Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 10

11-1-2005

Type I Error Of Four Pairwise Mean Comparison Procedures Conducted As Protected And Unprotected Tests

J. Jackson Barnette
University of Alabama at Birmingham, barnette@uab.edu

James E. McLean
University of Alabama, Tuscaloosa

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 446-459

1538 – 9472/05/$95.00

Type I Error Of Four Pairwise Mean Comparison Procedures Conducted As Protected And Unprotected Tests

J. Jackson Barnette
Department of Biostatistics
University of Alabama at Birmingham

James E. McLean
Program of Educational Research
University of Alabama, Tuscaloosa

Type I error control accuracy of four commonly used pairwise mean comparison procedures, conducted as protected or unprotected tests, is examined. If error control philosophy is experimentwise, Tukey's HSD, as an unprotected test, is most accurate and if philosophy is per-experiment, Dunn-Bonferroni, conducted as an unprotected test, is most accurate.

Key words: Type I error control, experimentwise vs. per-experiment error, protected vs. unprotected tests, pairwise comparisons, Tukey's HSD, Dunn-Bonferroni, Dunn-Sidak, Holm's sequentially rejective

Introduction

Whenever a researcher has more than two comparisons to test, control of the Type I error-rate becomes a concern. Soon after Fisher developed the process of analysis of variance (ANOVA), he recognized the potential problem of the error-rate becoming inflated when multiple *t* tests were performed on three or more groups.

He discussed this problem in the 1935 edition of his famous book, *The Design of Experiments*. His recommendation of using a more stringent alpha when performing his Least Significant Difference Procedure (LSD) is based on this concern. However, researchers still criticized the LSD as providing inadequate control of Type I error. This early recognition of the problem has resulted in hundreds of multiple comparison procedures being developed over the years.

J. Jackson Barnette is Senior Associate Dean for Academic Affairs and Professor of Biostatistics in the School of Public Health at the University of Alabama at Birmingham. Email: barnette@uab.edu. James E. McLean is University Professor and Dean in the College of Education at The University of Alabama, Tuscaloosa.

The earliest example of what is now known as a multiple comparison procedure could be found in 1929, when Working and Hotelling applied simultaneous confidence intervals to regression lines. The Fisher (1935) reference cited earlier was the first application to the process of ANOVA. The Type I error-rate control problem was also referred to by Pearson and Sekar in 1936 and Newman in 1939. Newman described a multiple comparison test that used the "Studentized Range Statistic." It is said that his work was prompted by a discussion he had with Student. Years later, Keuls published an updated version of the procedure (1952) using the Studentized range. That multiple comparison procedure is now known as the Student-Newman-Keuls procedure.

Most studies of Type I error rates for follow-up of pairwise mean differences have been based on what is referred to as experimentwise or familywise error control philosophies. These terms were more extensively described by Ryan (1959) and Miller (1966). Experimentwise (EW) Type I error relates to finding at least one significant difference by chance for the specified alpha level. In these cases, the only difference of concern is the largest mean difference. Experimentwise Type I error control ignores the possibility of multiple Type I errors in the same experiment. The pairwise mean differences for

those other than the largest mean difference are not considered. Type I error control is such that not all possible Type I errors are evaluated. In these cases, many procedures such as Tukey's HSD are considered to have conservative Type I error control since the actual probabilities of finding at least one Type I error are lower than the nominal alpha level.

Per-experiment (PE) Type I error control considers all the possible Type I errors that can occur in a given experiment. Thus, more than one Type I error per experiment is possible and reasonably likely to occur if there is an experimentwise Type I error on the highest mean difference. Klockars & Hancock (1994) pointed out the importance and risks associated with this distinction. They found, using a Monte Carlo simulation, that there was a difference of .0132 in the per-experiment and experimentwise Type I error rates for Tukey's HSD when alpha was set at .05. This discussion was expanded in their 1996 review titled "The Quest for α" (Hancock & Klockars). Thus, when one has exact control of Type I error in the experimentwise situation, the per-experiment Type I error probability is higher. One of the purposes of this research was to examine how much of a difference there may be between experimentwise and per-experiment Type I error rates for four of the most commonly used pairwise multiple comparison procedures when used with alpha levels of .10, .05, and .01, and to determine the relative influence on this difference of number of groups and number of subjects per group. While most Type I error research is based on an experimentwise mode, the per-experiment Type I error is more consistent with the reality of pairwise hypothesis testing. It considers not only the largest mean difference subjected to error control, but all the pairwise differences.

There seems to be an inconsistency of logic when comparing the power of various methods and manners of Type I error control. When it is stated that the Student-Newman-Keuls is more powerful than Tukey's HSD or Holm's procedure is more powerful than Dunn-Bonferroni; the notion is that one method leads to more rejections of partial null hypotheses. However, if one considers the notion of experimentwise Type I error (the largest pairwise difference or more being rejected), then SNK and HSD have the same power and Dunn-Bonferroni and Holm have the same power. Differences in power only come when considering pairwise differences that are found beyond the k number of means steps. Thus, should not error rate take into account the possible false rejections in the entire structure of mean differences, not just the largest one? Per-experiment Type I error control is more consistent with actual pairwise hypothesis decision-making.

Four multiple comparison procedures were selected for this research: Dunn-Bonferroni, Dunn-Sidak, Holm's sequentially rejective, and Tukey's HSD. Based on a review of current literature and commonly used statistical texts, it was concluded that these are among the most frequently used pairwise procedures and represent a variety of approaches to control for Type I error. Since the names of these procedures tend to vary slightly in texts, statistical software, and in the literature, each is described briefly below:

The Dunn-Bonferroni procedure uses the Bonferroni inequality ($\alpha_{PE} \leq \Sigma\alpha_{PC}$) as authority to divide equally the total a priori error among the number of tests to be completed, often following the application of the Fisher LSD procedure. The LSD procedure is equivalent to conducting all pairwise comparisons using independent *t* tests with the MS_{error} as the common pooled variance estimate (Kirk, 1982). An example of the application of the Dunn-Bonferroni would be identifying the a priori α as .05 where tests are required to compare means of five groups using 10 comparisons, running each individual test at the .05/10= .005 level (Hays, 1988). Sidak's modification of the Dunn-Bonferroni procedure, referred to as the Dunn-Sidak procedure substituted the multiplicative computation of the exact error-rate, $\alpha_{PE} = 1 - (1 - \alpha_{PC})^c$ where c is the number of comparisons for the Bonferroni Inequality ($\alpha_{PE} \leq \Sigma\alpha_{PC}$), otherwise following the same procedures (Kirk, 1982).

A procedure proposed by Holm in 1979, Holm's Sequentially Rejective procedure is also referred to as the Sequentially Rejective Bonferroni procedure. Assuming a maximum of

c comparisons to be performed, the first null hypothesis is tested at the α/c level. If the test is significant, the second null hypothesis is tested at the $\alpha/(c - 1)$ level. If this is significant, the testing continues in a similar manner until all c tests have been completed or until a nonsignificant test is run. The testing stops when the first nonsignificant test is encountered (Hancock & Klockars, 1996).

Tukey's Honestly Significant Difference procedure (HSD) was presented originally in a non-published paper by Tukey in 1953. Its popularity has grown to the point where it is, possibly, the most widely used multiple comparison procedure. The HSD is based on the Studentized Range Statistic originally derived by Gossett (a.k.a., Student) (1907-1938). This statistic, unlike the *t* statistic, takes into account the number of means being compared, adjusting for the total number of tests to make all pairwise comparisons (Kennedy & Bush, 1985).

Many researchers follow the practice of conducting post-hoc pairwise multiple comparisons only after a significant omnibus *F* test. Protected tests are conducted only after a significant omnibus *F* test, while unprotected tests are conducted without regard to the significance of the omnibus *F* test. Many common statistical texts either recommend or imply the use of a protected test for all post-hoc multiple comparison procedures (e.g., Hays, 1988; Kennedy & Bush, 1985; Kirk, 1982; Maxwell & Delaney, 1990). While these texts provide a logical basis for this, and excellent reviews of multiple comparison procedures are available (e.g., Hancock & Klockars, 1996; Toothaker, 1993), little empirical evidence is presented, either analytically or empirically, to justify this practice.

The research questions addressed in this research are:

1. Which of these four multiple comparison procedures has the most accurate control of Type I error across the three alpha conditions?

2. Does error control accuracy differ when tests are conducted as protected or unprotected tests?

3. Do methods differ relative to accuracy when conducted as experimentwise vs. per-experiment control?

Methodology

Monte Carlo methods were used to generate the data for this research. All data comprising the groups whose means were compared were generated from a random normal deviate routine, which was incorporated into a larger compiled QBASIC program that conducted all needed computations. The program was written by the senior author. All sampling and computation, conducted with double-precision, routines were verified using SAS® programs. Final analysis of the summary statistics and correlations was conducted using SAS®.

Several sample size and number of groups arrangements were selected to give a range of low, moderate, and large case situations. The numbers of groups were: 3, 4, 5, 6, 8, and 10 and the sample sizes for each group were: 5, 10, 15, 20, 30, 60, and 100, which when crossed gave 42 experimental conditions. This was replicated for three nominal alphas of .10, .05, and .01. The approach used was to determine what number of replications would be needed to provide an expected .95 confidence interval of +/- .001 around the nominal alpha.

This is an approach to examination of how well observed Type I error proportions are reasonable estimates of a standard nominal alpha. In other words, if alpha is the standard, what proportion of the estimates of actual Type I error proportions can be considered accurate, as evidenced by them being within the expected .95 confidence interval around nominal alpha?

This was based on the assumption that errors would be normally distributed around the binomial proportion represented by nominal alpha. Thus, when alpha was .10, 345742 replications were needed to have a .95 confidence interval of +/- .001 or between .099 and .101. When alpha was .05, 182475 replications were needed to have a .95 confidence interval of +/- .001 or between .049 and .051 and when alpha was .01, 38032 replications were needed to have a .95 confidence interval of +/- .001 or between .009

and .011. Observed Type I error proportions falling into the respective .95 confidence intervals are considered to be accurate estimates of the expected Type I error rate.

Within each nominal alpha/sample size/number of groups configuration, the number of ANOVA replications were generated. Each replication involved drawing of elements of the sample from a distribution of normal deviates, computation of sample means, and the omnibus *F* test. Error rates were determined for protected and unprotected tests for each of the four multiple comparison procedures. While Dunn-Bonferroni, Dunn-Sidak, and HSD use only one critical value for all differences, the pairwise differences were recorded in a hierarchical fashion to determine pairwise differences significant at each of the numbers of steps between means from k down to 2. This approach permitted determination of experimentwise Type I error (at least one Type I error per experiment) or a Type I error for the largest mean difference, and per-experiment Type I errors or the total number of Type I errors observed regardless of where they are in the stepwise structure.

Summary statistics were computed for each alpha level for experimentwise and per-experiment conditions including: the mean proportion of Type I errors, standard deviation of the proportion of Type I errors, and the percentage of those proportions falling in the three regions associated with the .95 confidence interval of nominal alpha +/- 0.001.Additional analysis included computation of differences between per-experiment proportions and experimentwise proportions (PE-EW).

Preliminary analyses were run using the Monte Carlo program to test its accuracy. First, 500,000 standard normal scores (*z* scores) were generated and the statistics for the distribution were computed. This resulted in a mean = -.00096, variance = 1.0013, skewness = .00056, kurtosis = .00067, and the Wilk-Shapiro D = .000734 (nonsignificant). Thus, we concluded that the program generates reasonable normal distributions. Second, 900,000 cases were computed with k ranging from 2 to 10 and n ranging from 5 to 100 with no differences between the group means. In each case, the proportions of significant *F* statistics were computed corresponding to preset alphas of .25, .10, .05, .01, .001, and .0001. The resulting proportions of rejected null hypotheses were .24989, .10106, .05071, .01022, .001004, and .000103 respectively. These results support the accuracy of the Monte Carlo program.

Results

The first research question is: Which of these four multiple comparison procedures has the most accurate control of Type I error across the three alpha conditions? The results for each of the three alpha conditions are presented in Tables 1 through 3 and Figures 1 through 3. Table 1 and Figure 1 present results when nominal alpha is set at .10, Table 2 and Figure 2 present results when nominal alpha is set at .05, and Table 3 and Figure 3 present results when nominal alpha is set at .01.

When alpha is set at .10, if the Type I error rate philosophy is experimentwise, the most accurate of these four procedures is clearly Tukey's HSD, conducted as an unprotected test, with a mean observed Type I error rate of .09940 and with 78.6% of the observed Type I errors being in the range of .099 to .101. The HSD conducted as a protected test with an experimentwise control philosophy had a mean of .08134, somewhat conservative. All of the other procedures conducted, based on the experimentwise philosophy are conservative procedures with mean Type I error rates in the range of .07239 to .07535 when conducted as unprotected tests and .06695 to .06885 when conducted as protected tests.

If the Type I error control philosophy is per-experiment, the most accurate procedure is clearly the Dunn-Bonferroni, conducted as an unprotected test with a mean observed Type I error rate of .10011 and 85.7% of the observed Type I errors in the range of .099 to .101. When the philosophy is per-experiment and conducted as unprotected tests, the other three methods tend to be liberal with the mean error rate for the Dunn-Sidak at .10481 and the Holm procedure at .10582. Tukey's HSD was very liberal in this situation with a mean error rate of .14579. When conducted as protected tests, HSD was slightly liberal with a mean error of .12741 and the other three methods were reasonably accurate with mean errors of .09466 for the Dunn-Bonferroni,

.09834 for the Dunn-Sidak, and .10036 for Holm's procedure.

When nominal alpha was set at .05, the results were very similar. If the Type I error rate philosophy is experimentwise, the most accurate of these four procedures is clearly Tukey's HSD, conducted as an unprotected test, with a mean observed Type I error rate of .04993 and with 97.6% of the observed Type I errors being in the range of .049 to .051. The HSD conducted as a protected test with an experimentwise control philosophy had a mean of .03865, somewhat conservative. All of the other procedures conducted, based on the experimentwise philosophy are conservative procedures with mean Type I error rates in the range of .03864 to .03943 when conducted as unprotected tests and .03352 to .03395 when conducted as protected tests.

If the Type I error control philosophy is per-experiment, the most accurate procedure is clearly the Dunn-Bonferroni, conducted as an unprotected test with a mean observed Type I error rate of .04998 and 92.9% of the observed Type I errors in the range of .049 to .051. When the philosophy is per-experiment and conducted as unprotected tests, the other three methods tend to be liberal with the mean error rate for the Dunn-Sidak at .05110 and the Holm procedure at .05208. Tukey's HSD was very liberal in this situation with a mean error rate of .06674. When conducted as protected tests, HSD was slightly liberal with a mean error of .05531 and the other three methods were slightly conservative with mean errors of .04483 for the Dunn-Bonferroni, .04560 for the Dunn-Sidak, and .04696 for Holm's procedure.

When nominal alpha was set at .01, the patterns of results were very similar to the .10 and .05 nominal alpha conditions. If the Type I error rate philosophy is experimentwise, the most accurate of these four procedures is clearly Tukey's HSD, conducted as an unprotected test, with a mean observed Type I error rate of .01002 and with 100.0% of the observed Type I errors being in the range of .009 to .011. The HSD conducted as a protected test with an experimentwise control philosophy had a mean of .00702, somewhat conservative. All of the other procedures conducted, based on the experimentwise philosophy are conservative procedures with mean Type I error rates in the range of .00860 to .00865 when conducted as unprotected tests and .00647 to .00649 when conducted as protected tests. If the Type I error control philosophy is per-experiment, the most accurate procedure is clearly the Dunn-Bonferroni, conducted as an unprotected test with a mean observed Type I error rate of .01003 and 97.6% of the observed Type I errors in the range of .009 to .011.

When the philosophy is per-experiment and conducted as unprotected tests, the Dunn-Sidak outcome is very close to the Dunn-Bonferroni outcome with a mean error rate of .01007 and 92.9% of the observed errors in the .009 to .011 range. The other two methods tend to be liberal with the mean error rate for the Holm procedure at .01026 and Tukey's HSD with a mean error rate of .01181. When conducted as protected tests, all four methods were conservative with Tukey's HSD slightly less conservative with a mean error rate of .00878. The other three methods were slightly more conservative with mean errors of .00790 for the Dunn-Bonferroni, .00793 for the Dunn-Sidak, and .00814 for Holm's procedure.

In summary, relative to research question 1 (Which of these four multiple comparison procedures has the most accurate control of Type I error across the three alpha conditions?), if the most accurate control of per-experiment Type I error is desired, the Dunn-Bonferroni, conducted as an unprotected test, is the most accurate across all three levels of alpha. It consistently provides a mean Type I error rate closest to nominal alpha, has the lowest variance, and captures the highest proportion of observed Type I errors in the expected +/- .001 interval. Although the Dunn-Sidak and Holm provide values that are reasonably close, they tend to be slightly more liberal and less accurate, particularly with higher nominal alpha. As alpha decreases, both the Dunn-Sidak and Holm approach the level of accuracy of the Dunn-Bonferroni. Tukey's HSD is liberal as an unprotected test in control of per-experiment Type I error, although this decreases as alpha decreases. If the error control philosophy is experimentwise, Tukey's HSD is the most accurate, conducted as an unprotected test. It has a mean error closest to nominal alpha, the lowest

variance, and the highest proportion of observed Type I errors in the expected +/- .001 interval. When alpha is .10, HSD is slightly less accurate than when alpha is .05 or .01. The other three methods are conservative, with the Dunn-Sidak being slightly less conservative compared with Dunn-Bonferroni and Holm.

The second research question is: Does error control accuracy differ when tests are conducted as protected or unprotected tests? If the interest is in using any of these methods as a protected test, a practice not generally supported by these data, the HSD provides the most accurate control of experimentwise Type I error although it is very conservative at all alpha levels. The other three methods are very conservative in control of experimentwise Type I error. If per-experiment control of Type I error is the philosophy, HSD is liberal when alpha is .10 or .05 but becomes more accurate, even somewhat conservative, when alpha is .01. Of the remaining three, Holm's procedure tends to be more accurate across the three alpha levels. It is clear and expected that unprotected tests are more powerful than protected tests.

The third research question is: Do methods differ relative to accuracy when conducted as experimentwise vs. per-experiment control? It seems pretty clear that the results vary a great deal depending on the Type I error control philosophy. By the very nature of these philosophies, there will be a higher proportion of Type I errors in the per-experiment condition compared with the experimentwise condition. In every case, across alpha levels and for both protected and unprotected tests, the lowest difference between these rates was for the Dunn-Bonferroni, followed relatively closely by the Dunn-Sidak, Holm's procedure has next highest, and the highest difference was for the HSD. Thus, the issue is more a concern if one is using the HSD as compared with the other three methods.

Conclusion

These results provide insights on two major controversies. One is the need for a significant omnibus F test as the gateway for conducting pairwise follow-ups (i. e., the protected test). Is it not possible, as Hancock & Klockars (1996) pointed out, that this requirement overprotects against finding pairwise differences? These results certainly support that claim, particularly when experimentwise Type I error is the control philosophy. Protected tests were more conservative in every case. It can clearly be concluded that none of these four tests should be used as protected tests when experimentwise error control is used. If per-experiment error control is desired, only the Holm procedure with alpha of .10 was more accurate as a protected test than as an unprotected test. However, that accuracy difference was lower when alpha was .05 or .01.

The other controversy is the use of experimentwise vs. per-experiment Type I error control. Clearly there is a difference in the error rates of these philosophies. The authors of this article contend that per-experiment mode is closest to the realities of pairwise hypothesis testing, because more than just the largest pairwise difference is of interest and all pairwise comparisons are tested. The conventional wisdom, based on experimentwise Type I error control, is that the Dunn-Bonferroni is very conservative and that the HSD is conservative, but less so.

The HSD is often recommended because it is conservative, yet provides reasonable power for finding significant differences; but this relates to experimentwise control and a protected test. Yet, arguments could be made that the HSD gets its power from a higher-than-nominal alpha level. In this research, when HSD is used as a protected test with alpha of .10 or .05, the actual per-experiment Type I error rates are .12741 and .05531 respectively and actual experimentwise Type I error rates were much lower at .08134 and .03865. Thus, the operational alpha level is not the nominal level, but a higher level.

If one is truly interested in maintaining an accurate level of control of Type I error, then methods which are shown to provide accurate actual controls should be used, and the power available can be determined by other comparison conditions: sample size, effect size, number of groups, and error variance. This research indicates that Tukey's HSD, conducted as an unprotected test, is the most accurate control of experimentwise Type I error. If it is

desired that accurate, as advertised, control of per-experiment Type I error be the primary criterion, there is one method that seems to provide that regardless of alpha level and that is the Dunn-Bonferroni conducted as an unprotected test.

These findings are not consistent with common wisdom or with recommendations found or implied in most statistics texts. However, it is hoped that this research influences others to replicate this work, possibly using other methods. Only when one is willing to question our current practice can one be able to improve on it.

Additional study of the discrepancy between experimentwise and per-experiment Type I errors is needed. Determining the importance of this discrepancy is required. The current study did not consider the case of unequal sample sizes or heterogenous variances. Is it the same under conditions of unequal sample sizes and/or variances? While it might be useful to include other procedures such as the Student-Newman-Keuls, Scheffé, and modifications of Holm's procedure, it is believed that it is unlikely that any of these methods will fare better as methods of Type I error control than Tukey's HSD when experimentwise is the control philosophy, or the Dunn-Bonferroni when per-experiment is the control philosophy and unprotected tests are used.

Table 1. Observed Per-Experiment and Experimentwise Type I Error Rates for Selected Multiple Comparison Procedures when Conducted as Protected and Unprotected Tests with Alpha= .10

		Protected Test			Unprotected Test		
		Per-Experiment Error (PE)	Experiment-wise Error (EW)	PE - EW Difference	Per-Experiment Error (PE)	Experiment-wise Error (EW)	PE - EW Difference
Dunn-Bonferroni	M	.09466	.06695	.02771	.10011	.07239	.02772
	M – α	–.00534	–.03305		+.00011	–.02767	
	SD	.00427	.00962		.00075	.00626	
	% in α +/-.001	19.0	0		85.7	0	
Dunn-Sidak	M	.09834	.06885	.02949	.10481	.07535	.02946
	M – α	–.00166	–.03115		+.00481	–.02465	
	SD	.00401	.00972		.00093	.00625	
	% in α +/-.001	19.0	0		0	0	
Holm	M	.10036	.06695	.03341	.10582	.07239	.03343
	M – α	+.00036	–.03305		+.00582	–.02761	
	SD	.00739	.00962		.00346	.00626	
	% in α +/-.001	2.4	0		7.1	0	
HSD	M	.12741	.08134	.04607	.14579	.09940	.04639
	M – α	+.02741	–.01866		+.04579	–.00060	
	SD	.00906	.00755		.01472	.00102	
	% in α +/-.001	0	0		0	78.6	

Figure 1

Accuracy of Type I Error Control with Experimentwise and Per-Experiment Control Conducted as Protected and Unprotected Tests when Nominal Alpha= .10 and % in .10 +/- 0.001

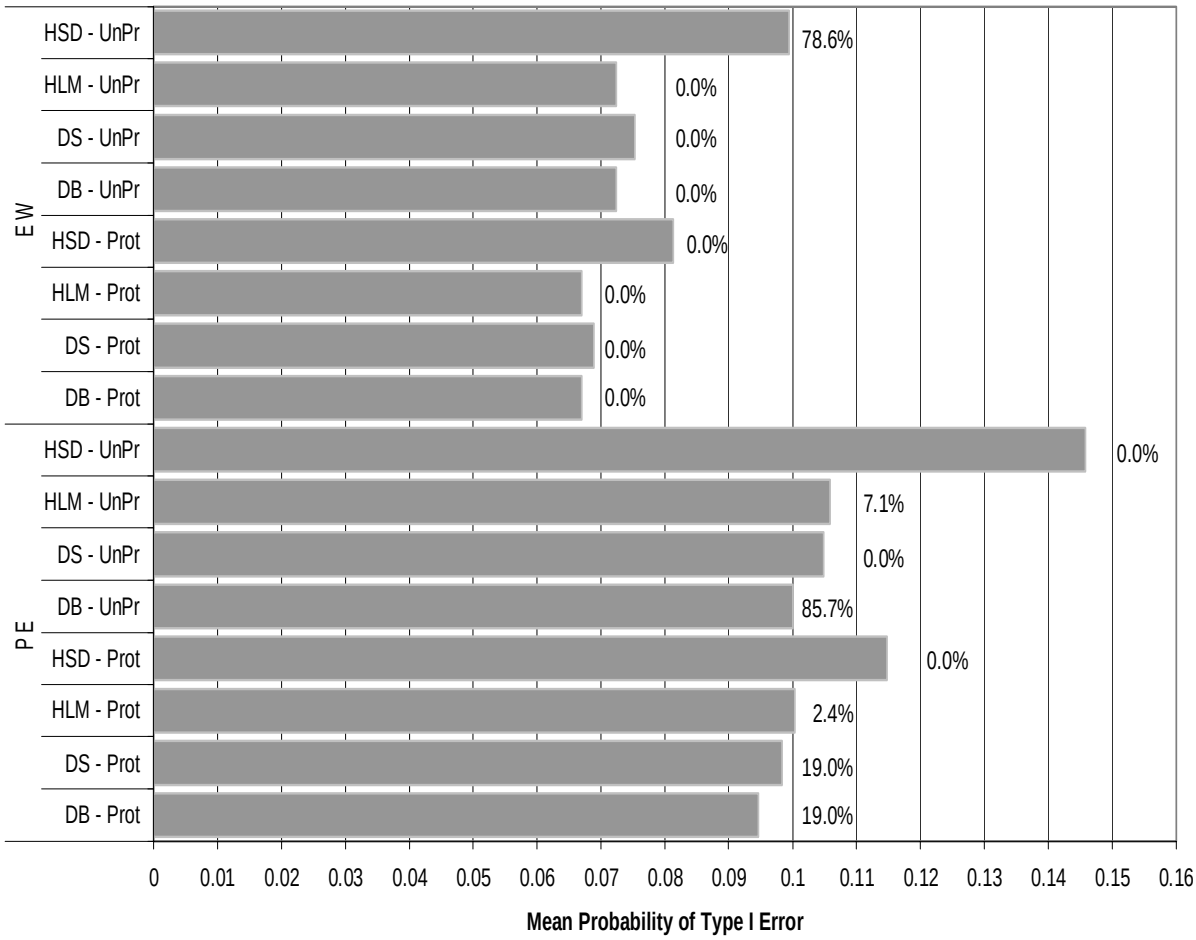


Table 2. Observed Per-Experiment and Experimentwise Type I Error Rates for Selected Multiple Comparison Procedures when Conducted as Protected and Unprotected Tests with Alpha= .05

		Protected Test			Unprotected Test		
		Per-Experiment Error (PE)	Experiment-wise Error (EW)	PE - EW Difference	Per-Experiment Error (PE)	Experiment-wise Error (EW)	PE - EW Difference
Dunn-Bonferroni	M	.04483	.03352	.01113	.04998	.03864	.01134
	M – α	–.00517	–.01648		–.00002	–.01136	
	SD	.00315	.00534		.00054	.00294	
	% in α +/-.001	7.1	0		92.9	0	
Dunn-Sidak	M	.04560	.03395	.01165	.05110	.03943	.01167
	M – α	–.00440	–.00405		+.00110	–.01057	
	SD	.00308	.00536		.00052	.00291	
	% in α +/-.001	16.7	0		50.0	0	
Holm	M	.04696	.03352	.01344	.05208	.03864	.01344
	M – α	–.00304	–.01648		+.00208	–.01136	
	SD	.00433	.00535		.00146	.00294	
	% in α +/-.001	19.0	0		33.3	0	
HSD	M	.05531	.03865	.01666	.06674	.04993	.01681
	M – α	+.00531	–.01135		+.01674	–.00007	
	SD	.00324	.00458		.00541	.00048	
	% in α +/-.001	2.4	0		0	97.6	

Figure 2
Accuracy of Type I Error Control with Experimentwise and Per-Experiment Control Conducted as Protected and Unprotected Tests when Nominal Alpha= .05 and % in .05 +/- 0.001

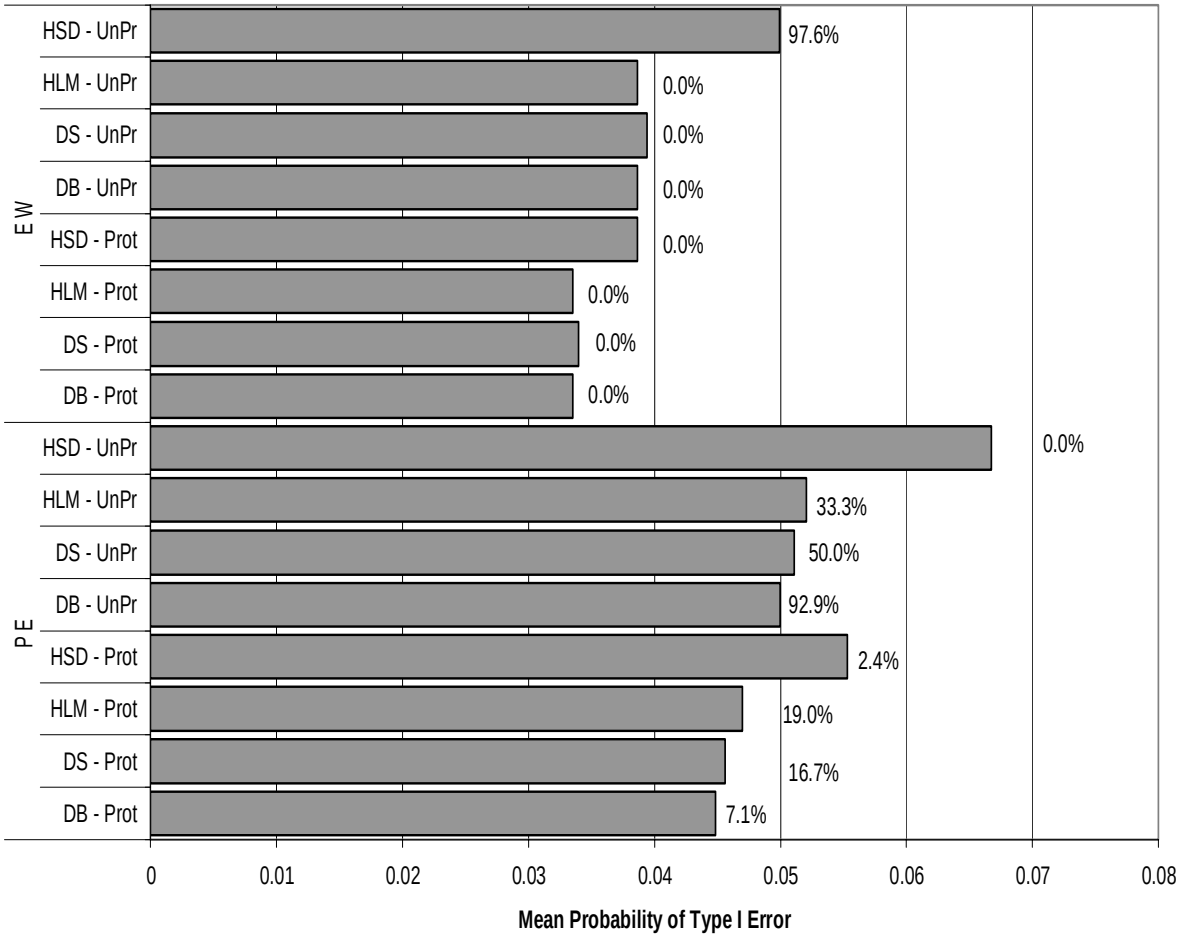
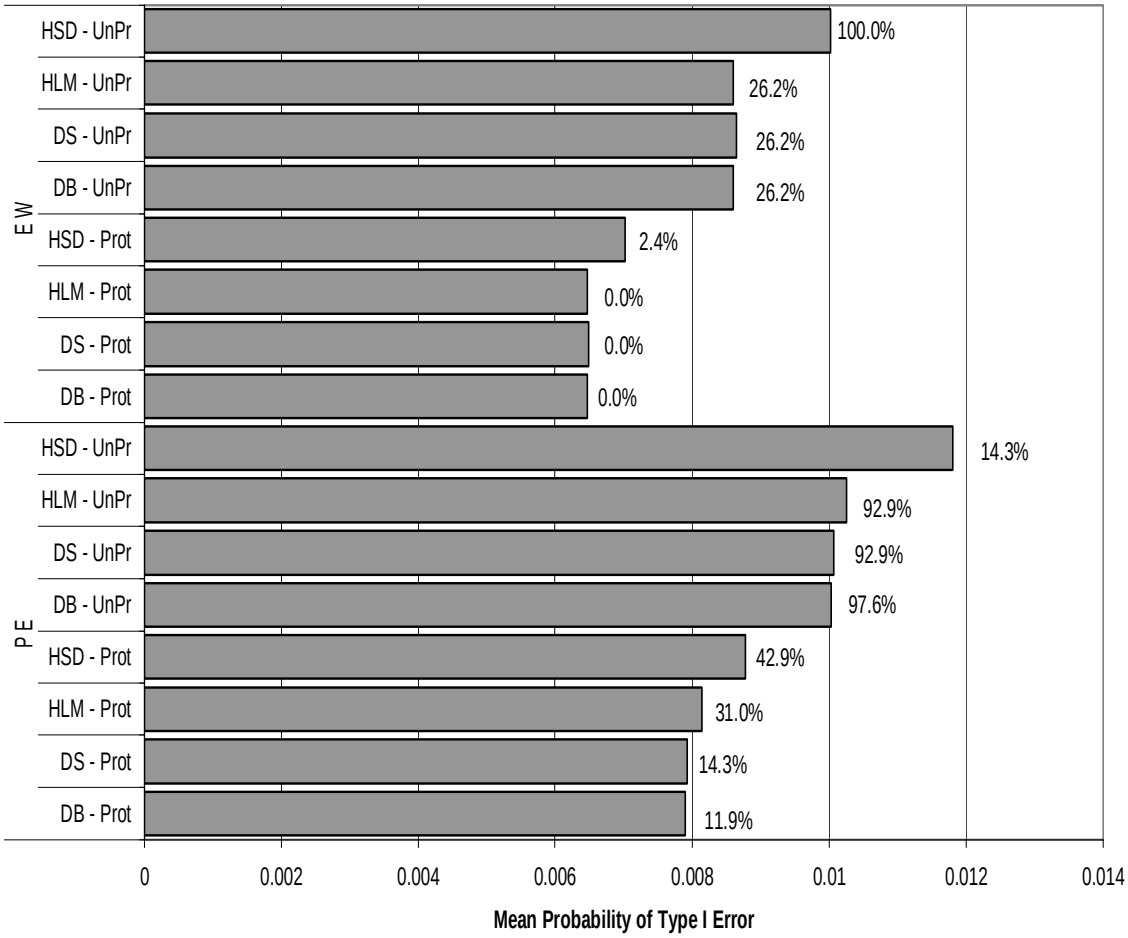


Table 3. Observed Per-Experiment and Experimentwise Type I Error Rates for Selected Multiple Comparison Procedures when Conducted as Protected and Unprotected Tests with Alpha= .01

		Protected Test			Unprotected Test		
		Per-Experiment Error (PE)	Experiment-wise Error (EW)	PE - EW Difference	Per-Experiment Error (PE)	Experiment-wise Error (EW)	PE - EW Difference
Dunn-Bonferroni	M	.00790	.00647	.00143	.01003	.00860	.00143
	M – α	–.00210	–.00353		+.00003	–.00140	
	SD	.00103	.00123		.00048	.00059	
	% in α +/-.001	11.9	0		97.6	26.2	
Dunn-Sidak	M	.00793	.00649	.00144	.01007	.00865	.00142
	M – α	–.00207	–.00351		+.00007	–.00135	
	SD	.00103	.00122		.00049	.00058	
	% in α +/-.001	14.3	0		92.9	26.2	
Holm	M	.00814	.00647	.00167	.01026	.00860	.00166
	M – α	–.00186	–.00353		+.00026	–.00140	
	SD	.00119	.00123		.00054	.00059	
	% in α +/-.001	31.0	0		92.9	26.2	
HSD	M	00878	.00702	.00176	.01181	.01002	.00179
	M – α	–.00122	–.00298		+.00181	+.00002	
	SD	.00097	.00116		.00080	.00043	
	% in α +/-.001	42.9	2.4		14.3	100.0	

Figure 3
Accuracy of Type I Error Control with Experimentwise and Per-Experiment Control Conducted as Protected and Unprotected Tests when Nominal Alpha= .01 and % in .01 +/- 0.001



References

Fisher, R. A. (1935, 1960). *The design of experiments*, 7th ed. London: Oliver & Boyd; New York: Hafner.

Gossett, W. S. (1907-1938) (1943). *Student's collected papers*. (E. S. Pearson & Wishart, J., editors). London: University Press, Biometrika Office.

Hancock, G. R. & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971), *Review of Educational Research, 66*, 269-306.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, and Winston, Inc.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Kennedy, J. J. & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America, Inc.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. (2nd ed.). Belmont, CA: Brooks Cole.

Klockars, A. J. & Hancock, G. R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement, 54* (2), 292-298.

Keuls, M. (1952). The use of "Studentized range" in connection with an analysis of variance. *Euphytica, 1*, 112-122.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth Publishing Company.

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

Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika, 31*, 20-30.

Pearson, E. S. & Sekar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika, 28*, 308-320.

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

Toothaker, L. E. (1993). *Multiple comparison procedures*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-089. Newbury Park, CA: Sage.

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

Working, H. & Hotelling, H. (1929). Application of the theory of error to the interpretation of trends. *Journal of the American Statistical Association, 35*, 73-85.