EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1974, 34, 511–519.

COMPARISON OF TUKEY'S T-METHOD AND SCHEFFÉ'S S-METHOD FOR VARIOUS NUMBERS OF ALL POSSIBLE DIFFERENCES OF AVERAGES CONTRASTS UNDER VIOLATION OF ASSUMPTIONS[1]

H. J. KESELMAN[2]
University of Manitoba

LARRY E. TOOTHAKER
University of Oklahoma

Empirical .05 and .01 rates of Type I error were compared for the Tukey and Scheffé multiple comparison techniques. The experimentwise error rate was defined over five sets of the all possible 25 differences of averages contrasts. The robustness of the Tukey and Scheffé statistics was not only related to the type of assumption violation, but also to the sets containing different numbers of contrasts. The Tukey method could be judged as robust a statistic as the Scheffé method.

THE analysis of variance (ANOVA) is often used to test whether sample means are indicative of experimental treatment effects or merely chance variation. Since the ANOVA only signifies the presence of overall treatment effects, experimenters usually follow a significant F-test with a multiple comparison statistic, enabling the researcher to locate the specific mean differences which have caused the ANOVA F-test to be significant (Scheffé, 1959).

From the arsenal of multiple comparison statistical techniques,

[1] The authors wish to express their appreciation to David Koulack for his comments on an earlier draft of the manuscript.

[2] Requests for reprints should be sent to H. J. Keselman, Department of Psychology, University of Manitoba, Winnipeg, Canada. The .01 estimates were collected with the support provided by The Canada Council (Grant No. S72-0533) and The Research Board of The University of Manitoba (Grant No. 431-1665-15).

the Tukey and Scheffé methods are very often used by psychological researchers. Tukey's technique, the T-method, utilizes the distribution of the studentized range (q_{k,ν_2}), the error degrees of freedom (ν_2), and the mean square error (MS_e) from the ANOVA to investigate differences among means, following the rejection of the ANOVA null hypothesis. Scheffé (1959, p. 74) wrote that for Tukey's T-method the probability is $1 - \alpha$ that the values of all contrasts $\Psi = \Sigma c_k \mu_k$ simultaneously satisfy (1)

$$\hat{\psi} - q_{\alpha;k,\nu_2}(MSe/n)^{1/2}(\tfrac{1}{2}\Sigma\ |c_k|) \leq \Psi \leq \hat{\psi} + q_{\alpha;k,\nu_2}(MSe/n)^{1/2}(\tfrac{1}{2}\Sigma\ |c_k|), \qquad (1)$$

where μ_k is the mean of $k = 1, \cdots, K$ independent samples of $i = 1, \cdots, n$ independently normally distributed random X_{ik} variables, c_k is a weight assigned to the parameters μ_k, $\Psi = \Sigma c_k \mu_k$ and $\hat{\psi} = \Sigma c_k \bar{X}_{\cdot k}$ are linear combinations of population and sample means, respectively, for which $\Sigma c_k = 0$, n specifies equal numbers of observations for the k samples, and $\bar{X}_{\cdot k} = \Sigma X_{ik}/n$. The Tukey method was derived under the restriction that the variances of the means are equal, and therefore the number of observations per sample, n, must be equal. In repeated experiments the probability is $1 - \alpha$ that all intervals simultaneously cover the true values of the population contrasts.

To circumvent the limited applicability of Tukey's T-method, Scheffé (1953, 1959) formulated the S-method which is a generalized version of Tukey's method, but uses the sampling distribution of F. For all possible contrasts of the form $\hat{\psi} = c\bar{X}_{\cdot 1} + c\bar{X}_{\cdot 2} + \cdots + c_K\bar{X}_{\cdot K}$ the probability is $1 - \alpha$ that all contrasts on K means simultaneously satisfy the relationship in (2),

$$\hat{\psi} - [(K-1)F_{\nu_1,\nu_2}]^{1/2}[MS_e\Sigma c_k{}^2/n_k]^{1/2} \leq \Psi \leq \hat{\psi} + [(K-1)F_{\nu_1,\nu_2}]^{1/2}[MS_e\Sigma c_k{}^2/n_k]^{1/2}, \qquad (2)$$

where F_{ν_1,ν_2} is the tabled α F-value with ν_1 and ν_2 degrees of freedom, c_k and n_k are the weight and sample size for the kth sample, and Ψ and $\hat{\psi}$ are defined as in (1). Scheffé's S-method is not dependent upon equal variances of the means nor consequently upon equal sample sizes for its validity, and is robust to non-normality and heterogeneity of variance (Scheffé, 1959, p. 77).

There have been few comparisons between Tukey's T-method and Scheffé's S-method. Miller (1966) pointed out that most of the

work dealing with the effects of departures from assumptions has focused on statistical techniques utilizing the F distribution. Therefore, the robustness of the ANOVA F test and the Scheffé S-method, which utilize the F distribution, is well documented (Scheffé, 1959). However, the evidence concerning the robustness of the Tukey multiple comparison statistic, which utilizes the Studentized Range distribution, is limited (Petrinovich and Hardyck, 1969; Smith, 1971).

Although Miller (1966) suggested that Tukey's T-method would be more sensitive to assumption violations than Scheffé S-method, no studies have examined this possibility when sampling from a skewed distribution.

Smith (1971) examined the robustness of Tukey's multiple comparison statistic to Type I errors with groups of unequal sample size using the following procedures: (1) the harmonic mean of the group sizes, (2) the sample sizes of the two groups comprising the linear comparison, and (3) the average value of the groups sizes. He found that methods (1) and (2) yielded empirical estimates that were more consistently congruent with the theoretical alpha values than did the average size approximation.

Petrinovich and Hardyck (1969) investigated the robustness of the Tukey and Scheffé methods for varied population shapes, population variance, for differing numbers of treatment levels and sample sizes, and concluded that multiple comparison procedures, like the ANOVA F test, suffer from the same lack of robustness for certain combinations of unequal n's and unequal variances. Similarly, the empirical probabilities were affected when sampling from exponential distributions.

Petrinovich and Hardyck's (1969) comparisons of the Tukey and Scheffé methods were only for pairwise contrasts however, and therefore do not provide a just base for comparison. The Scheffé method was derived to control alpha for all types of contrast, while the Tukey method as used by Petrinovich and Hardyck, was intended only for pairwise contrasts and as expected, Petrinovich and Hardyck found the Scheffé technique overly conservative. It is evident that since the probability of a Type I error would be related to the number of contrasts investigated, the results of comparing the Tukey and Scheffé methods are biased and misleading.

Generally it would be useful for the researcher if he knew the extent to which the probability of a Type I error could deviate from

the theoretical alpha as a function of the number of contrasts that were computed.

The purpose of the present study was to investigate the Tukey and Scheffé methods for empirical probabilities of a Type I error under violation of assumptions. For unequal numbers of observations per treatment cell and for unequal population variances, these methods were compared when sampling from a normal and a skewed population for various numbers of comparisons.

Procedure

Pseudo-random numbers were selected, using a pseudo-random number generator and distributed to the four treatment levels of a one-way fixed effects analysis of variance.

The observations from the normal distribution were generated by means of GAUSS (IBM, 1967), which generates pseudo-random normal deviates with $\mu_1 = 0$ and $\mu_2 = \sigma_2^2 = 1$. The skewed population was derived from a chi-square distribution with three degrees of freedom, and hence with moments $\mu_1 = 3$, $\mu_2 = \sigma^2 = 6$, $\mu_3 = 24$, $\mu_4 = 252$, a skewness measure $\gamma_1 = 1.663$, and a kurtosis measure $\gamma_2 = 4$. The 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 a 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.

While there are an infinite number of contrasts on K means, many of these contrasts are of no interest to the typical experimenter. Of those that might be of interest, the group which will be called differences of averages (DA) was used in the present research. Six differences between pairs of means (pairwise contrasts, or $\Psi = \mu_2 - \mu_1$) are included in the group of DA contrasts in addition to 12 which compare one mean to the average of two means ($\Psi = \mu_1 - \frac{1}{2}[\mu_2 + \mu_3]$), 4 comparing one mean to the average of three means ($\Psi = \mu_1 - \frac{1}{3}[\mu_2 + \mu_3 + \mu_4]$) and 3 comparing the average of two means to the average of two means ($\Psi = [\mu_1 + \mu_2] - [\mu_3 + \mu_4]$). Other contrasts such as $\Psi = -3\mu_1 - 1\mu_2 + 1\mu_3 + 3\mu_4$ might be of in-

terest to a researcher, but were not included in the present research.

Given four levels of the treatment variable there were thus 25 possible DA contrasts that could be computed from the data. Tukey's T-method and Scheffé S-method statistics were calculated in order to determine the number of contrasts which bracketed zero for each of the five sets of DA contrasts. Set I defined the experimentwise error rate over all of the 25 DA contrasts, while Sets II, III, and IV contained 18 (72%), 12 (48%), and 6 (24%) contrasts, respectively, that were randomly selected from Set I. For Set V the experimentwise error rate was defined over the six simple pairwise comparisons that are included in Set I. The procedure of generating four random samples with n_k observations for the kth sample and calculating Tukey's and Scheffé's multiple comparison procedures constituted one experiment; the procedure was repeated 1,000 times for each of the two populations and the sets of sample sizes to be described later.

The five combinations examined when sampling from a normal distribution were (A) $n_1 = n_2 = n_3 = n_4$; $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, (B) $n_1 = n_2 = n_3 = n_4$; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$, (C) $n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, (D) $n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$ (positively related to sample size) and (E) $n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$ (negatively related to sample size). These five conditions were also investigated for the non-normal skewed population. For comparisons involving unequal variances, the variances chosen were .4, .8, 1.2, and 1.6. The harmonic mean was used with the Tukey method when there were an unequal number of observations per cell, and for the samples of sizes 8, 9, 11, and 16 was 10.269. The sample sizes and unequal variances are enumerated in Table 1.

TABLE 1

Samples Sizes and Variances

Sample Sizes:	11	11	11	11	Harmonic Mean 11		8	9	11	16	Harmonic Mean 10.269
	A	1	1	1	1		C	1	1	1	1
σ^2							D	.4	.8	1.2	1.6
	B	.4	.8	1.2	1.6						
							E	1.6	1.2	.8	.4

Results

Normal Distribution

A. *Equal n's (11); Equal σ^2's (1):*

For the 25 DA contrasts (Set I) used in this study 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 protect alpha for all possible pairwise and complex type contrasts. Tukey's method however, was originally derived to control the probability of a Type I error for pairwise contrasts (Scheffé, 1959) and therefore the Type I empirical estimates should be consistent with alpha under Set V. The Scheffé Type I estimates though, should be less than alpha for Set V since the pairwise contrasts are a small subset of the all possible 25 DA contrasts. The tabled probabilities for condition A do conform to theory (Table 2).

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

Variance heterogeneity does not affect the multiple comparison Type I probabilities. The Scheffé .05 Set I estimate of .046 and the Tukey Set V value of .052 are in accord with theoretical alpha. These multiple comparison statistics, like the ANOVA F test, are not substantially affected by variance heterogeneity when there are an equal number of observations per cell. The S and T tests are also robust for the .01 level of significance.

C. *Unequal n's (8, 9, 11, 16), Equal σ^2's (1):*

The Scheffé .05 empirical probabilities are generally invariant from the estimates for conditions A and B, while the Tukey values are larger. The probability of a Type I error for the S-method is in agreement with .05 for all 25 DA contrasts, but the estimates progressively deviate from .05 for the remaining sets of contrasts. The probability statement for the Tukey statistic stipulates that the number of observations per cell must be equal. The effect of violating this restriction is reflected in the .05 Set I and Set V estimates which are 1 σ greater than theoretical alpha. Violating this restriction though does not affect the .01 estimates.

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

For this condition samples of unequal size were sampled from normal populations having unequal variances. The sample sizes

TABLE 2

Monte Carlo Type I Experimentwise Errors for the Scheffé (S) and Tukey (T) Statistics

			*	Normal Distribution					Skewed Distribution				
				A	B	C	D	E	A	B	C	D	E
$\alpha = .05$	S	Set I		.051	.046	.048	.030	.081	.041	.043	.031	.032	.083
		Set II		.047	.043	.044	.030	.076	.038	.039	.031	.031	.076
		Set III		.039	.039	.037	.023	.062	.034	.036	.026	.023	.066
		Set IV		.025	.027	.030	.017	.046	.026	.026	.014	.019	.043
		Set V		.034	.031	.031	.023	.061	.029	.036	.020	.025	.052
	T	Set I		.053	.052	.060	.030	.110	.041	.050	.032	.030	.103
		Set II		.045	.044	.050	.028	.094	.036	.042	.029	.030	.085
		Set III		.033	.035	.038	.022	.077	.029	.036	.024	.016	.073
		Set IV		.021	.026	.026	.011	.054	.023	.024	.014	.012	.049
		Set V		.053	.052	.060	.030	.110	.041	.050	.032	.030	.103
$\alpha = .01$	S	Set I		.012	.012	.009	.004	.022	.007	.008	.008	.011	.019
		Set II		.012	.012	.009	.004	.021	.007	.007	.008	.010	.018
		Set III		.009	.012	.008	.004	.017	.005	.005	.008	.008	.013
		Set IV		.006	.009	.003	.002	.011	.002	.003	.004	.008	.007
		Set V		.011	.011	.004	.003	.015	.005	.003	.006	.008	.012
	T	Set I		.014	.017	.009	.003	.036	.008	.007	.009	.009	.032
		Set II		.012	.016	.008	.003	.032	.007	.005	.009	.008	.026
		Set III		.011	.013	.009	.003	.021	.005	.002	.008	.006	.019
		Set IV		.006	.005	.005	.002	.015	.002	.003	.008	.005	.012
		Set V		.014	.017	.009	.003	.036	.008	.007	.009	.009	.032

* (A) $n_1 = n_2 = n_3 = n_4$; $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, (B) $n_1 = n_2 = n_3 = n_4$; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$, (C) $n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, (D) $n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$ (postively related to sample size) and (E) $n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$ (negatively related to sample size).

Sets: (I) error rate defined over all of the 25 DA contrasts (II) 18 contrasts randomly selected from Set I (III) 12 contrasts randomly selected from Set I (IV) 6 contrasts randomly selected from Set I (V) the 6 simple pairwise comparisons.

were positively related to the variances such that 8, 9, 11, and 16 cases were sampled from the populations having variance .4, 1.8, 1.2, and 1.6, respectively. Consistent with previous findings (Box, 1954 a,b; Box and Anderson, 1955); the empirical probabilities of a Type I error are less than theoretical alpha, when positively pairing unequal variances and unequal sample sizes.

The Scheffé and Tukey .05 and .01 estimates are also very similar to one another for all five sets of DA contrasts.

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

For the negatively related pairings of unequal variances and unequal sample sizes, the empirical probabilities for the S-method

exceed alpha, particularly the .05 estimates, as has been found with the ANOVA F test (Box, 1954 a,b). Even for the pairwise contrasts, the probability of a Type I error is larger than alpha. The Tukey pairwise estimates are much larger than .05 and .01, even more so than the values for the Scheffé method.

Skewed Distribution

The empirical probabilities of a Type I error when sampling from the skewed distribution are also enumerated in Table 2 for the same five conditions that were investigated when sampling from a normal distribution. The degree of correspondence of the Scheffé and Tukey empirical probabilities to one another and the variability of each statistic's estimates as a function of the number of contrasts computed, is similar to the normal distribution estimates for all five assumption violating conditions. The empirical probabilities for conditions A–D though are generally less than the normal distribution estimates, whereas the condition E estimates are quite similar to the normal distribution probabilities.

Discussion

Though the robustness of a statistic may connote something different to each reader, the authors believe that for the user of the statistic the *absolute* deviation of empirical alpha from the size of the statistic set by the experimenter could be crucial. For the nine assumption violating conditions, the Scheffé statistic is closer to the alpha size of .05 for four of the conditions (Normal distribution: C, E; Skewed distribution: D, E) when comparing the Scheffé Set I values with the Tukey Set V estimates. For this same comparison, at the .01 level, the Scheffé test is in closer agreement with alpha for five conditions (Normal distribution: B, D, E; Skewed distribution: B, E). When considering just pairwise .05 estimates, the Tukey statistic is closer for seven of the nine population conditions (Normal distribution: B, C, D; Skewed distribution: A, B, C, D), while closer for five (Normal distribution: C; Skewed distribution: A, B, C, D) at the .01 level. From the data it could be concluded that the Tukey method is as robust a statistic as the Scheffé method. It is also apparent that the empirical probability of a Type I error varies with the different number of contrasts computed and as hypothesized, the empirical estimates for the Scheffé method for the pair-

wise contrasts were less than the estimates for the all 25 DA contrasts.

REFERENCES

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

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,* 1954, 25, 484–498. (b)

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 B,* 1955, 17, 1–34.

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

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences.* California: Brooks/Cole Publishing Co., 1968.

Miller, R. G., Jr. *Simultaneous statistical inference.* New York: McGraw-Hill Book Co., 1966.

Petrinovich, L. R. and Hardyck, C. D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin,* 1969, 71, 43–54.

Ryan, T. A. Multiple comparisons in psychological research. *Psychological Bulletin,* 1959, 56, 26–47.

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

Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika,* 1953, 40, 87–104.

Scheffé, H. *The analysis of variance.* New York: John Wiley & Sons, 1959.

Smith, R. 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, 1953, Princeton University.