comparison procedures are designed to fetter out the exact mean differences, perhaps the choice of the number of observations per cell should be geared to finding multiple comparison differences, not overall ANOVA

differences. The ANOVA F test should, therefore, be considered as a preliminary first step to the more important step of computing multiple comparisons (Gabriel, 1964, p. 472; Scheffé, 1959, p. 71).

The second approach for increasing the power of the S-method would be to increase the alpha level. For the alpha purists who are stuck at the .01 and .05 significance levels, the idea of increasing the level of significance is also shared by others (Aitkin, 1969, p. 193; Scheffé, 1959, p. 71).

Increasing the mean differences, as Petrinovich and Hardyck (1969) had done, is an unrealistic procedure for increasing the power of the Scheffé method. The behavioral scientist is not in the same privileged position as the researcher of statistical methods in that the behavioral scientist must work with the mean differences that have been brought about by his independent variable(s) and cannot build differences a priori into his data. Also as Games (1971) points out, for large σ-unit differences the means would obviously be different, so different that it would not be necessary for a researcher to perform a statistical multiple comparison test under conditions in which the means were indeed so divergent.

Therefore, the first major phase of this investigation manipulated the level of significance (.05, .10, ..., .25) for the Scheffé test and incremented the number of observations per cell to determine the optimum alpha level and sample size that would be sufficient to substantially increase the power of Scheffé's S-method, from the low level reported when alpha was set at the traditional .05 level and when there were seven observations per cell (Keselman and

Toothaker, 1971).

After determining the optimum alpha level and sample size, all possible of the contrasts, 75, 50, and 25 per cent of the contrasts, and all pairwise contrasts were computed to determine the empirical probabilities of a Type I and Type II error. The empirical probabilities were examined when the assumptions of the Scheffé test were met and under conditions of assumption violations to determine whether the pattern of relationship between the empirical probabilities and the number of contrasts computed, found by Keselman and Toothaker (1971), would also be found for large alpha and for larger sample sizes.

Procedure

Pseudo-random numbers were selected, using a pseudo random number generator. Depending upon the assumption violation, the numbers were selected from either a normal or skewed distribution with $\mu = 0$ and $\sigma = 1$. The random numbers were distributed to the four treatment levels that comprise a one-way fixed effects analysis of variance. Mean differences (differences between adjoining means expressed in standard deviation units) were set at .75 σ-unit differences. Therefore, $\mu_1 = 0$, $\mu_2 = .75$, $\mu_3 = 1.50$, and $\mu_4 = 2.25$.

The observations from the normal distribution were generated by means of GAUSS (IBM, 1967), which generates pseudo-random deviates with $\mu = 0$ and $\sigma = 1$. The skewed population was derived from a chi-square distribution with three degrees of freedom, having mean three, variance six, third moment 24, fourth moment 252, skewness measure $\gamma_1 = \left(\mu_3^2/\mu_2^3\right)^{1/2} = 1.663$ and kurtosis measure

$\gamma_2 = \mu_4 / \mu_2^2 - 3. = 4$ (Kendall and Stuart, 1969).

Pseudo random chi-square variables with three degrees of freedom were generated by summing the squares of three N(0, 1) variables. These numbers were then scaled so that the mean and variance of the skewed population would be the same as the mean and variance of the normal population, first by subtracting three from each score and then multiplying by $1/\sigma$ where $\sigma^2 = 6$. The resulting skewed population has mean zero, variance one, skewness measure $\gamma_1 = 1.663$ and kurtosis measure $\gamma_2 = 4$, as γ_1 and γ_2 are invariant under additive and multiplicative transformations.

Since population mean values were prespecified, determining $\sum_k^K \alpha_k^2$ was straightforward and consequently, the sample size could be calculated such that the power was at least .90 for detecting .75 σ-unit differences with the ANOVA $\underline{F}$ test.

For any one sampling from the random number generator, a one-way fixed effects ANOVA was calculated. If the $\underline{F}$ value equalled or exceeded the critical .05 $\underline{F}$ value, the Scheffé S-method was initiated. When the obtained $\underline{F}$ value failed to reach significance, the program returned to the sampling procedure and an ANOVA test was once again performed. For any given set of data in which the null hypothesis of equal mean values was rejected, Scheffé's multiple comparison statistic was computed on the significant data.

Once the program passed from the ANOVA to the stage of computing the S-method, the levels of significance for the Scheffé test, not the $\underline{F}$ test, were manipulated. That is, the Scheffé test statistic uses a value from the sampling distribution of $\underline{F}$, e.g., $\left[(J-1)\ {}_{\alpha}F_{\nu_1\nu_2}\right]^{\frac{1}{2}}$. This critical $\underline{F}$ value is determined by three

values: (1) the numerator degrees of freedom (ν_1) from the ANOVA test, (2) the denominator degrees of freedom (ν_2) from the ANOVA test and, (3) the probability of a Type I error that the experimenter sets a priori. In addition to setting the Scheffé test statistic at the traditional .05 level, i.e., $\left[(J-1)\ {}_{.05}F_{\nu_1\nu_2}\right]^{\frac{1}{2}}$, the Scheffé statistic was also set at nontraditional levels of significance (.10, .15, ..., .25). That is, the overall ANOVA F test was calculated with $\alpha = .05$ but, the α chosen for the Scheffé procedure ranged from .05 to .25.

Also manipulated was the sample size per cell. The original sample size of seven observations per cell was increased 75 per cent (12 observations per cell), 150 per cent (17 observations per cell), and 200 per cent (21 observations per cell).

For each set of contrasts, all possible, 75 per cent, 50 per cent, 25 per cent, and all pairwise contrasts, Scheffé's S-method was calculated to determine the number of contrasts which did or did not exceed the S critical value for the statistic, e.g., $\left[(J-1)_{\alpha}F_{\nu_1,\nu_2}\right]^{\frac{1}{2}}\left[MS_{\varepsilon}\sum^{K}\left(c_1^2/n_1 + \ldots + c_k^2/n_k\right)\right]^{\frac{1}{2}}$. The procedure of generating random samples (K=4) with n_k observations per cell and thereafter, if the F test was significant, calculating the S-method constituted one single experiment; the procedure was repeated for 1000 experiments.

Unequal variances and unequal sample sizes were combined when sampling from a normal distribution to explore the two types of error of Scheffé's method under conditions of assumption violations. The five conditions examined when sampling from a normal distribution

were: (1) equal observations per treatment level - equal variances, (2) equal observations per treatment level - unequal variances, (3) unequal observations per treatment level - equal variances, (4) unequal observations per treatment level - unequal variances (proportionately paired), and (5) unequal observations per treatment level - unequal variances (inversely proportionately paired). These five conditions were also investigated for the non-normal skewed distribution.

The first criterion for selecting the unequal sample sizes was to get ϕ (non centrality parameter for the non-central F distribution as close to the value of ϕ for the equal sample case. The second criterion was to have the sample sizes divergent enough to be interesting as a possible assumption violation.

For comparisons involving unequal variances, the variances were specified to be in the ratio of 1:2:3:4. A further qualifier placed upon the choice of values for the variances was that the average of the variances should equal one, thereby not differentially affecting the original calculations of sample size for a desired power, which was calculated for the ANOVA F test.

Results

Normal Distribution, Equal n's (7), Equal σ^2's (1):

The empirical probabilities of a Type II error for Scheffé's multiple comparison method are contained within Table 1. The probabilities are tabled for the experimentwise, per comparison, and per experiment error rates with varied alpha levels (.05, .10, ... , .25). At the traditional .05 significance level, the empirical

Table 1. Type II Error Rates for Scheffé's S-Method: .75 MD, Normal Distribution, Equal n's (7), Equal σ^2's (1).

		Alpha Levels of Significance												
		.05	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20	.25
Error Rates	Contrasts													
Experimentwise	All Poss.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	75%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	25%	1.00	.99	.99	.98	.99	.99	.99	.99	.98	.99	.97	.98	.97
	Pairwise	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Comparison	All Poss.	.64	.56	.55	.54	.53	.52	.51	.50	.49	.49	.48	.47	.44
	75%	.64	.56	.55	.54	.53	.52	.51	.50	.50	.48	.48	.47	.44
	50%	.64	.56	.55	.53	.53	.53	.50	.50	.50	.49	.48	.47	.44
	25%	.62	.56	.55	.53	.53	.52	.52	.51	.50	.49	.48	.47	.44
	Pairwise	.65	.57	.56	.54	.53	.53	.51	.50	.49	.48	.48	.47	.43
Experiment	All Poss.	15.90	13.95	13.65	13.37	13.18	13.02	12.71	12.50	12.36	12.16	12.04	11.79	11.02
	75%	11.45	10.04	9.84	9.67	9.54	9.38	9.14	8.93	8.90	8.68	8.70	8.47	7.90
	50%	7.64	6.71	6.60	6.41	6.34	6.31	6.04	6.00	5.94	5.83	5.77	5.68	5.32
	25%	3.74	3.35	3.30	3.17	3.21	3.15	3.09	3.06	3.00	2.96	2.86	2.84	2.66
	Pairwise	3.90	3.42	3.33	3.24	3.21	3.18	3.09	3.02	2.96	2.88	2.87	2.84	2.57

probabilities of a Type II error are excessive for the three error rates.

When the error rate is set experimentwise, the likelihood that the Scheffé method would commit a Type II error for all possible comparisons is 100 per cent with .75 σ-unit differences in the means. For the per comparison error rate, as with the experimentwise error rate, the probabilities are very large. For all possible contrasts with $\alpha = .05$, the empirical probability of a Type II error is .64, while for the pairwise contrasts the probability is .65. The long run average for committing a Type II error, the per experiment error rate, is also large.

Table 1 indicates that the empirical probabilities of a Type II error for the S-method can be decreased by increasing the alpha level. For the largest alpha investigated, .25, the probabilities decrease, but the probability of a Type II error is still in excess of the probability of a Type II error for the ANOVA $\underline{F}$ test, ≈ .10. Increasing the alpha level is not sufficient to reduce the probability of a Type II error to a level where the power of the Scheffé test would compare favorably to the power that the ANOVA $\underline{F}$ test would have for detecting the same .75 σ-unit differences.

For seven observations per cell, the probability of a Type II experimentwise error is still about one with alpha set at .25. The probability of a Type II error is reduced for the per comparison error rate but the probabilities are still excessive; there is approximately a 44 per cent chance of committing a Type II error. The power

therefore, of detecting .75 σ-unit mean differences, when counting with a per comparison rule, is approximately .56, even though the significance level had been set at $\alpha = .25$ for the Scheffé test. Consequently, it would appear that not only must the alpha level be increased but sample size must as well be increased to reduce the probability of a Type II error for the Scheffé multiple comparison statistic.

Tables 2-4 contain the empirical probabilities of a Type II error for the Scheffé method when the original sample size of seven observations per cell is incremented 75, 150, and 200 per cent. Counting errors with an experimentwise rule results in very large Type II probabilities for the Scheffé S-method even with large n_k and large alpha. Glancing from Table 1 to Table 4, the reader should note that the probability of a Type II error is not substantially affected for the experimentwise error rate. When there are seven observations per cell and alpha is set at .05 the probabilities for all possible of the contrasts, 75, 50, 25 per cent and for the pairwise contrasts are 1.00, 1.00, 1.00, 1.00, and 1.00, respectively. Incrementing the sample size to 21 observations per cell and the alpha level to .25, the probabilities are 1.00, 1.00, .99, .86, and .81; certainly not much of a change. As Miller (1966, p. 32) states, multiple comparison techniques that count errors with an experimentwise rate are appropriate for controlling only Type I errors and would therefore be most apropos for those researchers who want to increase the protection of their null hypothesis. Correspondingly, the experimentwise error rate appears to be very inappropriate when

Table 2. Type II Error Rates for Scheffé's S-Method: .75 MD, Normal Distribution, Equal n's (12), Equal σ^2's (1).

		Alpha Levels of Significance												
Error Rates	Contrasts	.05	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20	.25
Experimentwise	All Poss.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	75%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00	.99
	25%	.99	.97	.97	.96	.96	.97	.96	.95	.94	.94	.94	.95	.93
	Pairwise	1.00	1.00	1.00	1.00	1.00	1.00	.99	.99	1.00	.99	.99	.99	.98
Comparison	All Poss.	.49	.44	.42	.42	.41	.41	.40	.39	.38	.38	.38	.37	.34
	75%	.49	.44	.42	.42	.41	.41	.40	.38	.38	.39	.38	.37	.34
	50%	.49	.44	.43	.42	.41	.40	.40	.39	.38	.38	.38	.36	.34
	25%	.48	.44	.43	.42	.39	.41	.40	.38	.38	.38	.38	.37	.34
	Pairwise	.50	.43	.42	.41	.39	.40	.38	.36	.36	.36	.35	.34	.31
Experiment	All Poss.	12.20	10.90	10.60	10.45	10.15	10.17	10.00	9.68	9.46	9.53	9.45	9.20	8.61
	75%	8.78	7.87	7.66	7.54	7.34	7.31	7.22	6.94	6.89	6.94	6.75	6.60	6.21
	50%	5.84	5.22	5.12	5.02	4.90	4.85	4.82	4.64	4.55	4.60	4.59	4.35	4.11
	25%	2.91	2.62	2.56	2.54	2.33	2.45	2.37	2.30	2.25	2.28	2.26	2.20	2.06
	Pairwise	2.99	2.55	2.49	2.45	2.37	2.38	2.27	2.16	2.14	2.16	2.11	2.04	1.86

Table 3. Type II Error Rates for Scheffè's S-Method: .75 MD, Normal Distribution, Equal n's (17), Equal σ^2's (1).

		Alpha Levels of Significance												
Error Rates	Contrasts	.05	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20	.25
Experimentwise	All Poss.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	75%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	50%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	.99
	25%	.97	.95	.94	.94	.94	.93	.92	.91	.92	.92	.91	.91	.90
	Pairwise	1.00	.99	.99	.98	.98	.97	.97	.97	.96	.95	.95	.95	.91
Comparison	All Poss.	.40	.36	.36	.35	.35	.34	.34	.33	.33	.32	.31	.31	.30
	75%	.40	.36	.36	.35	.34	.34	.33	.33	.32	.32	.31	.31	.29
	50%	.40	.36	.36	.34	.35	.34	.34	.32	.32	.32	.32	.31	.29
	25%	.41	.37	.36	.35	.35	.34	.33	.33	.32	.32	.31	.31	.30
	Pairwise	.39	.33	.32	.31	.30	.29	.29	.28	.28	.26	.26	.26	.22
Experiment	All Poss.	10.01	9.08	8.88	8.64	8.64	8.50	8.39	8.19	8.14	7.88	7.82	7.86	7.37
	75%	7.15	6.53	6.43	6.24	6.22	6.08	5.96	5.90	5.82	5.71	5.59	5.64	5.29
	50%	4.83	4.38	4.28	4.14	4.15	4.12	4.02	3.87	3.87	3.82	3.80	3.76	3.51
	25%	2.46	2.24	2.14	2.08	2.09	2.04	2.00	1.98	1.93	1.89	1.86	1.88	1.80
	Pairwise	2.35	2.00	1.92	1.83	1.82	1.76	1.71	1.69	1.67	1.57	1.54	1.54	1.35

Table 4. Type II Error Rates for Scheffé's S-Method: .75 MD, Normal Distribution, Equal n's (21), Equal σ^2's (1).

		Alpha Levels of Significance												
Error Rates	Contrasts	.05	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20	.25
Experimentwise	All Poss.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	75%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	50%	1.00	1.00	.99	.99	.99	.99	1.00	.99	.99	.99	1.00	1.00	.98
	25%	.94	.91	.92	.91	.92	.89	.90	.89	.90	.88	.88	.88	.86
	Pairwise	.99	.97	.95	.95	.95	.94	.91	.92	.91	.89	.88	.88	.81
Comparison	All Poss.	.37	.32	.32	.32	.31	.30	.30	.30	.29	.29	.28	.28	.26
	75%	.37	.32	.32	.32	.31	.30	.30	.30	.29	.29	.29	.28	.26
	50%	.37	.32	.32	.32	.31	.30	.30	.30	.28	.29	.28	.28	.26
	25%	.37	.32	.32	.32	.31	.30	.30	.29	.29	.29	.28	.28	.25
	Pairwise	.35	.28	.26	.26	.25	.24	.24	.23	.22	.21	.21	.21	.17
Experiment	All Poss.	9.23	8.13	7.95	7.92	7.73	7.53	7.52	7.44	7.23	7.18	7.12	7.07	6.48
	75%	6.61	5.84	5.66	5.68	5.57	5.44	5.37	5.36	5.23	5.17	5.14	5.06	4.68
	50%	4.40	3.88	3.80	3.80	3.69	3.56	3.56	3.54	3.38	3.48	3.39	3.40	3.08
	25%	2.21	1.92	1.89	1.89	1.87	1.77	1.77	1.77	1.75	1.73	1.70	1.70	1.50
	Pairwise	2.08	1.65	1.55	1.55	1.48	1.44	1.42	1.39	1.32	1.28	1.24	1.24	1.04

the alternative hypothesis is true. That is, when the alternative hypothesis is true, the probability of a Type II error is very high with an experimentwise error rate.

The empirical probabilities of a Type II error for the per comparison error rate are reduced from .64 for all possible contrasts and .65 for the pairwise contrasts, when there are seven observations per cell and alpha is set at .05, to, .26 for all possible contrasts and .17 for pairwise contrasts, with 21 observations per cell and alpha set at .25. The empirical probabilities of a Type II error indicate that for the per comparison error rate, the power of Scheffé's S-method can be increased substantially by increasing the level of significance and the sample size per cell. The power of Scheffé's method, when the Type II empirical estimates are counted with a per comparison rule, is graphed in Figures 1-4. The four graphs vividly illustrate the effects of sample size and varied alpha levels on the power of the S-method. Also, for the small sample condition, (seven observations per cell), the S-method has greater power when all of the possible contrasts are computed than when only the pairwise contrasts are computed. Increasing the number of observations per cell caused a reversal of the above stated finding. That is, for the larger sample conditions, 12, 17, and 21 observations per cell, the S-statistic is more powerful for the pairwise contrasts. Consequently, when the Type II errors are counted with a per comparison rule, the power of the Scheffé method for detecting pairwise contrasts can be increased by increasing the number of observations per cell.

The per experiment error rate is likewise amenable to changes

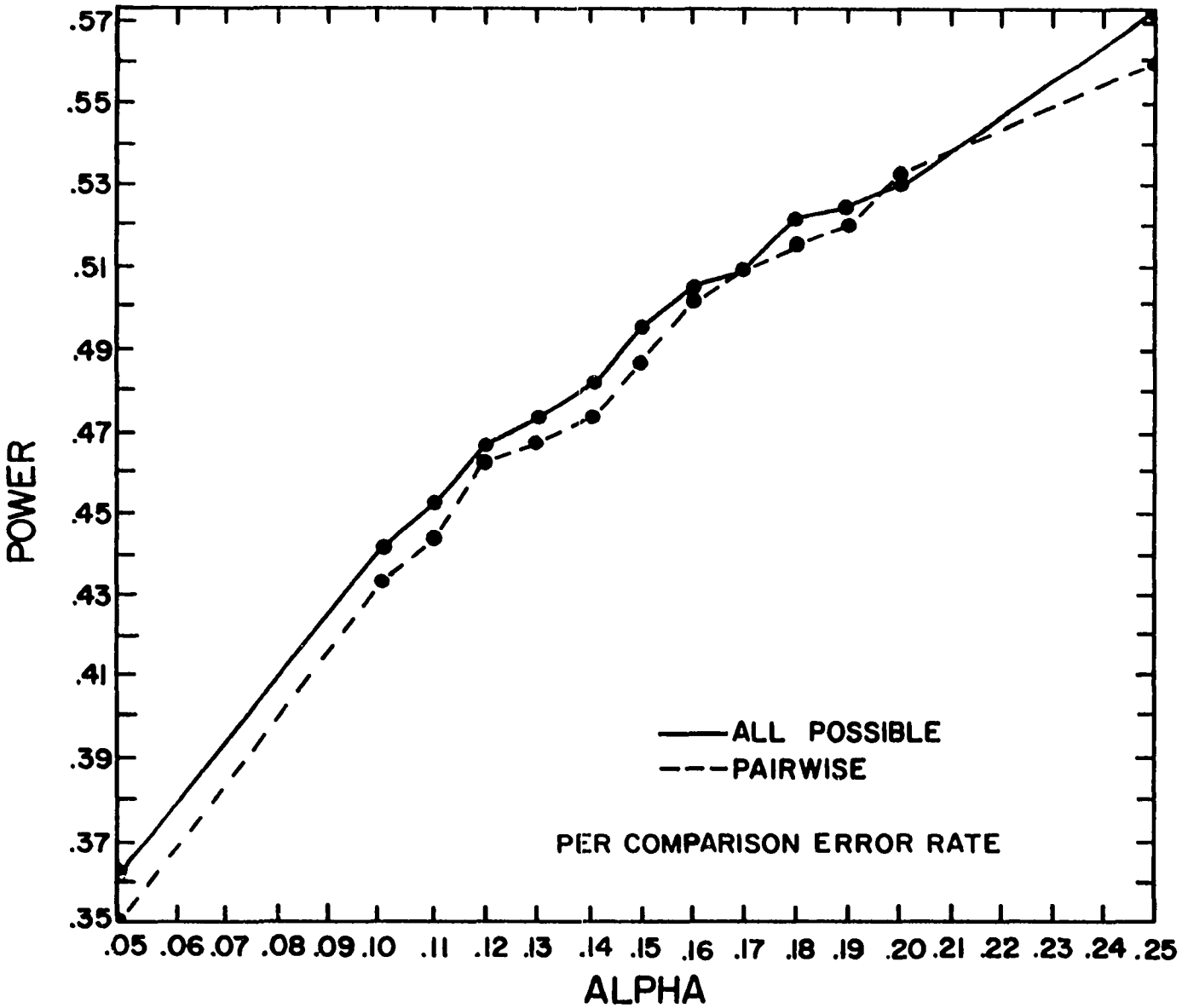


Fig.1 Scheffé's S-method: .75 M.D. Normal Distribution, Equal σ^2's(1), Equal n's(7)

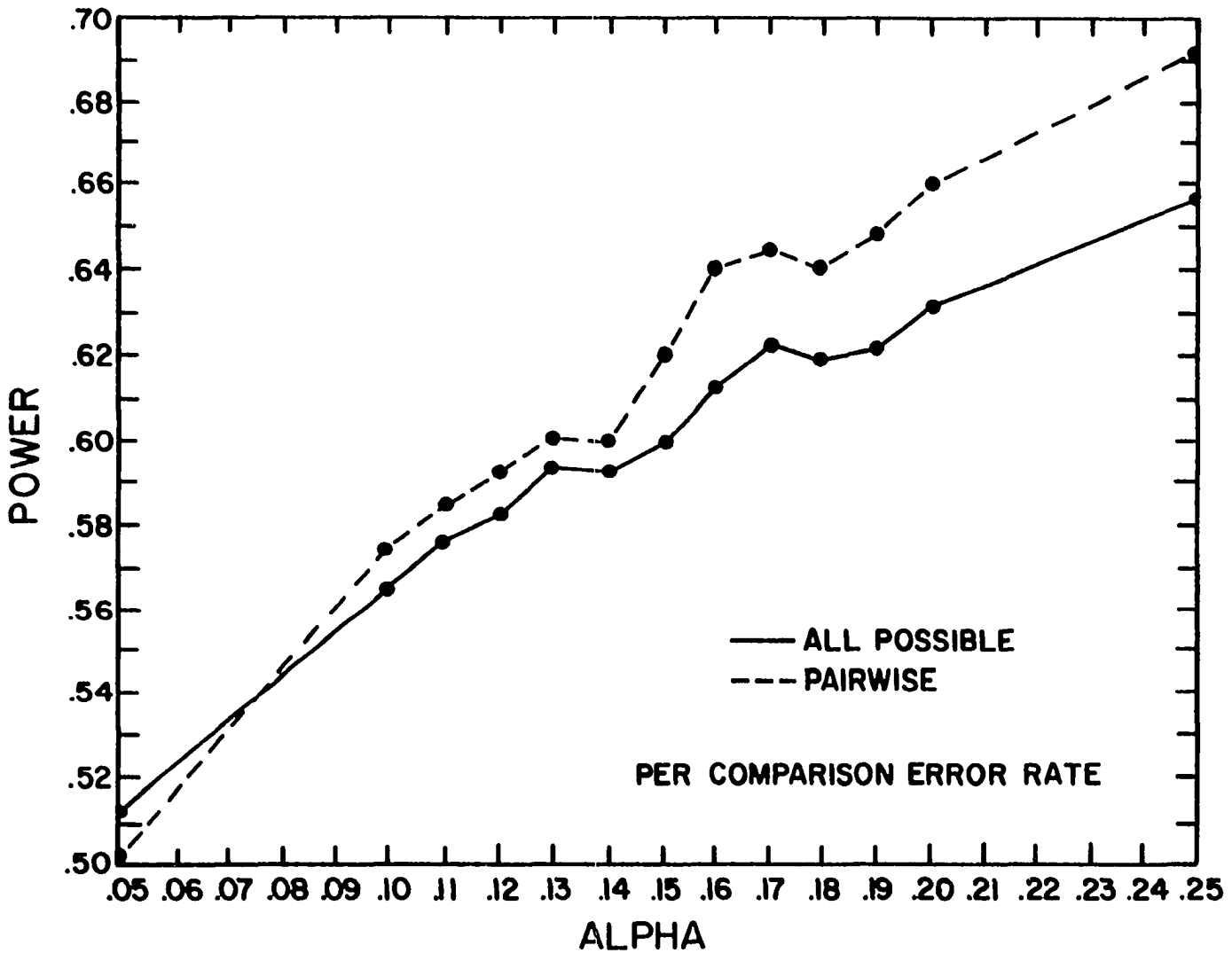


Fig.2 Scheffé's S-method: .75 M.D. Normal Distribution, Equal σ^2's(1), Equal n's(12)

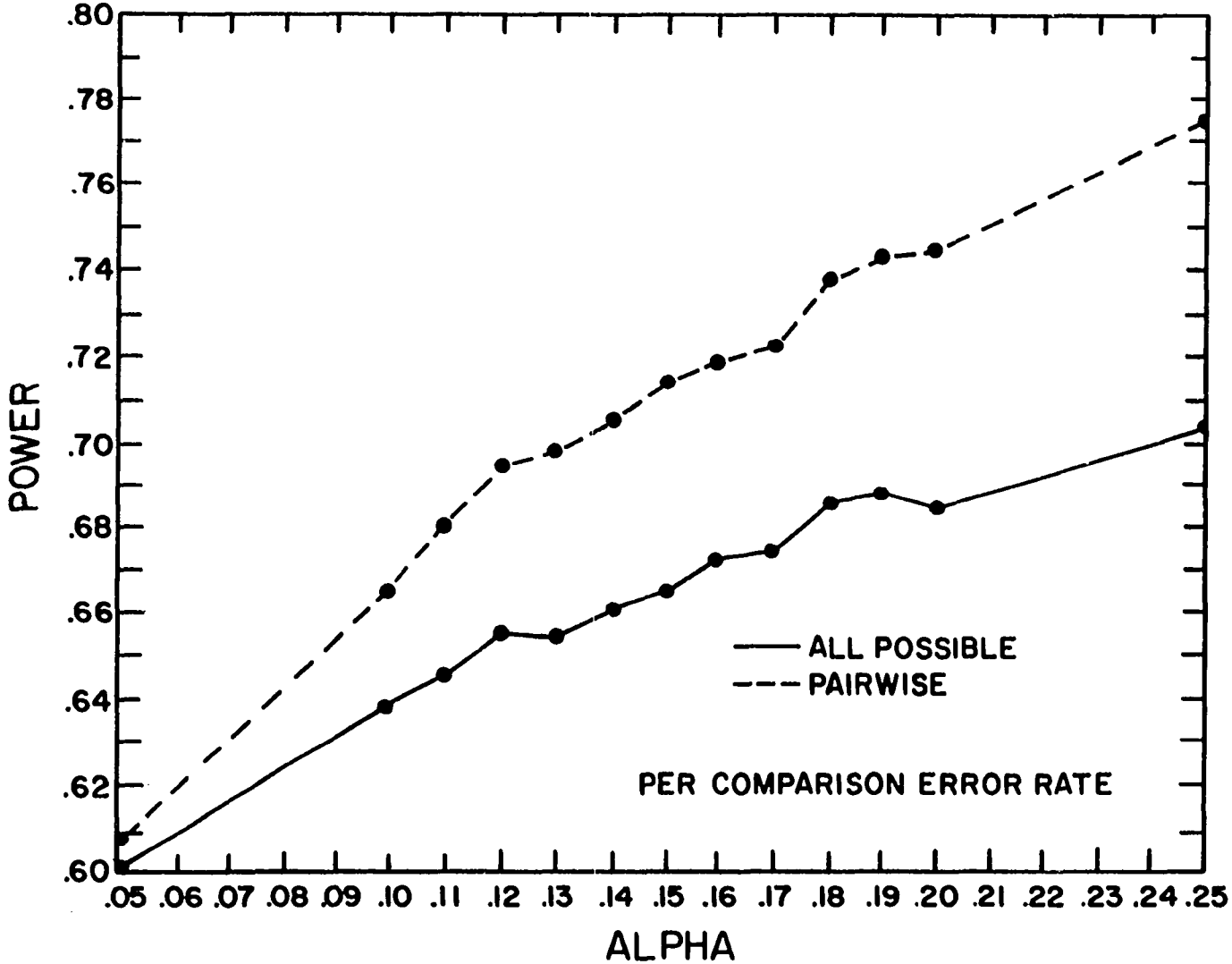


Fig.3 Scheffé's S-method: .75 M.D. Normal Distribution, Equal σ^2's (1), Equal n's (17)

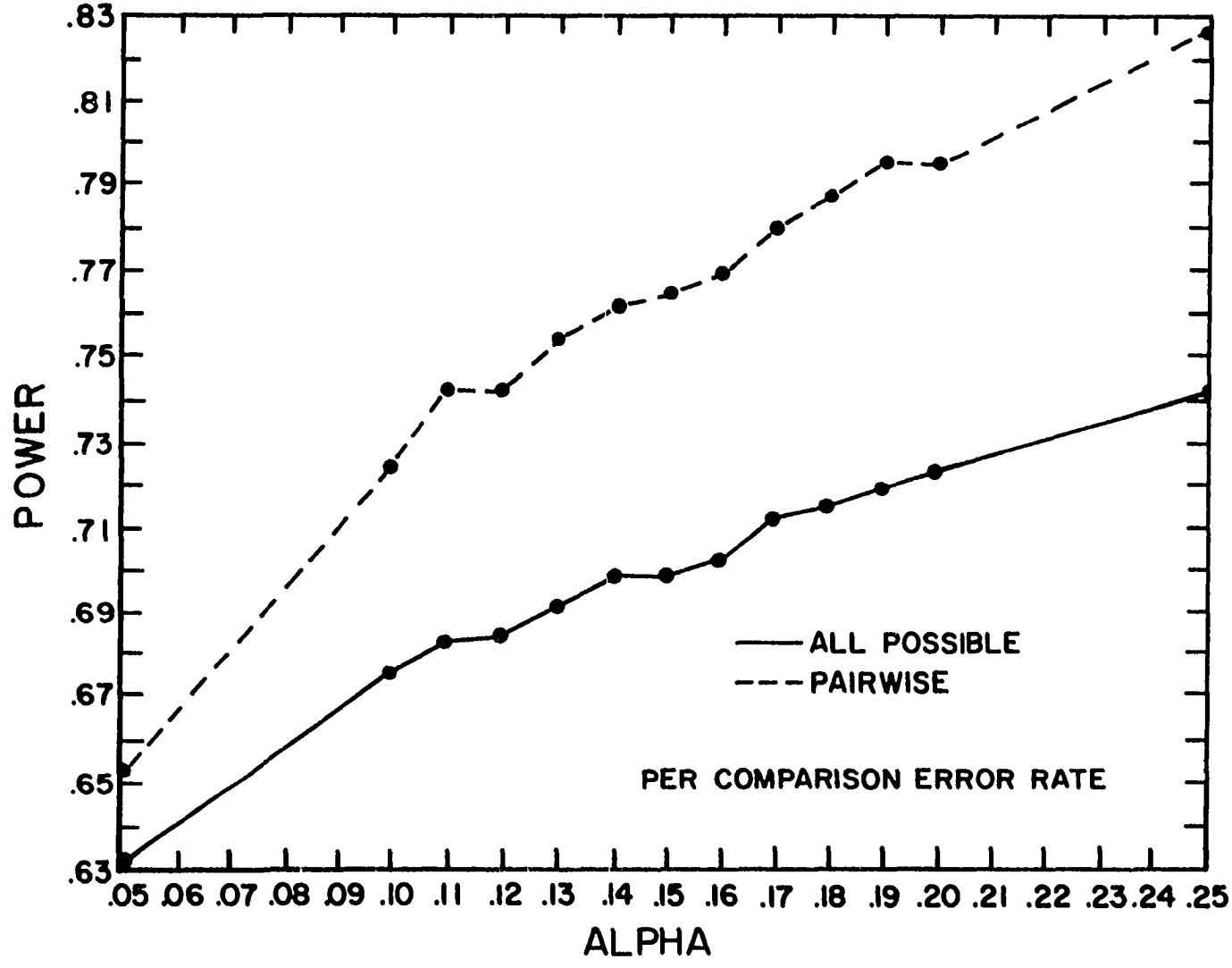


Fig. 4 Scheffé's S-method: .75 M.D. Normal Distribution, Equal σ^2's(1), Equal n's(21)

in the significance level and the sample size. The long run average of committing a Type II error decreases when the sample size and alpha level are incremented.

Increasing the sample size per cell and increasing the level of significance, do appreciably reduce the probability of a Type II per comparison error and the long run average of the per experiment Type II error. These two approaches for increasing the power of the Scheffé test are therefore effective for a per comparison and per experiment rate, but not for the experimentwise error rate.

Since the effects of sample size and varied alpha levels on the power of the S-method have been determined, the pattern of relationship between the empirical probabilities of a Type I and Type II error for different number of contrasts computed (Keselman and Toothaker, 1971) can be evaluated with alpha set at .20 and with 21 observations per cell.

Normal Distribution, Equal n's (21), Equal σ^2's(1):

The empirical probabilities of a Type I and Type II error and the number of contrasts computed are generally related as can be seen from Table 5. When the error rate is set experimentwise, the probability of a Type I error is very close to theoretical alpha for all possible of the contrasts. The empirical probabilities decrease for each succeeding subset of contrasts sampled except for the pairwise contrasts. That is, the probability of a Type I error when 75 per cent (18) of the contrasts are randomly sampled is larger than the probability when 50 per cent (12) of the contrasts are sampled. The empirical probability of a Type I error is also larger when 50 per cent

Table 5. Type I and Type II Error Rates for Scheffé's S-Method:

α = .20, .75 MD, Normal Distribution, Equal n's (21), Equal σ^2's (1).

Error Rates	Contrasts	Empirical Estimates	
		Type I Errors	Type II Errors
Experimentwise	All Possible	.201	1.000
	75 per cent	.183	1.000
	50 per cent	.168	.996
	25 per cent	.128	.884
	Pairwise	.143	.878
Comparison	All Possible	.035	.283
	75 per cent	.034	.281
	50 per cent	.033	.284
	25 per cent	.035	.283
	Pairwise	.035	.206
Experiment	All Possible	.868	7.071
	75 per cent	.610	5.058
	50 per cent	.395	3.403
	25 per cent	.210	1.697
	Pairwise	.208	1.237

$.20\sigma_P$ = .013

of the contrasts are sampled than when 25 per cent (six) of the contrasts are sampled. The Type I estimate for the six pairwise contrasts is consistently higher than the probability when 25 per cent of the contrasts are computed. Apparently, the likelihood of committing a Type I error is greater when the six pairwise contrasts are computed than when six contrasts are randomly sampled from the all possible contrasts.

The empirical probability of a Type I error shows very little variability with the per comparison error rate. Scheffé's method is not designed to control the probability of a Type I error on a per comparison basis. The empirical probabilities reflect the probability of a Type I error for any one t test, if, for a set of multiple t tests one wanted to control the probability experimentwise at .20 (Aitkin, 1969, p. 195). That is, instead of using the S-method, one could compute multiple t tests and control the error rate experimentwise for the set of multiple t tests, by setting a conservative level of significance for each of the individual tests. The researcher though, would not know exactly what his experimentwise error rate is, unless calculated by the Aitkin procedure. Scheffé's method gives the advantage of controlling the error rate experimentwise. Therefore, the behavioral scientist would not have to second guess what the probability of an error would be for the experiment; that is, for the set of multiple comparison tests. The probability is set a priori by the experimenter when using the Scheffé method. Referring back to Table 5 the Type I empirical per comparison probabilities are fairly close to the theoretical value of .026.

The long run average of committing a Type I error, the per experiment error rate, also fluctuates with the number of contrasts computed.

For an experimentwise error rate, the empirical probability of a Type II error is excessive, regardless of the number of contrasts computed. Indeed, the probability of a Type II error in any one experiment is one hundred per cent, or close to it. The probabilities of a Type II error for the per comparison error rate, reflect the effects of increasing the level of significance and the effects of incrementing the sample size. That is, the probability of an error is reduced considerably from what had been found when there were seven observations per cell and when alpha was set at the traditional .05 level. For a per comparison rule of counting the errors, the power of the Scheffé method for detecting mean differences approaches a respectable level, ≃ .72. For the per experiment error rate, the number of Type II errors also varies with the number of contrasts computed.

The relationship between the probability of committing an error varying with the number of contrasts computed for large alpha and large sample size per cell, is consistent with the data reported when alpha was set at .05 and when the number of observations per cell was considerably less than the number used in this investigation (Keselman and Toothaker, 1971). Therefore, unless the relationship should change, this finding will not again be discussed for the following violation of assumption conditions.

The effect of the assumption violation in each condition may

be readily assessed by remembering that the empirical probability of a Type I error for Scheffé's method, for all possible contrasts, should coincide with theoretical alpha ($\alpha = .20$).

Normal Distribution, Equal n's (21), Unequal σ^2's (.4, .8, 1.2, 1.6):

The empirical probabilities of a Type I and Type II error when sampling from populations with unequal variances are different from the data reported when the populations have equal variances. Table 6 contains the probability estimates. Variance heterogeneity does not cause the probability of a Type I error to substantially deviate from theoretical alpha for all possible of the contrasts with the experimentwise error rate, if there are an equal number of observations per cell. The correspondence between the empirical probability and the theoretical probability is restricted to all possible contrasts. For any number less than all possible of the contrasts, the empirical probability of a Type I error is less than theoretical alpha, i.e., Scheffé's method is a conservative test.

The empirical probabilities of a Type II error for the per comparison error rate are somewhat higher than the values when there are no assumption violations; the Scheffé test therefore, loses some of its power for detecting mean differences when the variances are not equal.

Normal Distribution, Unequal n's (17, 19, 22, 26), Equal σ^2's (1):

Table 7 contains the Type I and Type II empirical probabilities for the unequal n's case. The data from Table 7 does differ from the

Table 6. Type I and Type II Error Rates for Scheffé's S-Method: α = .20, .75 MD, Normal Distribution, Equal n's (21), Unequal σ^2's (.4, .8, 1.2, 1.6).

Error Rates	Contrasts	Empirical Estimates	
		Type I Errors	Type II Errors
Experimentwise	All Possible	.191	1.000
	75 per cent	.175	1.000
	50 per cent	.173	.997
	25 per cent	.137	.902
	Pairwise	.151	.933
Comparison	All Possible	.038	.310
	75 per cent	.038	.308
	50 per cent	.040	.309
	25 per cent	.040	.314
	Pairwise	.040	.249
Experiment	All Possible	.954	7.741
	75 per cent	.691	5.546
	50 per cent	.481	3.706
	25 per cent	.243	1.885
	Pairwise	.237	1.496

$.20\sigma_P$ = .0.13

Table 7. Type I and Type II Error Rates for Scheffé's S-Method: α = .20, .75 MD, Normal Distribution, Unequal n's (17, 19, 22, 26), Equal σ^2's (1).

		Empirical Estimates	
Error Rates	Contrasts	Type I Errors	Type II Errors
Experimentwise	All Possible	.177	1.000
	75 per cent	.169	1.000
	50 per cent	.154	.990
	25 per cent	.116	.877
	Pairwise	.133	.868
Comparison	All Possible	.031	.280
	75 per cent	.030	.277
	50 per cent	.032	.280
	25 per cent	.033	.280
	Pairwise	.031	.200
Experiment	All Possible	.771	7.013
	75 per cent	.548	4.980
	50 per cent	.380	3.359
	25 per cent	.196	1.681
	Pairwise	.184	1.201

$.20\sigma_P = .013$

data that was found for the equal n's case; the Type I estimates are slightly more conservative. Consequently, the Scheffé method is to an extent affected by an unequal number of observations per cell and the probability of a Type I error changes according to the number of contrasts computed.

Normal Distribution, Unequal n's (17, 19, 22, 26), Unequal σ^2's (.4, .8, 1.2, 1.6):

For this assumption violation, unequal sample sizes and unequal variances have been combined proportionately. That is, the smallest sample was paired with the population with the largest variance. As has been found with the ANOVA F test (Box, 1954a, b; Box and Anderson, 1955; Horsnell, 1953) for this form of assumption violation, the empirical probability of a Type I error is considerably less than theoretical alpha. The Scheffé method is conservative with the probability of a Type I error but, the magnitude of the disparity between the empirical probabilities and the theoretical value is again a function of the number of contrasts computed. The empirical probabilities of a Type I and Type II error, for the three error rates, are contained within Table 8. The empirical probabilities of a Type II error are larger than the probabilities when there are no assumption violations. The Scheffé method therefore, loses some of its power for detecting mean differences when proportionately pairing unequal variances and unequal sample sizes.

Normal Distribution, Unequal n's (17, 19, 22, 26), Unequal σ^2's (1.6, 1.2, .8, .4):

The empirical probabilities of a Type I and Type II error when

Table 8. Type I and Type II Error Rates for Scheffé's S-Method: $\alpha = .20$, .75 MD, Normal Distribution, Unequal n's (17, 19, 22, 26), Unequal σ^2's (.4, .8, 1.2, 1.6).

		Empirical Estimates	
Error Rates	Contrasts	Type I Errors	Type II Errors
Experimentwise	All Possible	.130	1.000
	75 per cent	.119	1.000
	50 per cent	.100	.995
	25 per cent	.072	.931
	Pairwise	.090	.974
Comparison	All Possible	.020	.328
	75 per cent	.020	.325
	50 per cent	.020	.328
	25 per cent	.016	.330
	Pairwise	.020	.290
Experiment	All Possible	.500	8.197
	75 per cent	.354	5.858
	50 per cent	.242	3.930
	25 per cent	.099	1.977
	Pairwise	.119	1.739

$.20\sigma_P = .013 = .013$

inversely pairing unequal variances and unequal sample sizes are enumerated in Table 9. The Type I estimates are considerably larger than when there were no assumption violations. The magnitude of discrepancy between the empirical estimates and the theoretical value, again, varies with the number of contrasts computed. The Type II estimates, on the other hand, do not substantially differ from the data found when there are no assumption violations.

Skewed Distribution:

The empirical probabilities for the five conditions examined when sampling from the skewed distribution are contained within Tables 10-14. The probabilities found when sampling from the skewed distribution differ quite often from the data when the distribution is normal in shape.

When the population variances are equal and the number of observations per treatment level are also equal, the empirical probability of a Type I experimentwise error is less, and also more deviant from theoretical alpha when sampling from the skewed distribution (Table 10). The Type II estimates are basically the same regardless of population shape.

Another instance in which the empirical probabilities differ due to the shape of the distribution, is when the population variances are unequal (Table 11). The probability of a Type I experimentwise error is less than the probability when sampling is from a normal distribution. Again, the probabilities are farther from theoretical alpha. The Type I estimates for the per comparison and per experiment error rates, on the other hand, do not differ substantially

Table 9. Type I and Type II Error Rates for Scheffé's S-Method: α = .20, .75 MD, Normal Distribution, Unequal n's (17, 19, 22, 26), Unequal σ^2's (1.6, 1.2, .8, .4).

Error Rates	Contrasts	Empirical Estimates: Type I Errors	Empirical Estimates: Type II Errors
Experimentwise	All Possible	.286	1.000
	75 per cent	.269	1.000
	50 per cent	.250	.991
	25 per cent	.215	.875
	Pairwise	.229	.877
Comparison	All Possible	.068	.290
	75 per cent	.069	.287
	50 per cent	.068	.290
	25 per cent	.072	.297
	Pairwise	.068	.211
Experiment	All Possible	1.713	7.252
	75 per cent	1.238	5.159
	50 per cent	.814	3.474
	25 per cent	.430	1.783
	Pairwise	.411	1.266

$.20\sigma_P$ = .013

Table 10. Type I and Type II Error Rates for Scheffé's S-Method: σ = .20, .75 MD, Skewed Distribution, Equal n's (21), Equal σ^2's (1).

Error Rates	Contrasts	Empirical Estimates: Type I Errors	Empirical Estimates: Type II Errors
Experimentwise	All Possible	.177	1.000
	75 per cent	.169	.999
	50 per cent	.155	.988
	25 per cent	.118	.861
	Pairwise	.143	.817
Comparison	All Possible	.034	.275
	75 per cent	.034	.270
	50 per cent	.033	.271
	25 per cent	.034	.275
	Pairwise	.034	.186
Experiment	All Possible	.847	6.871
	75 per cent	.610	4.868
	50 per cent	.396	3.250
	25 per cent	.207	1.652
	Pairwise	.207	1.118

$.20\sigma_P = .013$

Table 11. Type I and Type II Error Rates for Scheffé's S-Method: α = .20, .75 MD, Skewed Distribution, Equal n's (21), Unequal σ^2's (.4, .8, 1.2, 1.6).

Error Rates	Contrasts	Empirical Estimates	
		Type I Errors	Type II Errors
Experimentwise	All Possible	.185	1.000
	75 per cent	.170	1.000
	50 per cent	.155	.989
	25 per cent	.126	.875
	Pairwise	.137	.854
Comparison	All Possible	.035	.272
	75 per cent	.034	.272
	50 per cent	.036	.269
	25 per cent	.035	.275
	Pairwise	.033	.186
Experiment	All Possible	.875	6.790
	75 per cent	.613	4.901
	50 per cent	.429	3.233
	25 per cent	.211	1.650
	Pairwise	.197	1.117

$.20\sigma_P = .013$

Table 12. Type I and Type II Error Rates for Scheffé's S-Method: α = .20, .75 MD, Skewed Distribution, Unequal n's (17, 19, 22, 26), Equal σ^2's (1).

		Empirical Estimates	
Error Rates	Contrasts	Type I Errors	Type II Errors
Experimentwise	All Possible	.201	1.000
	75 per cent	.188	1.000
	50 per cent	.168	.988
	25 per cent	.124	.875
	Pairwise	.154	.824
Comparison	All Possible	.036	.276
	75 per cent	.036	.276
	50 per cent	.036	.274
	25 per cent	.033	.270
	Pairwise	.038	.191
Experiment	All Possible	.905	6.906
	75 per cent	.642	4.976
	50 per cent	.433	3.295
	25 per cent	.200	1.623
	Pairwise	.227	1.145

$.20\sigma_P$ = .013

Table 13. Type I and Type II Error Rates for Scheffè's S-Method: α = .20, .75 MD, Skewed Distribution, Unequal n's (17, 19, 22, 26), Unequal σ^2's (.4, .8, 1.2, 1.6).

		Empirical Estimates	
Error Rates	Contrasts	Type I Errors	Type II Errors
Experimentwise	All Possible	.168	1.000
	75 per cent	.161	1.000
	50 per cent	.140	.993
	25 per cent	.106	.882
	Pairwise	.133	.887
Comparison	All Possible	.030	.286
	75 per cent	.030	.288
	50 per cent	.030	.288
	25 per cent	.030	.284
	Pairwise	.031	.208
Experiment	All Possible	.747	7.141
	75 per cent	.548	5.176
	50 per cent	.354	3.458
	25 per cent	.179	1.702
	Pairwise	.186	1.246

$.20\sigma_P$ = .013

Table 14. Type I and Type II Error Rates for Scheffé's S-Method: α = .20, .75 MD, Skewed Distribution, Unequal n's (17, 19, 22, 26), Unequal σ^2's (1.6, 1.2, .8, .4).

Error Rates	Contrasts	Empirical Estimates: Type I Errors	Empirical Estimates: Type II Errors
Experimentwise	All Possible	.207	1.000
	75 per cent	.193	.999
	50 per cent	.189	.982
	25 per cent	.145	.875
	Pairwise	.166	.743
Comparison	All Possible	.043	.262
	75 per cent	.043	.260
	50 per cent	.042	.257
	25 per cent	.044	.268
	Pairwise	.044	.169
Experiment	All Possible	1.082	6.542
	75 per cent	.775	4.673
	50 per cent	.509	3.087
	25 per cent	.262	1.607
	Pairwise	.266	1.012

$.20\sigma_P = .013$

as a result of the form of the distribution when the variances are unequal.

When the number of observations per treatment level is unequal (Table 12), the experimentwise Type I empirical probabilities are larger and closer to theoretical alpha than the probabilities when sampling is from the normal distribution. The discrepancy between the empirical probabilities and theoretical alpha is conditional upon the number of contrasts computed. The per comparison and per experiment Type I estimates are likewise somewhat larger.

For the case in which both the variances and sample sizes are unequal and proportionately paired (Table 13), the experimentwise Type I probabilities are larger and closer to theoretical alpha.

When inversely pairing unequal variances and unequal sample sizes (Table 14), the empirical probabilities of a Type I experimentwise error are reduced considerably when sampling from the skewed distribution. The probabilities are also closer to theoretical alpha. The Type I estimates for the per comparison and per experiment rates are also reduced.

Sampling from this skewed chi-square distribution, caused at times, the empirical probabilities of a Type I error to substantially differ from the probabilities found when the observations were obtained from the normal distribution.

Discussion

Keselman and Toothaker (1971) have pointed out that Scheffé's S-method is not a powerful statistic for detecting reasonable mean differences when the alpha level is set at the traditional five per

cent level. Though the analysis of variance F test had a power of approximately .90 for detecting .75 σ-unit differences when there were seven observations per treatment cell, the power of the Scheffé method for detecting these .75 σ-unit differences, was considerably less than the power that Keselman and Toothaker had "built into" the ANOVA F test.

Two reasonable approaches that can be employed to increase the power of the Scheffé method would be to increase the significance level for the Scheffé test and to increase the sample size per cell. Increasing the mean differences as Petrinovich and Hardyck (1969) had done is an unrealistic procedure for increasing the power. The behavioral scientist must work with the differences in the means that he has found via the manipulation of his independent variable(s).

After rejecting the ANOVA null hypothesis at the .05 level, Scheffé's multiple comparison statistic was computed for all possible of the contrasts, 75, 50, and 25 per cent of the contrasts, and also for the pairwise contrasts. Type I and Type II estimates were were tabulated for different values of significance levels (.05, .10, ... , .25) and for different sample sizes per cell (7, 12, 17, 21).

Under conditions of assumption violations, the probability of a Type I error for the S-method did depart from theoretical alpha, though the degree of departure was most definitely related to the number of contrasts computed.

Regardless of the sample size or the level of significance, the empirical probabilities of a Type II error for the Scheffé method

was extremely large when counting with the experimentwise error rate. Regardless of the number of contrasts computed, the probability of a Type II error was extremely excessive.

The empirical probability of a Type II error, on the other hand, was amenable to changes in alpha and sample size for the per comparison and per experiment error rates.

The power that had been "built into" the ANOVA F test is not the same power that the Scheffé statistic had for detecting mean differences. The experimenter who anticipates using the S-method should consider increasing his sample size per cell, to a number considerably larger than what would be required to detect prespecified mean differences with an ANOVA F test. He should also use an alpha level, for the Scheffé test, larger than the traditionally accepted .05 level. For those who are alpha purists and have only considered guarding against false rejections, the empirical probabilities should be consoling in that the probabilities of a Type I error were generally less than theoretical alpha and for the pairwise contrasts, considerably less.

If a multiple comparison procedure is to follow a significant ANOVA F test (Scheffé, 1959, p. 66) it seems that greater concern should be given to protecting against an excessive number of false acceptances, not false rejections. Rejecting the ANOVA null hypothesis indicates that there were treatment effects; the Scheffé multiple comparison procedure should be adjusted to detect those differences. Increasing the sample size and increasing the alpha level are two reasonable approaches for increasing the power of Scheffé's S-method

for detecting reasonable mean differences.

References

Aitkin, M. A. Multiple comparisons in psychological experiments. The British Journal of Mathematical and Statistical Psychology, 1969, 22, 193-198.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. I, Effect of inequality of variances in the one-way classification. Annals of Mathematical Statistics, 1954a, 25, 290-302.

Box, G. E. P. 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, 1954b, 25, 484-498.

Box, G. E. P., and Anderson, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, Series B, 1955, 17, 1-26.

Gabriel, K. R. A procedure for testing the homogeneity of all sets of means in analysis of variance. Biometrics, September, 1964, 459-477.

Games, Paul, A. Inverse relation between the risks of Type I and Type II errors and suggestions for the unequal n case in multiple comparisons, Psychological Bulletin, 1971, 75, 97-102.

Horsnell, G. The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 1953, 40, 128-136.

IBM, 1130 Scientific Subroutine Package (1130-CM-02X) Programmer's Manual, H20-0252-1, International Business Machines Corporation, 1967.

Kendall, M. A., and Stuart, Alan. The Advanced Theory of Statistics, Vol. 1. New York: Hafer Publishing Company, 1969.

Keselman, H. J., and Toothaker, L. E. A comparison of Scheffé's S-method and Tukey's T-method for various numbers of all possible contrasts under violation of assumptions. Unpublished doctoral dissertation, University of Oklahoma, 1971.

Miller, Rupert, G., Jr. Simultaneous Statistical Inference. New York: McGraw-Hill Book Company, 1966.

Petrinovich, Lewis, F., and Hardyck, Curtis, D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

Ryan, Thomas, A. Multiple comparisons in psychological research. Psychological Bulletin, 1959, 56, 26-47.

Ryan, Thomas, A. The experiment as a unit for computing rates of error. Psychological Bulletin, 1962, 59, 301-305.

Scheffé, Henry. A method for judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheffé, Henry. The Analysis of Variance. New York: John Wiley & Sons, Inc., 1959.

Appendix II

DISSERTATION PROSPECTUS

The researcher in the behavioral sciences intending to explore multiple treatment-effects, where there are two or more levels of the treatment variable, has available to him a most versatile statistical tool to aid him in evaluating his data. Indeed, the analysis of variance (ANOVA) and its various theoretical-methematical models have at last become a primary statistical tool to aid behavioral researchers evaluate their data.

A one-way fixed effects ANOVA is one of the models typically employed by behavioral scientists. This model is most applicable to those experimental settings in which the researcher expects to examine sample mean ($\bar{X}.$) differences associated with different treatment conditions or levels. In the fixed effects model the researchers' experimental interests are for those, and only those, fixed treatment levels manipulated in the actual experiment; this means, that in any replication of the experiment these same treatment levels will once again be manipulated. The one-way fixed effects ANOVA model only permits the researcher to <u>statistically</u> generalize to those treatment levels manipulated within the experiment (Hays, 1963).

The hypothesis of interest usually is that $\mu_1 = \mu_2 = \ldots = \mu.$ That is, the null hypothesis subjected to a statistical test is that the population means for the various treatment levels are equal; hence the observations within each treatment level have been randomly

sampled from one population with mean μ. Consequently, when this hypothesis cannot be rejected the researcher concludes that statistically there appears to be no treatment effects. Naturally, such a statement is made in probabilistic terms.

For the null hypothesis to be rejected at some predetermined probability level (specified by α, the probability of a Type I error), at least two of the treatment means must be statistically different. A treatment effect is defined as $\alpha_1 = \mu_1 - \mu$. The expected value[1] (long run average) of $\hat{\alpha}_1$ is equal to $\mu_1 - \mu$. Therefore, the unbiased estimate of the population treatment effect $\mu_1 - \mu$ is $\bar{X}._1 - \bar{X}..$, the difference of the grand mean $(\bar{X}..)$ and the mean of the treatment level $(\bar{X}._1)$. If the number of treatment levels, K, equals four, $\alpha_1 = \mu_1 - \mu$, $\alpha_2 = \mu_2 - \mu$, $\alpha_3 = \mu_3 - \mu$, and $\alpha_4 = \mu_4 - \mu$. The null hypothesis of no treatment effects can be stated as $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ or $(\mu_1 - \mu) = (\mu_2 - \mu) = (\mu_3 - \mu) = (\mu_4 - \mu) = 0$, or in general, $\alpha_1 = \alpha_2 = \ldots = \alpha_K = 0$. Previously the null hypothesis (H_o) was specified as $\mu_1 = \mu_2 = \ldots = \mu$. It can be shown that the two ways so far presented as expressing the null hypothesis are equivalent.[2] Therefore, for the null hypothesis to be rejected at least any two α_k's must be statistically different.

Having rejected the null hypothesis for the one-way fixed effects ANOVA, the researcher can conclude in probabilistic terms that the means differ statistically. If the experimenter was interested in determining whether any treatment effects existed, then the one-way ANOVA is indeed a most convenient and versatile statistical tool to

detect such differences and this hypothetical experimenter could consider that his statistical question had been answered. On the other hand, the above example is indeed hypothetical in that it is a rare occasion when the experimenter is content in just being able to state that the treatment levels were different. Of course, the experimenter's interest then is for further exploration of these different means. The one-way fixed effects ANOVA merely reflects whether at least any two of the treatment levels differ. Did treatment level one differ from level two or level three or level four or perhaps the combination of treatment levels one and two differed from the combination of three and four, etc., etc.? These are the type of questions that are generally of interest. How many researchers are content in merely being able to say that there were differences without being able to specify exactly where the differences lie?

Because behavioral scientists are usually interested in digging deeper into their data, probing techniques were developed to be used following the rejection of the analysis of variance null hypothesis. Prior to a discussion of post-hoc probing procedures [post-hoc indicating probing without any pre-existing statistical questions in mind but rather exploring any or all aspects of the data after examining the data (Hays, 1963)] it would be beneficial to consider the statistical technique of individual comparisons, of which post-hoc is a subcategory.

A population contrast among means is defined $\Psi_l = c_1 u_1 + c_2 u_2 + \ldots + c_k u_k$ where the $\sum_{k}^{K} c_k = 0$ and not all the c_k's equal zero. To estimate a population contrast, the population means are replaced by their sample

counterparts, $\bar{X}_{.k}$. That is, a sample contrast is defined as $\hat{\psi}_1 = c_1\bar{X}_{.1} + c_2\bar{X}_{.2} + \ldots + c_K\bar{X}_{.K}$, again with the restriction that the $\sum_{k}^{K} c_k = 0$, where not all the c_k's are zero. This sample contrast unbiasedly estimates the population contrast.[3] Returning to the previous example of four treatment levels, if an experimenter had specific questions to ask of the data before running the experiment and seeing the actual data then his comparisons would be "planned contrasts" or in Haysian dialect "planned comparisons." When the behavioral scientist has specific questions in mind prior to the actual running of the experiment the technique of planned comparisons is used instead of the ANOVA F-test. Before exploring the question of planned comparisons versus the general over-all ANOVA test, further consideration must be given to the planned comparison technique.

Given four levels of a treatment variable there are twenty-five ways in which the means can be compared.[4] The data from any one of the twenty-five comparisons is dependent or related to the data from some of the other comparisons. The sample means from a given set of data are used in more than one of the comparisons, hence, the comparisons cannot be considered statistically independent but, to the contrary, the comparisons are obviously statistically related.

The planned comparison technique can be used to circumvent the problem of statistical dependency. Whether the researcher wants to limit his comparisons to only independent comparisons or perform the more voluminous nonindependent comparisons, basically, boils down to a value judgement as to the merits of the two approaches. Strong

arguments can be offered in defense and against each approach. To assure himself that each of his comparisons are independent of one another and independent of the grand mean[5] the researcher must restrict the number of comparisons to K-1 (the degrees of freedom for the between-groups sums of squares). In our example there would therefore be three independent comparisons for the set of four means. By restricting the number of comparisons to K-1 the researcher guarantees himself that he is working with non-redundant independent combinations of the data. Here is one apparent advantage of the independent comparison technique.

Before performing a statistical test on any given comparison, e.g., $\hat{\Psi}_1 = c_1\overline{X}._1 + c_2\overline{X}._2 + \ldots + c_K\overline{X}._K$, as in most statistical tests, the researcher must have an estimate of the sampling or error variance of whatever it is that he is interested in testing. Therefore, it is necessary to check the linear comparison of a given set of means with the sampling fluctuation expected merely by chance. What is needed then is the standard-error of the sampling distribution of a sample comparison. It can be shown (Hays, 1963), that the variance of a sample contrast[6] is estimated by the product of the MS error and $\sum_K c_k^2/n_k$. At this point we have the necessary components to perform a statistical test. For planned comparisons the statistical hypothesis that is put to the test is $H_o: \Psi = 0$; that is, in the population the contrast is equal to zero.

Post-hoc techniques are also considered as a subcategory of

the individual comparison procedures. As the name implies, and as this author stated previously, this technique follows the analysis of variance, specifically after a significant ANOVA F test. This author believes that the primary question of whether the data should be analyzed with a planned comparison technique or a post-hoc comparison technique is determined by whether the researcher has specific questions in mind prior to the actual running of the experiment and data collection. Opposed to the planned technique, post-hoc techniques are to be used when the experimenter does not have specific a priori questions in mind prior to the collection of his data but, decides on avenues of exploration after running the experiment and looking at the data (Hays, 1963; Miller, 1966; Scheffé, 1959; Stanley, 1957). Because this author prefers to distinguish the two techniques, in terms of the appropriateness of their use, primarily on the above qualification this author finds the usual arguments in favor of each of these techniques not substantially pertinent to the question of when each should be used. Nonetheless, for those who do not choose to accept this writer's evaluation, the usually cited advantages for each technique will be enumerated so the reader can decide and choose for himself one of the two techniques for any of the reasons that he believes to be most pertinent.

Since there can be only K-1 independent planned comparisons there are therefore a restricted number of questions that can be asked of the data. On the other hand, the number of questions, for

most post-hoc techniques are in many and most instances unrestricted. The advantage of the post-hoc techniques consequently is that they are not limited to just K-1 comparisons; this advantage can be considered at the same time as a disadvantage. That is, the comparisons are not independent, and nonindependence can be considered a disadvantage. Also, for any contrast of the means the planned technique is generally more powerful than the post-hoc technique for detecting mean differences. Along with greater power there is necessarily the increased likelihood of a Type I error; these techniques are considered, therefore, generally less conservative than the post-hoc comparison procedure. When the experimenter's concern is for interval estimation and the accompanying confidence coefficient, rather than with tests of significance (Scheffé's method was originally developed for interval estimation), the length of the confidence interval is generally shorter for the planned technique.

The question of choosing between the two techniques is not just one of deciding between the more powerful, limited number of planned independent comparisons or the less powerful, generally unlimited number of post-hoc comparisons; the experimenter can also, if he so chooses, calculate nonindependent planned comparisons. Nonindependent planned comparisons, as the name implies, permits the researcher to carry out comparisons which exceed, in number, the restrictions of independence. In effect, the experimenter is unrestricted in the number of contrasts that he may choose to

explore; he no longer is restricted to just K-1 comparisons. A problem inherent to nonindependent planned comparisons is the lack of control for the Type I error. Each contrast is typically tested at a predetermined level of significance. Yet, for the set of nonindependent contrasts, the probability of a Type I error is greater than the level for any one of the contrasts and would even be greater than $1-(1-\alpha)^c$ (the probability of independent multiple t-tests), where c is the number of comparisons. For small α, $1-(1-\alpha)^c$, is approximately equal to $c\alpha$. To control the number of Type I errors, Dunn's (1961) procedure is often used. Dunn's procedure is quite simple in logic in that the technique merely divides alpha by the number of comparisons in the set and consequently sets the alpha level for each test at the new partitioned level of significance. For example, if one wanted to test four linear contrasts and at the same time guard against the number of Type I errors exceeding the 5 per cent value for the complete set of four contrasts, one could use Dunn's technique to divide alpha among the four contrasts. An alpha of .05 can most easily be split among the four contrasts simply by dividing .05 by the number of contrasts. Therefore, each contrast to be significant would now have to equal or exceed the critical value for the new alpha now set at .0125 (.05/4 = .0125). Dunn's procedure, also called Bonferroni t statistics (Miller, 1966), can also set the alpha level for each contrast, not only by equally dividing up the level of significance, but rather as a function of

the experimenter's concern for the control of a Type I error he wishes to assign to a particular contrast. That is, each contrast can be tested at any preconceived level of significance, so long as the total alpha for the set of nonindependent contrasts does not exceed a specified level of significance, (e.g., $\alpha_1 = .005$, $\alpha_2 = .015$, $\alpha_3 = .02$, $\alpha_4 = .01$; $\Sigma\alpha = .05$).

The restrictive rationale of limiting one's planned comparisons to K-1 and thereby maintaining independence among the contrasts has recently been challenged by Davis (1969). Davis questions and challenges the necessity of independence for the planned technique since independence is left by the wayside in the post-hoc procedures. Davis feels if the experimenter does not become overly concerned about the lack of independence for post-hoc comparisons why then the concern over independence when the comparisons happen to be planned? Based on this argument then, Davis offers a procedure and index whereby the experimenter can measure the power of his planned nonindependent comparisons relative to Scheffé's post-hoc procedure allowing the researcher. . . "to trade off between number of comparisons and power while holding the Type I error rate constant" (Davis, 1969).

Dunn's technique and Davis' recommendations have been offered so that the reader can get a feel for the divergence of technique and opinion in regard to the topic of individual comparisons. Like most statistical procedures, there are varied opinions concerning the best means of exploring a particular statistical question.

Therefore, there is not usually just one statistical technique available to explore one's data with, as a new student in statistics might erroneously suspect, but oftentimes many procedures to choose from.

The reader can take his pick as to the advantages and disadvantages in deciding which technique to use, but, this writer again suggests that choosing one of the individual comparison techniques for reasons related to (1) the number of comparisons that can be examined or, (2) because of the question of the number of Type I and Type II errors or, perhaps (3) because of the length of the confidence interval, should not be the primary determinant of the decision, but, at the crux of the decision should be the consideration of whether the questions to be asked of the data have been formulated prior to data collection or, after running the experiment and examining the data.

Another statistical issue that is pertinent to the research to be presented and to individual comparison procedures is the issue of error rates. The various techniques of planned and post-hoc comparisons control different types of error rates. Ryan (1953) and Kirk (1968) offer the following definitions to the various error rates:

$$\text{Error rate per comparison} = \frac{\text{number of comparisons falsely declared significant}}{\text{total number of comparisons}}$$

Error rate per hypothesis =

$$\frac{\text{number of hypothesis falsely declared significant}}{\text{total number of hypotheses}}$$

Error rate per experiment =

$$\frac{\text{number of comparisons falsely declared significant}}{\text{total number of experiments}}$$

Error rate experimentwise =

$$\frac{\text{number of experiments with at least one statement falsely declared significant}}{\text{total number of experiments}}$$

Error rate per family =

$$\frac{\text{number of comparisons falsely declared significant}}{\text{total number of families}}$$

Error rate familywise =

$$\frac{\text{number of families with at least one statement falsely declared significant}}{\text{total number of families}}$$

The following example will illustrate how three of the above error-rates differentially count Type I errors. Suppose there were 1000 experiments and for each experiment six linear contrasts were each subjected to a test of significance. For the total 6,000 statements of significance 480 are truly false and these false statements are contained within just 100 experiments. The error-rates per comparison, per experiment, and experimentwise are:

1. Error rate per comparison: 480/6,000 = .08

2. Error rate per experiment: 480/1,000 = .48

3. Error rate experimentwise: 100/1,000 = .10

Among psychologists, there is presently a controversy as to which error rate should be used in psychological research. Ryan (1959, 1963) and Wilson (1960) perhaps are most representative of the two positions currently competing for the forefront. Ryan is for setting the error rate for the experiment while Wilson believes the error rate should be the traditionally accepted error rate per comparison or per hypothesis. After bantering over equally logical arguments in favor of their respective error rates, both gentlemen are in the final analysis concerned that the number of Type I errors, no matter how they are counted, should be adequately controlled. Ryan and Wilson are rightfully concerned about the number of false statements but they implicitly address their arguments to the need of controlling Type I errors. It is certainly commendable to guard against an excessive number of Type I errors but like the sacred 5 and 1 per cent criteria levels, one should not prima facie always assume that one's duty is to protect Type I errors. That is, for certain circumstances greater concern should be paid to the number of Type II errors and perhaps the extreme attitude could be taken: damn the number of Type I errors!

Individual comparison techniques are intricately tied to the concept of error rates, or at least they should be. In deciding upon an appropriate individual comparison procedure the behavioral

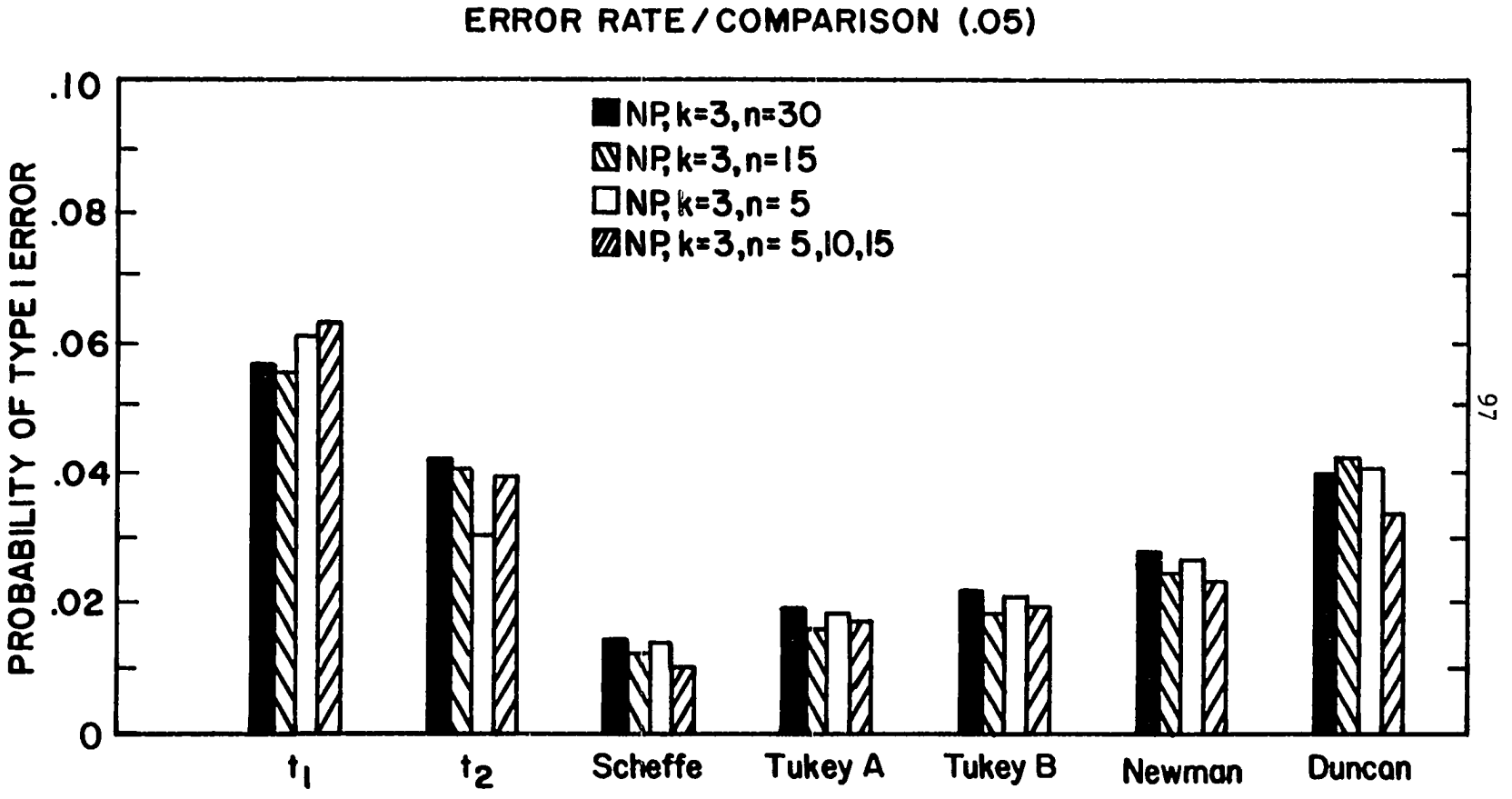
scientist should consider (1) whether he is familiar enough with the area of research to ask sensitive a priori questions, (2) the potential advantages and disadvantages of planned versus post-hoc comparisons, and (3) how the individual comparison technique, whether planned or post, controls for the number of false statements. Both techniques that will be explored in this research, Scheffé's S-Method and Tukey's T-Method, control the error rate experimentwise under the null hypothesis.

There is another equally important question, or better described as a controversy, that is very pertinent to the research explored in this dissertation. This question is whether post-hoc procedures, Scheffé's S-Method and Tukey's T-Method in particular and post-hoc techniques in general, are to follow the analysis of variance only when the ANOVA hypothesis has been rejected or whether the post-hoc techniques may be applied even if the null hypothesis has not been rejected. This author cannot offer a definitive answer to this question; there are those on both sides of the issue and others who ignore the question entirely. Hays (1963) states that post-hoc techniques are to follow a significant $\underline{F}$ as does McNemar (1962; whereas Edwards (1960), Ryan (1959a), and Kirk (1968), state that a significant $\underline{F}$ is not necessary to use the post-hoc technique. Federer (1955) and Winer (1962) make no mention as to whether a significant $\underline{F}$ is first required or not.

Recent empirical investigators explored Type I and Type II errors for various multiple comparison techniques under conditions

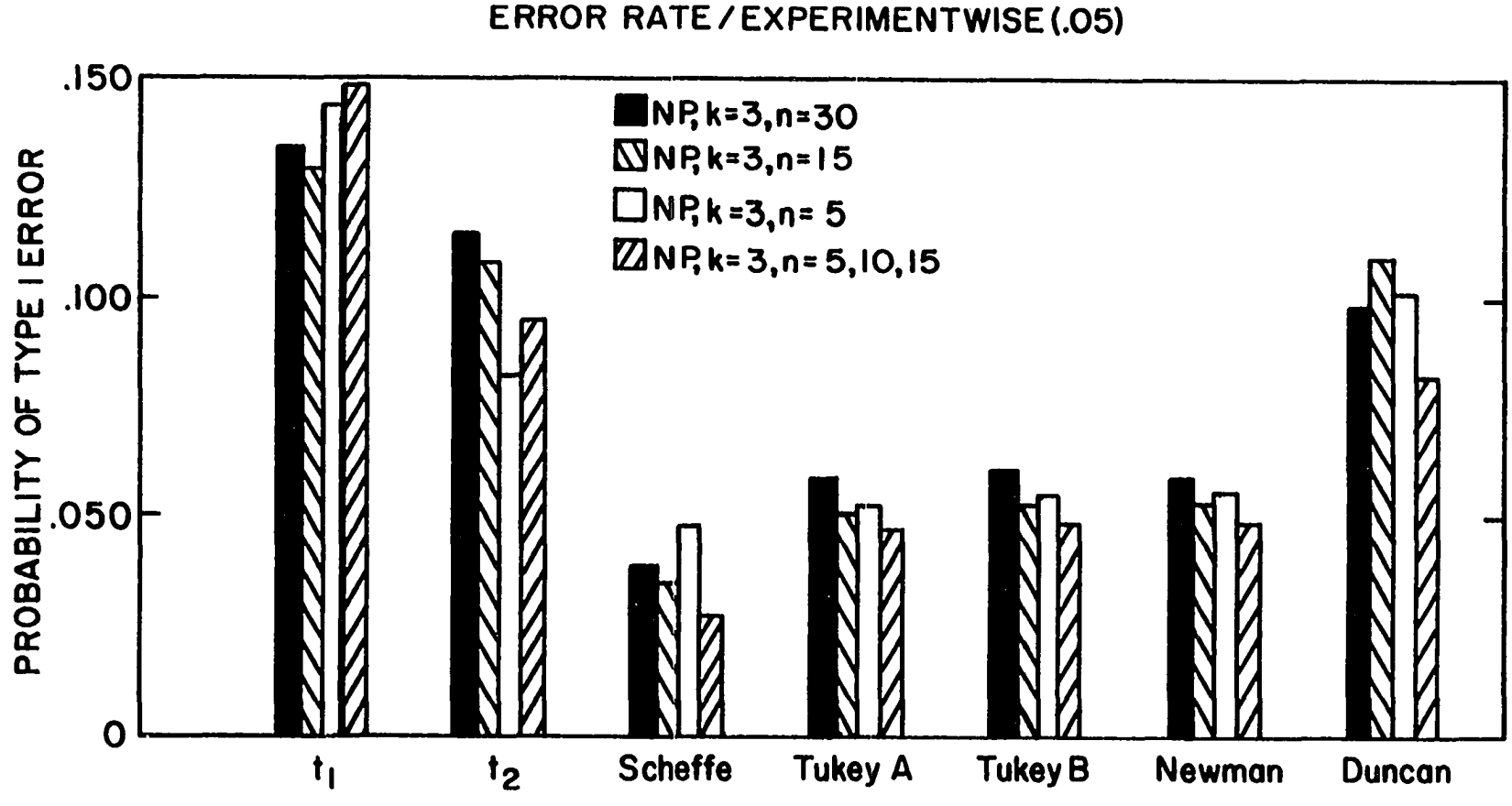
of no differences, without performing an ANOVA and checking to see whether the null hypothesis of $\mu_1 = \mu_2 = \ldots \mu$ was tenable or not, and under conditions of real differences, the alternative case, (Petrinovich and Hardyck, 1969). Under their zero difference conditions (the true null case), Petrinovich and Hardyck empirically investigated Tukey's T-Method and Scheffé's S-Method along with other multiple comparison procedures to compare the frequency of the Type I error. As Miller (1966) points out "the basic premise of simultaneous statistical inference and multiple comparisons is to given increased protection to the null hypothesis and bears no head to the number of errors that may occur under the alternative." (Miller, 1966, p. 32). Petrinovich and Hardyck are checking the number of Type I errors to theoretical alpha for the null case.

Under varied conditions of population shape and variance and for different number of treatment levels and sample sizes and for the different error rates, Tukey's and Scheffé's methods were generally found to be conservative. When the error rate was set per comparison Scheffé's and Tukey's techniques were found to be approximately .02 or less (reading from Petrinovich and Hardyck's Figure 1) for the varied conditions of different numbers of treatment levels and different numbers of sample sizes. For an experimentwise error rate, empirical alpha for Scheffé's technique was .05 and less, while Tukey's technique fluctuated slightly above and below the 5 per cent level (reading from Petrinovich and Hardyck's Figure 2 and Figure 4). The only exception of the number of Type I



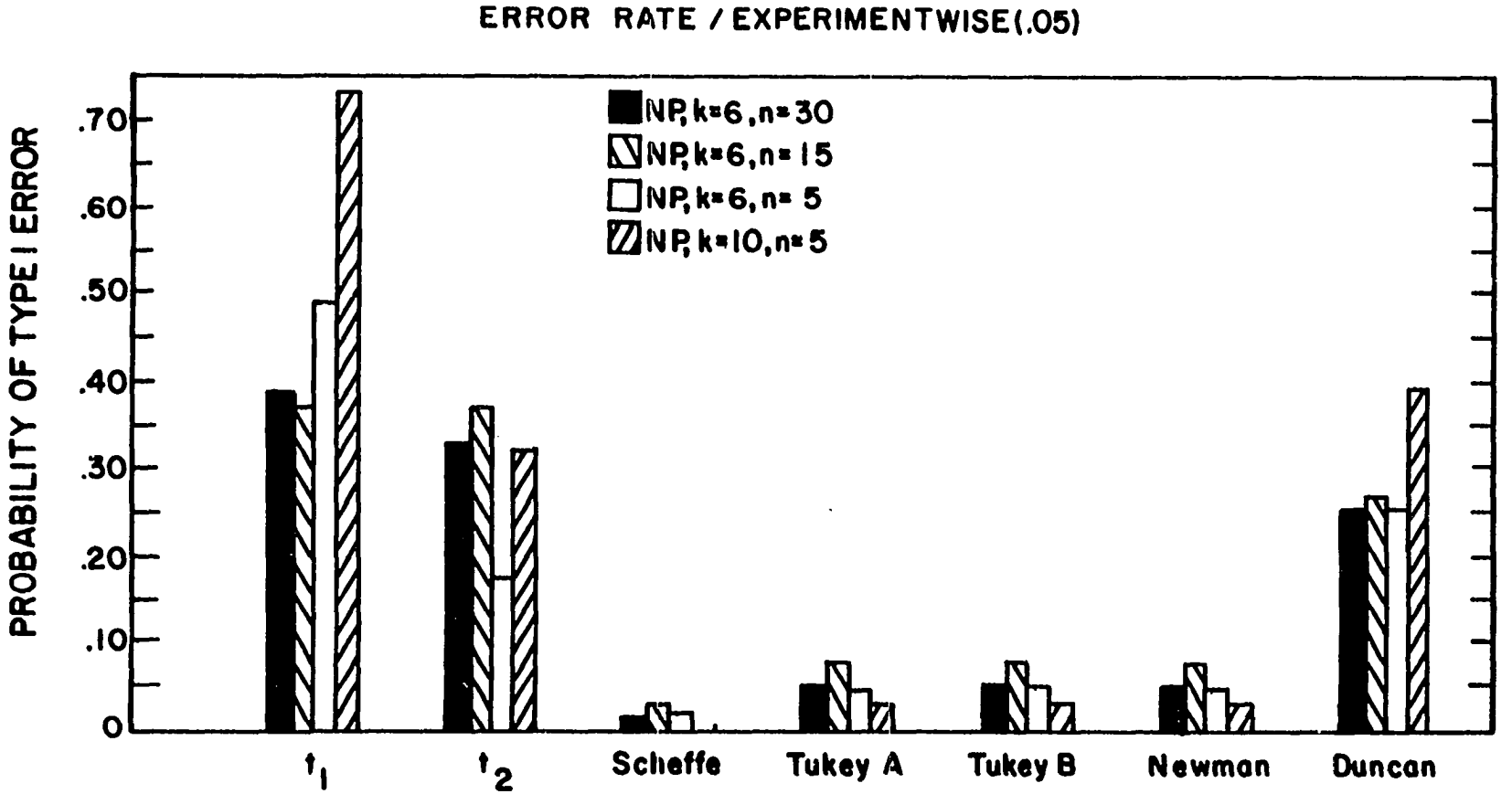
Petrinovich and Hardyck's Fig. 1

Error rate per comparison for all methods when three samples are compared.



Petrinovich and Hardyck's Fig. 2

Error rate experimentwise for all methods when three samples are compared.



Petrinovich and Hardyck's Fig. 4

Error rate experimentwise for all methods when three samples are compared.

errors being less than or equal to theoretical alpha, occurred when unequal variances were combined with unequal sample sizes. As has been found for the robustness studies of the ANOVA F test, the combination of unequal variances and unequal sample sizes affects the probability of a Type I error: Boneau (1960), Box (1954a, b), Box and Anderson (1955), Cochran (1947), Godard and Lindquist (1940), Horsnell (1953), Hsu (1938), Lindquist (1953), Scheffé (1959), and Welch (1937).

General conclusions that can be drawn from the research investigating variance heterogeneity are that for equal number of observations per treatment cell, heterogeneous variances do not substantially affect the probability of the Type I error, but, when the number of observations per cell is not equal, variance heterogeneity will substantially affect the probability of a Type I error. When the sample sizes and variances are unequal and the smaller samples have been selected from the populations with the small variances than the probability of a Type I error is less than alpha. For the conditions of variance heterogeneity investigated by Petrinovich and Hardyck, the number of Type I errors for Scheffé's method when the error rate was set experimentwise, per experiment, and per comparison were .016, .018, and .006, respectively, for the proportional case (e.g., n = 5, 10, 15, V = 1, 2, 4). For this same proportional case the number of Type I errors for Tukey's technique for the three error rates was found to be .012, .013, and .004 (reading from Table 1). Just as has

Error Rates for Multiple Comparisons

Petrinovich's and Hardyck's Table 1

Type I Error Rate (.05) for Multiple Comparison Methods

Population	Error Rate [a]	t_1	t_2	Scheffé	Tukey A	Tukey B	Newman Keuls	Duncan
NP, k = 3, n = 5, 10, 15 V = 1, 2, 4	1REW	.119	.038	.016	.012	.012	.012	.030
	1REX	.114	.046	.018	.013	.016	.018	.038
	1RC	.048	.015	.006	.004	.005	.006	.013
NP, k = 3, n = 5, 10, 15 V = 4, 2, 1	1REW	.157	.233	.100	.141	.141	.141	.212
	1REX	.216	.341	.141	.192	.213	.235	.319
	1RC	.072	.114	.047	.064	.071	.078	.106

Note.--The percentage of F values significant at $p < .05$ is 2.5% for the first population and 11.7% for the second.

[a] 1REW = Type 1 error rate experimentwise; 1REX = Type I error rate per experiment; 1RC = Type I error rate per comparison.

been found with the robust studies for the ANOVA F test the combination of the smaller samples with the smaller variances will cause the number of Type I errors to be less than theoretical alpha (i.e., conservative test). When the smaller samples are selected from the populations with the larger variances (e.g., n = 5, 10, 15, V = 4, 2, 1) the number of Type I errors for Scheffé's technique was .100 experimentwise, .141 per experiment and .047 per comparison. The frequency of Type I errors found for Tukey's method were .141 experimentwise, .192 per experiment, and .063 per comparison. The general conclusion to be derived is, like the ANOVA F test, when the larger of the variances is paired with the smaller of the samples the number of Type I errors will generally be greater than theoretical alpha. The multiple comparison procedures, like the ANOVA F test, suffer the same lack of robustness for certain combinations of unequal n's and unequal variances.

The frequency of Type I errors for the Tukey and Scheffé procedures are generally less than theoretical alpha; that is, these techniques were found to be conservative under the null case. One could say that not only do these techniques, Tukey's somewhat, Scheffé's generally, afford ample protection to falsely rejecting the null hypothesis but perhaps it should be said that at times they over-protect.

The two types of errors in hypothesis testing are related in that an increase or decrease in one of them will cause a decrease

or increase in the other. This relationship is borne out in the research of Petrinovich and Hardyck (1969). Table 2 presented by these authors contains the Type II error rates.

Type II errors were investigated for varied conditions of population shape, population variance, and for differing numbers of cells and sample sizes. Another factor manipulated by the authors was the mean differences between adjoining means ranging from .6 σ-unit differences to 2.6 σ-unit differences. For example, in order to build in .6 σ-unit differences between the three means (when k = 3) μ_1 is set equal to zero, μ_2 is set equal to .6 and μ_3 is set at 1.2. By adding these values to the observations for each respective treatment level, differences are "built-in" between the population means. Inspection of Table 2 should upset the researcher. What is so upsetting is the inordinate number of Type II errors. With the number of beta errors as large as they were generally found to be, the power of the multiple comparison techniques is generally disappointingly low! Only under conditions of large sample size and large mean differences does the number of Type II errors become respectable and the power of these post-hoc techniques also attains a respectable level. At this point Miller must be requoted . . . "the basic premise of simultaneous statistical inference and multiple comparisons is to give increased protection to the null hypothesis and bears no head to the number of errors that may occur under the alternative." (Miller, 1966, p. 32). The research reported by Petrinovich

Error Rates for Multiple Comparisons
Petrinovich's and Hardyck's Table 2
Type II Error Rates (α = .05) for Increasing Mean Differences

Population	Error Rate[a]	MD[b]	t_1	t_2	Scheffé	Tukey A	Tukey B	Newman Keuls	Duncan
NP, k = 3, n = 30	2REW	.6	.545	.543	.781	.740	.652	.548	.547
		1.0	.065	.067	.174	.148	.100	.067	.067
		1.3	.003	.003	.008	.007	.003	.003	.003
		1.6	.000	.000	.000	.000	.000	.000	.000
	2RC	.6	.190	.189	.305	.279	.235	.191	.190
		1.0	.022	.022	.058	.049	.033	.022	.022
		1.3	.001	.001	.003	.002	.001	.001	.001
		1.6	.000	.000	.000	.000	.000	.000	.000
NP, k = 3, n = 15	2REW	.6	.875	.874	.974	.973	.930	.875	.874
		1.0	.482	.474	.740	.685	.578	.477	.476
		1.6	.018	.017	.060	.049	.026	.017	.017
		2.0	.000	.000	.004	.003	.001	.000	.000
	2RC	.6	.356	.356	.488	.460	.412	.363	.356
		1.0	.170	.167	.285	.259	.210	.168	.167
		1.6	.006	.006	.020	.016	.009	.006	.006
		2.0	.000	.000	.001	.001	.000	.000	.000
NP, k = 3, n = 5	2REW	.6	.979	.976	.995	.993	.992	.976	.976
		1.0	.931	.912	.970	.963	.943	.912	.912
		1.6	.615	.557	.799	.764	.675	.557	.557
		2.0	.389	.337	.605	.558	.446	.337	.337
		2.6	.086	.058	.171	.147	.094	.058	.058
	2RC	.6	.457	.432	.594	.588	.494	.483	.462
		1.0	.419	.392	.541	.509	.456	.401	.392
		1.6	.233	.207	.353	.325	.268	.207	.207
		2.0	.138	.118	.237	.214	.163	.118	.118
		2.6	.029	.019	.061	.051	.031	.019	.019

[a] 2REW = Type II error rate experimentwise; 2RC = Type II error rate per comparison.

[b] Mean difference.

and Hardyck offers startling evidence as to what happens to error rates under the alternative distribution. Scheffé's and Tukey's techniques, multiple comparison techniques designed for simultaneous inference, are indeed prone to an excessive number of Type II errors, more so than were the other multiple-comparison techniques investigated by Petrinovich and Hardyck. When the samples were from populations with different variances, the multiple comparison procedures became even less sensitive, i.e., the frequency of Type II errors increased. The number of Type II errors was found to be generally invariant for the different combinations of unequal variances and unequal sample size.

Petrinovich and Hardyck in summary felt that unless mean differences are very large it is fruitless to pursue multiple comparisons when sample sizes are less than ten, for the power of these comparison procedures was in most cases practically nill or generally quite low by most accepted standards.

As Reese (1970) points out, Petrinovich and Hardyck did not, unfortunately, place enough emphasis in the right direction. Reese feels that in evaluating the merits of the multiple comparison methods, Petrinovich and Hardyck appeared to show a greater concern for Type I errors rather than Type II errors and consequently, advocated the use of Scheffé's and Tukey's techniques. Yet, Reese feels that Scheffé's and Tukey's techniques perhaps should not be considered so favorable if a researcher has a greater concern for protecting the number of Type II errors rather than Type I errors.

Reese further points out that if post-hoc multiple comparison techniques are to follow a significant ANOVA F test than it seems logical that Type II errors are indeed the errors to guard against.

This author agrees with Reese that post-hoc techniques should follow the analysis of variance when the null hypothesis has been rejected. Post-hoc comparisons are probing, postmortem techniques (Stanley, 1957). The name itself implies that the technique is to follow something; in this case follow an analysis of variance. But why would one follow up a non-significant ANOVA F value? If the null hypothesis was not rejected, statistically, the researcher should conclude that there are no differences in the means for the treatment levels; why then would one pursue something that was just shown not to be present--treatment effects? Then again, by name, a probing procedure would search for something that one believed to be present. When one fails to reject the null hypothesis one must therefore conclude that nothing is happening; consequently, this writer feels that there is no logic in probing for something that one has in the last breath said did not exist, at least by the test performed. Therefore, from at least this line of reasoning (additional support will follow) this author takes the position that post-hoc techniques are to follow the analysis of variance when the null hypothesis has been rejected. The technique permits the researcher to explore all possible contrasts in an attempt to detect those contrasts that lead to the rejection of the null hypothesis of the analysis of variance.

Consequently, like Reese, this author feels Petrinovich and Hardyck too hastily jumped on the Scheffé band-wagon. If post-hoc techniques are to follow significant F values from the analysis of variance, then guarding against Type II errors could be considered the more serious error that needs to be controlled. In light of these considerations, judgement should be reserved regarding the general efficacy attributed to Scheffé's technique by Petrinovich and Hardyck until further research is available for the post-hoc procedures.

Tukey (1953, unpublished, privately circulated manuscript) is credited by Scheffé for devising a method to simultaneously estimate all contrasts (Scheffé, 1959). Tukey's technique, the T-Method, utilizes the Studentized Range to investigate differences among means following the rejection of the ANOVA null hypothesis. Scheffé (1959) states that for Tukey's T-Method the probability is $1-\alpha$ that the relationship (1) holds for all pairwise contrasts given that the restrictions on the method are satisfied. The restrictions are that the contrasts have equal variances, and that the number of observations per treatment level are equal.

$$(\bar{X}_k - \bar{X}_{k'}) - q_{\nu_1\nu_2} (MS_e/n)^{\frac{1}{2}} \leq (\mu_k - \mu_{k'}) \leq (\bar{X}_k - \bar{X}_{k'}) + q_{\nu_1\nu_2} (MS_e/n)^{\frac{1}{2}} \quad (1)$$

In repeated experiments therefore, the probability is $1-\alpha$ that all pairwise intervals simultaneously cover their true value of the population contrast. In its original formulation, according to Scheffé (1959), Tukey's method was designed to set limits around pairwise contrasts,

e.g., $\hat{\psi} = c_1\bar{X}_{.k} + c_2\bar{X}_{.k'}$. Scheffé (1959), Winer (1962), and Kirk (1968) present ammended procedures for Tukey's T-Method, sometimes called the Honestly Significant Difference technique, that are appropriate for contrasts other than pairwise contrasts and also when the number of observations per treatment level are not equal.

Smith (1971) empirically checked the robustness of Tukey's post-hoc procedure to Type I errors when the groups are of unequal sample size. The different procedures that have been suggested for this type of problem and therefore checked by Smith when sample sizes are unequal are: (1) the harmonic mean of the group sizes, (2) the harmonic mean of the two extreme range group sizes, and (3) the average value of the group sizes. Smith had found that the two harmonic mean approximations yielded Type I errors that were more consistently congruent with theoretical alpha than did the average group size approximation.

To circumvent the limited applicability of Tukey's T-Method, Scheffé (1953, 1959) formulated his S-Method which is a generalized version of Tukey's method but uses the sampling distribution of F. For all possible contrasts the probability is $1-\alpha$ that all contrasts simultaneously satisfy the relationship in (2).

$$\hat{\psi} - \sqrt{(k-1)\, F_{\nu_1 \nu_2}} \sqrt{MS_e\left(\Sigma\, c_k^2 / n_k\right)} \leq \Psi \leq \hat{\psi} + \sqrt{(k-1)\, F_{\nu_1 \nu_2}} \sqrt{MS_e\left(\Sigma\, c_k^2 / n_k\right)} \quad (2)$$

Here again the probability is $1-\alpha$ that the confidence intervals for

all contrasts will simultaneously cover their true psi values. For example, for four treatment levels there are twenty-five possible contrasts. In 1000 experiments the probability should be $(1-\alpha)\%$ that all contrasts simultaneously bracket their true psi values. If the 95% confidence limit was chosen, 950 experiments would have all twenty-five intervals bracketing their true psi values; 50 experiments will have at least one interval (experimentwise) not bracketing its true psi value.

Scheffé's S-Method is not dependent upon equal variances nor equal sample sizes for its validity. Also, Scheffé's technique is applicable to any form of contrast and not merely to pairwise contrasts. For these reasons and others, Glass & Stanley (1970) report that Scheffé's S-Method is generally preferred by mathematical statisticians.

In addition to setting limits around a contrast, Tukey's and Scheffé's techniques can be used to test the hypothesis that the contrast equals zero, e.g., $\Psi = 0$. Scheffé (1953, 1959) states that the hypothesis is tested according to whether his interval inclusively includes or excludes the value of $\Psi = 0$. That Scheffé's technique can be used to test the hypothesis that $\Psi = 0$ points out the relationship of the S-Method and the analysis of variance. According to Scheffé (1953, 1959) and Miller (1966), the null hypothesis for the ANOVA is equivalent to the statement that all the contrasts are zero.

> Scheffé states . . .if we say that an estimated contrast $\hat{\psi}$ is significantly different from zero or not according as the interval . . . (2) . . .excludes or includes the value $\Psi = 0$, we shall find . . .that the F-test rejects H if and only if some $\hat{\psi}$ are significantly different from zero. In other words, if (and only if) at the α level of significance, the F-test concludes that the true contrasts are not all zero, then the above method will find estimated contrasts which are significantly different from zero (Scheffé, 1959, pg. 67).

Both Scheffé (1953, 1959) and Miller (1966, p. 50 and 51) offer proofs that tests of significance can be made from inspection of the confidence interval.

There have been few comparisons between Tukey's T-Method and Scheffé's S-Method. Scheffé's (1953, 1959) Tables 3a and 3b compare the relative efficiencies of the two methods when the number of treatment levels is four and six. Scheffé's comparison was restricted to conditions of equal variances for the contrasts and equal observations per treatment conditions (these restrictions are those under which Tukey's T-Method was derived). The column headed 1/R gives the relative efficiency of Scheffé's S-Method as compared to Tukey's T-Method. On the other hand, the column headed R gives the relative efficiency of Tukey's T-Method as compared to Scheffé's S-Method. The column labeled Type of Contrast specifies the form of the contrast. That is, (1, 1) is a comparison of a single mean and another single mean, (1, 2) is a comparison of a single mean and the average of two other means while, (1, 3) is a contrast comparing a single mean and the average of three means. When k = 4 (reading from Scheffé's Table 3a)

Scheffé's Table 3a. Relative efficiency of two methods when $k = 4(\alpha = 0.05, \nu = \infty)$.

Type of Contrast	1/R	R
(1, 1)	0·84	
(1, 2)		0·89
(1, 3)		0·79
(2, 2), quadratic		0·59
Linear, cubic		0·74

Scheffé's Table 3b. Relative efficiency when $k = 6(\alpha = 0.05, \nu = \infty)$.

Type of Contrast	1/R	R
(1, 1)	0·73	
(1, 2)	0·98	
(1, 3)		0·91
(1, 4)		0·85
(1, 5)		0·82
(2, 2)		0·68
(2, 3)		0·57
(2, 4)		0·51
(3, 3)		0·45
Linear		0·59
Quadratic		0·57
Cubic		0·48
Quartic		0·53
Quintic		0·67

Tukey's method is superior for contrasts only of the form (1, 1) and for those of the form (1, 1) and form (1, 2) when k = 6 (reading from Scheffé's Table 3b). What can be gleaned from Scheffé's tables is that for pairwise contrasts Tukey's method is preferable while for the more complicated contrasts Scheffé's method is more efficient, and gives shorter intervals. Petrinovich and Hardyck (1969) while not specifically focusing on just Tukey's and Scheffé's methods, nevertheless provide data on the two techniques, enabling us to compare the two. Under the null hypothesis conditions, both techniques control the Type I error as they were designed to do, but the number of Type I errors for Scheffé's technique is consistently less than theoretical alpha (.05); it appears to overprotect. Consistent with this pattern of overprotection for the first kind of error Scheffé's method (Table 2) is found to commit a greater number of Type II errors than does Tukey's method and is as expected from the Type I estimate, generally less powerful. Since Petrinovich and Hardyck limited their study to pairwise contrasts (1, 1), their findings are consistent with Scheffé's analytical results. Specifically then, for contrasts of the form (1, 1), Tukey's technique sets shorter intervals and is more powerful in detecting differences for this type of contrast.

Scheffé (1953) points out that very little is known about how Tukey's T-Method works. Scheffé (1953) has contrasted his method to Tukey's but only under a few specific conditions. Petrinovich and Hardyck (1969) published a long awaited empirical

investigation and evaluation of various multiple comparison procedures. A question should be raised regarding computing post-hoc linear contrasts of the means without first checking to see whether the ANOVA null hypothesis can be rejected or not. When Petrinovich and Hardyck stated that there is disagreement and uncertainty as to the efficacy of using post-hoc techniques without first computing an ANOVA F value they failed to take into consideration the words of H. Scheffé. If Scheffé states that his technique and Tukey's T-Method are to be used following a significant F value then,to this author at least, the question is not so equivocal.

> If the hypothesis is rejected in actual applications of the F-test for equality of means in the one-way layout, the resulting conclusions that the means $\beta_1,\beta_2,\ldots\beta_I$, are not all equal would by itself usually be insufficient to satisfy the experimenter. Methods of making further inferences about the means are then desirable. A similar problem may arise after "the" F-test has been made for any H,. . . Simple answers to the question of what further inferences can be made about the means are offered by the methods of multiple comparisons which we shall call the S-Method and the T-Method (Scheffé, 1959, pg. 66).

The criticism that this auther would put forth concerning Petrinovich and Hardyck's study and conclusions coincides with the comments offered by Reese (1970). This author would add to Reese in stating more emphatically that multiple comparison procedures are to follow a significant ANOVA F value, and therefore perhaps Petrinovich and Hardyck did inadvertently place too much emphasis on the number of Type I errors rather than Type II errors. Consequently,

Scheffé's method need not be considered the most efficacous as Petrinovich and Hardyck insinuated it to be.

In the Petrinovich and Hardyck article the Scheffé and Tukey methods were found to be conservative for Type I errors. As Ryan (1959) and Wilson (1962) point out, and with which this author agrees, it is better to be conservative and stringent with Type I errors than to let alpha get out of hand in terms of an inordinate number of false rejections. In regard to error rates, the experimentwise error rate adequately checks and controls the number of Type I errors in a suitable manner in regards to psychological experimentation (Ryan, 1959, 1963). Tukey's and Scheffé's methods are two techniques which control the number of Type I errors. For a moment consider the probability statement that is made with regard to Scheffé's method: the probability is $1-\alpha$ that all contrasts simultaneously cover their true psi values. This statement is theoretically true for the null case. Petrinovich and Hardyck in the research they reported just investigated pairwise contrasts. What then may happen to the number of Type I errors when some number less than all possible contrasts are performed? Petrinovich and Hardyck give us clues as to what may possibly happen. They found that under the null, when just a subset of all possible contrasts are performed, i.e., pairwise contrasts, the number of Type I errors was generally less than alpha. Petrinovich and Hardyck, though, did not perform all possible contrasts but rather

limited their investigation to pairwise comparisons.

Generally, it is a fine state of affairs to guard against an excessive number of Type I errors and set alpha at a conservative value. But if the number of Type I errors is set at some conservative experimentwise level of significance, it would be imperative that the researcher know whether the actual number of Type I errors is now even more conservative due to the fact that some number less than all possible of the contrasts were investigated. Here then is one of the objectives for the following research: does the number of Type I errors fluctuate for Scheffé's and Tukey's post-hoc methods when certain percentages of all possible contrasts are being explored? That is, for the true zero-difference null hypothesis, for a one-way fixed effects ANOVA, with four levels of the treatment variable, Scheffé's and Tukey's methods will be investigated for the empirical number of Type I errors when all possible, 75 per cent, 50 per cent, 25 per cent, and pairwise contrasts are considered. The frequency of Type I errors for all possible, 75, 50, 25 per cent and pairwise contrasts will also be empirically checked under conditions of assumption violations, e.g., nonnormal populations, unequal variances, and for unequal observations per cell.

Recently, Aitkin (1969) has pointed out that the error rate per comparison and the error rate experimentwise can be calculated exactly rather than estimated via monte carlo sampling procedures.

By using Pearson's (1968) tables of the incomplete beta-function, Aitkin maintains that given an alpha level per comparison the corresponding alpha level experimentwise can be calculated directly, and, vice-versa, given an alpha level experimentwise, the alpha level per comparison is also directly obtainable from the tables. Aitkin cites a case from Petrinovich and Hardyck to demonstrate the use of the incomplete beta function. Petrinovich and Hardyck had found that when the error rate is set experimentwise, for three treatment levels with five observations per level, the comparable error rate per comparison was .013. That is, setting the error rate experimentwise for Scheffé's technique the error rate when counted by a per comparison rule was found to be .013 for Scheffé's test (reading from Petrinovich and Hardyck's Figure 1). Aitkin by using (3) and (4) calculated the exact probability to be .0164.

$$\xi = \left(1 + \frac{t}{\nu}\right)^1 \tag{3}$$

$$\alpha = I_{\xi}\left(\frac{\nu}{2}, \tfrac{1}{2}\right) \tag{4}$$

Here is how Aitkin determined the exact probability. For three groups each containing five observations per cell, the critical value from the F distribution would be 3.89. Adjusting alpha via Scheffé's technique, $\sqrt{(K-1)F_{\nu_1 \nu_2}}$, the adjusted critical value becomes 2.789. That is, if one were to compute multiple t-tests and set the error-rate experimentwise the critical value would be

2.789. Substituting this information into (3) and (4) and referring to the incomplete beta function for $q = \frac{1}{2}$ and $I_{\xi} = 6.01$ the exact probability per comparison would be .0164. The correspondence between the exact probability and the sampling probability found by Petrinovich and Hardyck is good.

Since exact probabilities per comparison can be calculated directly when an error rate is set experimentwise the sampling estimates will be checked for their goodness of correspondence via the incomplete beta function tables. The first minor phase of this research then is concerned with the number of Type I errors.

The second phase of this research will look at the sampling estimates of the relative frequency of Type II errors. This second phase is of the utmost concern and importance for it is my contention that for post-hoc multiple comparison procedures, the relative number of Type II errors is of paramount importance. If post-hoc multiple comparison procedures are to follow a significant ANOVA F test, it seems that the researcher should protect the possible number of false acceptances he is willing to commit rather than controlling too carefully the number of false rejections. A significant ANOVA F test is indicative of general mean differences and consequently treatment effects. Since the treatment effects are there, it seems logical that the post-hoc comparison procedures should be adjusted to find those differences which in the last breath were stated to be existent. The post-hoc multiple comparison

procedures should be adjusted, because as Petrinovich and Hardyck (1969), Aitkin (1969) and Scheffé (1959) imply, these techniques, Scheffé's and Tukey's in particular, are subject to an excessive number of Type II errors and consequently lack any substantial power. Scheffé states . . . "Robustness for Type I errors is not a sufficient recommendation for a test; the power must also be considered against some alternatives of interest (Scheffé, 1959, pg. 361). Again, the previous authors insinuate that the excessive number of Type II errors is not unalterably a function of the post-hoc comparison procedures but rather appears to be also a function of the stringent experimentwise error rate. Petrinovich and Hardyck's Table 2 really drives home the relative frequency of Type II errors and their dependence on the error rates. For their small sample condition (n = 5) the number of beta errors does not reach a respectable level (approximately .05) and is as large as .995 and even for 2.6 σ-unit differences between adjoining means the number of Type II errors is .171. When the sample size per treatment level is increased to fifteen, 1.6 σ-unit differences are required to bring the probability of a Type II error down to approximately .05. Even when the error rate is set per comparison the number of Type II errors is still somewhat excessive.

Consider this: if you, the experimenter, were interested in detecting 1.2 σ-unit differences between any two of your population means when k = 3 (this corresponds to Petrinovich and Hardyck's case of .6 σ-unit differences between adjoining means) referring

to the phi tables you would calculate that 1.2 σ-unit differences should be detected $\simeq$97 per cent of the time with five observations per treatment level. Therefore, for a one-way ANOVA with three treatment levels, five observations per cell, you, the experimenter should reject the ANOVA hypothesis, H_o: $\mu_1 = \mu_2 = \mu_3 = 0$, approximately 97 per cent of the time. Having rejected the null hypothesis for the ANOVA F test you are now interested in probing your data to determine which contrasts of the means had lead to the rejection of the ANOVA null hypothesis. You decide to do three pairwise contrasts among your three means, e.g.,

$\hat{\psi}_1 = (+1)\overline{X}._1 + (-1)\overline{X}._2$, $\hat{\psi}_2 = (+1)\overline{X}._1 + (-1)\overline{X}._3$ and $\hat{\psi}_3 = (+1)\overline{X}._2 + (-1)\overline{X}._3$.

Keeping in mind that initially you had "built-in" a power of approximately $\simeq$.97, when there are five observations per treatment cell, Petrinovich and Hardyck's results show that for your post-hoc comparisons on the three pairwise means, the power $(1-\beta)$ of detecting differences among the mean is $\simeq$.005 for Scheffé's method, and $\simeq$.007 for Tukey's method. Contrary to what might be expected, the power that was so called "built-into" the ANOVA F test does not carry over to your post-hoc comparisons, according to the results of Petrinovich and Hardyck. Accordingly, these authors state . . . "if sample size is less than 10, it scarcely seems worthwhile to carry out the computations for multiple comparisons since the power of any method to detect differences between small groups is extremely low" (p. 53). Aitkin (1969) and Scheffé (1959) suggest that the insensitivity of post-hoc techniques which control

the error rate experimentwise is a function of the dependence that researchers have for the .05 and .01 levels of significance. That is, the insensitivity (lack of power) of post-hoc techniques which control the Type I error experimentwise may be adjusted if alpha is set at some other level than the conventional 5 and 1 per cent levels. The third major phase of this study therefore will examine the Type II errors and consequently, the power of Scheffé's and Tukey's technique for different levels of alpha. Also, the data will be examined for the different error rates, for, perhaps when investigating linear contrasts in the post-hoc sense, setting the error rate experimentwise will prove to be untenable in regards to the number of Type II errors and consequently the error rate per comparison may prove to be more suitable to the purpose and rationale of post-hoc comparisons---finding the significant differences.

References

Aitkin, M. A. Multiple comparisons in psychological experiments. The British Journal of Mathematical and Statistical Psychology, 1959, 22, 193-198.

Boneau, C. A. The effects of violations of assumptions underlying the t-test. Psychological Bulletin, 1960, 57, 49-64.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. I, Effect of inequality of variances in the one-way classification. Annals of Mathematical Statistics, 1954a, 25, 290-302.

Box, G. E. P. 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, 1954b, 25, 484-498.

Box, G. E. P., and Anderson, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, Series B, 1955, 17, 1-26.

Cochran, William, G. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 1947, 3, 22-38.

Davis, Daniel, J. Flexibility and power in comparisons among means. Psychological Bulletin, 1969, 71, 441-444.

Dunnett, C. W. Multiple comparison tests. Biometrics, 1970, 26, 139-141.

Edwards, A. L. Experimental Design in Psychological Research. New York: Holt, Rinehart, & Winston, 1960.

Glass, Gene, V., & Stanley, Julian, C. Statistical Methods in Education and Psychology. New Jersey: Prentice-Hall, Inc., 1970.

Godard, R. H., and Lindquist, E. F. An empirical study of the effect of heterogeneous within-groups variance upon certain F-tests of significance in analysis of variance. Psychometrika, 1940, 5, 263-274.

Hays, William, L. Statistics. New York: Holt, Rinehart, & Winston, 1963.

Horsnell, G. The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 1953, 40, 128-136.

Hsu, P. L. Contribution to the theory of student's t-test as applied to the problem of two samples. Statistical Research Memoirs, Vol. 2, (353, 355, 356, 364), p. 1-24.

Kendall, M. A., & Stuart, Alan. The Advanced Theory of Statistics. Vol. 1. New York: Hafner Publishing Company, 1969.

Kirk, Roger, E. Experimental Design: Procedures for the Behavioral Sciences. California: Brooks/Cole Publishing Company, 1968.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin, 1953.

McNemar, Q. Psychological Statistics (3rd ed.), New York: Wiley & Sons, 1962.

Miller, Rupert, G., Jr. Simultaneous Statistical Inference. New York: McGraw-Hill Book Company, 1966.

Petrinovich, Lewis, F., and Hardyck, Curtis, D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

Reese, Hayne, W. Multiple comparison methods. American Psychologist, 1970, 25, 365-366.

Ryan, Thomas, A. Comments on orthogonal components. Psychological Bulletin, 1959, 56, 394-305 (a).

Ryan, Thomas, A. Multiple comparisons in psychological research. Psychological Bulletin, 1959, 56, 26-47.

Ryan, Thomas, A. The experiment as the unit for computing rates of error. Psychological Bulletin, 1962, 59, 301-305.

Scheffé, Henry. A method for judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheffé, Henry. The Analysis of Variance. New York: John Wiley & Sons, Inc., 1959.

Smith, Robert, A. The effect of unequal group size on Tukey's HSD procedure. Psychometrika, 1971, 36, 31-34.

Steel, R. G. D. Error rates in multiple comparisons. Biometrics, 1961, 17, 326-328.

Stanley, Julian, C. Additional "post-mortem" tests of experimental comparisons. Psychological Bulletin, 1957, 54, 128-130.

Welch, B. L. The significance of the difference between two means when the population variances are unequal. Biometrika, 1937, 29, 350-362.

Wilson, Warner. A note on the inconsistency, inherent in the necessity to perform multiple comparisons. Psychological Bulletin, 1962, 59, 296-300.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1962.

Footnotes

1 $E(\hat{\alpha}_1) = E(\bar{X}._1 - \bar{X}..) = E(\bar{X}._1) - E(\bar{X}..) = \mu_1 - \mu = \alpha_1$

2 $E(\hat{\alpha}_1) = \mu_1 - \mu$, similarly $\alpha_2 = \mu_2 - \mu$, $\alpha_3 = \mu_3 - \mu$, and $\alpha_4 = \mu_4 - \mu$. The null hypothesis can therefore be written as $(\mu_1 - \mu) = (\mu_2 - \mu) = (\mu_3 - \mu) = (\mu_4 - \mu) = 0$. Adding the constant $\mu..$ to each term gives $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$.

3 $$E(\hat{\Psi}) = E\left(\sum_{k=1}^{K} C_K \bar{X}._K\right) = \sum_{k=1}^{K} C_K E(\bar{X}._K) = \sum_{k=1}^{K} C_K \mu = \Psi$$

4 Six contrasts of the form (1, 1), twelve contrasts of the form (1, 2), three contrasts of the form (2, 2), and four contrasts of the form (1, 3).

5 Any two comparisons are independent if:

$$\sum_{k=1}^{K} C_K C_K'/n_K = 0$$

Independence of a comparison and the grand mean:

$$\bar{X}.. = \frac{\sum\sum^{IK} X_{ik}}{N} = \frac{\sum_i X_{i1} + \sum_i X_{i2} + \ldots + \sum_i X_{ik}}{n_1 + n_2 + \ldots + n_K} = \frac{n_1\bar{X}._1 + n_2\bar{X}._2 + \ldots + n_K\bar{X}._K}{n_1 + n_2 + \ldots + n_K}$$

$$= \frac{n_1\bar{X}._1}{\Sigma n_K} + \frac{n_2\bar{X}._2}{\Sigma n_K} + \ldots + \frac{n_K\bar{X}._K}{\Sigma n_K}$$

Now for any contrast: $\hat{\Psi} = C_1\bar{X}._1 + C_2\bar{X}._2 + \ldots + C_K\bar{X}._K$

Remembering the requirement for independence, that is,

$\sum_{k=1}^{K} C_K C_K'/n_K = 0$ we have $\sum^{K} [(n_K/\Sigma n_K)](C_K)]/n_K$

Multiplying numerator and denominator by $1/n_K$ we have:

$\sum_K C_K/n_K = \sum_K C_K/N = 1/N \sum_K C_K = 0$, thereby satisfying the requirement of independence.

6 $\bar{X}_{.K}$ is a linear combination of the X_i observations.

$\bar{X}_{.K} = 1/n_1 X_1 + 1/n_2 X_2 + \ldots + 1/n_K X_K$

The variance of $\bar{X}_{.K} = C_1^2$ var. $(\bar{X}_{.})$ + C_2^2 var. $(\bar{X}_{.})$ + ... + C_K^2 var. $(\bar{X}_{.K})$

The variance of the sampling distribution of means is σ^2/N, therefore

var. $(\hat{\Psi}) = C_1^2 (\sigma^2/n_1) + C_2^2 (\sigma^2/n_2) +$ ... $+ C_K^2 (\sigma^2/n_K) = \sum_K C_K^2(\sigma^2/n_K)$. Assuming equal variances we have $\sigma^2 \sum_K C_K^2/n_K$. Now est. σ^2 = MS error therefore est. var. $(\hat{\Psi})$ = MS error $\sum_K C_K^2/n_K$.

7 $\mu_1 = 0$, $\mu_2 = 2.6$, $\mu_3 = 5.2$, and $\mu_4 = 7.8$. Since $\alpha_1 = \mu_1 - \mu_K =$ -3.9, $\alpha_2 = 1.3$, $\alpha_3 = 1.3$, and $\alpha_4 = 3.9$. $\sum_K \alpha_K^2 = 33.80$

Therefore $\phi = \dfrac{I \sum_K \alpha_K^2}{J \sigma_e^2} = I\ (8.45)$

When I is chosen to be 11, $\phi = 9.64$, for $\nu_1 = 3$, $\nu_2 = 40$ the power is approximately .9999.