Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v12n1p55

Some clarifications regarding power and Type I error control for pairwise comparisons of three groups

By Frane

Published: 26 April 2019

Electronic Journal of Applied Statistical Analysis
Vol. 12, Issue 01, April 2019, 55-68
DOI: 10.1285/i20705948v12n1p55

Some clarifications regarding power and Type I error control for pairwise comparisons of three groups

Andrew V. Frane*

University of California, Los Angeles
UCLA Department of Psychology, Los Angeles, CA 90095, United States of America

Published: 26 April 2019

A previous study in this journal used Monte Carlo simulations to compare the power and familywise Type I error rates of ten multiple-testing procedures in the context of pairwise mean-comparisons in balanced three-group designs. The authors concluded that, of those ten procedures, the Benjamini–Hochberg procedure was the "best." However, they did not compare the Benjamini–Hochberg procedure to classical, commonly used multiple-testing procedures that were developed specifically for pairwise comparisons, such as Fisher's protected least significant difference procedure and Tukey's honest significant difference procedure. Simulations in the present study show that in the three-group case, Fisher's method is more powerful than both Tukey's method and the Benjamini–Hochberg procedure, in terms of both *per-pair power* (mean probability of significance across the tests of false null hypotheses) and *any-pair power* (probability of significance in at least one test of a false null hypothesis). Compared to the Benjamini–Hochberg procedure, Tukey's method is shown to have lower per-pair power, but slightly greater any-pair power. The maximum familywise Type I error rate of all three procedures (Benjamini-Hochberg, Fisher, and Tukey) was equal to the designated alpha level.

keywords: Type I error, multiple comparisons, multiple testing, multiplicity, power.

*Corresponding author: avfrane@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

One of the most common types of statistical analyses for experimental designs is pairwise comparison of group means. Often there are more than two groups, and hence more than one comparison. In that case, a multiple-testing procedure is typically required to prevent Type I error inflation. Presumably, in practice the most commonly encountered number of groups, besides two, is three. So it is worth asking which multiple-testing methods are preferable in the three-group case. Sections 1.1–1.3 below provide an inexhaustive list of well known multiple-testing methods that can be used in the three-group case to control the *familywise Type I error rate* (FWER; the probability of at least one Type I error). These methods can typically be executed with simple commands in standard statistical software.

1.1 General-purpose FWER-control procedures

Most FWER-control procedures can be applied not only to pairwise mean-comparisons in particular, but also to multiple-testing situations more generally. The best-known such method is the Bonferroni procedure. It controls not only the FWER, but also the *per-family Type I error rate* (PFER; the expected number of Type I errors), which is a stricter standard than FWER. Consequently, the Bonferroni procedure is often regarded as overly conservative, given that FWER has become a preferred standard over PFER in practice (but see Frane, 2015a; Frane, 2015b; Klockars and Hancock, 1994). By sacrificing PFER control, other general-purpose FWER-control procedures (e.g., Holm, 1979; Hochberg, 1988; Hommel, 1988) can provide greater *power* (probability or frequency of significance in tests of false null hypotheses) than the Bonferroni procedure.

The reason FWER control is more "powerful" than PFER control is that, unlike PFER, FWER does not count all Type I errors that occur. Instead, in a given *family* (e.g., in a given study), multiple co-occurring errors count the same as a single error. For instance, if 5 out of every 100 families each contained a single error, then the PFER and FWER would both be .05, but if 5 out of every 100 families each contained 10 errors, then the FWER would still be .05—even though the PFER would be inflated to .50. Thus, the more an FWER-control procedure can relax protection against co-occurring errors, the more frequently significances can occur, and hence the more powerful the procedure can be while still controlling the FWER.

1.2 FWER-control procedures specifically for pairwise comparisons

Given that the power of FWER-control procedures is largely based on avoiding "overprotection" against co-occurring errors, knowing how co-occurring errors behave for the specific type of tests at hand allows for the design of more powerful procedures. For instance, for pairwise comparisons of group means, there is an inherent logical relationship between the null hypotheses (e.g., for three group means $\{\mu_1, \mu_2, \mu_3\}$, if $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$, then $\mu_1 = \mu_3$). Additionally, there is correlation among the test statistics in pairwise comparisons, which increases the probability of co-occurrence among errors (e.g., the test of μ_1 versus μ_2 is correlated with the test of μ_1 versus any other mean). By using such information, FWER-control procedures that are designed specifically for pairwise comparisons can provide more power than general-purpose FWER-control procedures.

Tukey's (1953) *honest significant difference procedure* (HSD) is one of the oldest and most widely used methods of FWER control that has been designed for pairwise comparisons (Hancock and Klockars, 1996; Ramsey and Ramsey, 2008). It requires equal group sizes, but there are well known modifications of HSD that accommodate unequal group sizes (e.g., the Tukey–Kramer procedure; Kramer, 1956; Tukey, 1953). HSD also assumes that the population distributions have equal variance.

An even older procedure for pairwise comparisons is Fisher's (1935) *protected least significant difference procedure* (PLSD), which has been found to be more powerful than HSD (Ramsey, 1978; Seaman et al., 1991). PLSD works as follows: If the omnibus test (an ANOVA in classical PLSD, though other tests can be used) is statistically significant at the nominal alpha level, then the pairwise comparisons (Student t-tests in classical PLSD, though other tests can be used) are conducted without adjustment; otherwise, significance of all pairwise comparisons is forfeited. In general, PLSD only controls the FWER "in the weak sense," meaning that it controls the FWER when all null hypotheses are true, but can fail to control the FWER when only some null hypotheses are true (Hochberg and Tamhane, 1987, p. 3). However, procedures that generally only control the FWER in the weak sense also control the FWER "in the strong sense" (meaning regardless of what proportion of the null hypotheses are true) in a special case: pairwise comparison of three groups on a single outcome variable (Hayter, 1986; Shaffer, 1986). To understand why, consider the three possible mean configurations in the three-group case: (1) all means are equal ($\mu_1 = \mu_2 = \mu_3$), in which case weak control of the FWER is as effective as strong control because all null hypotheses are true; (2) only two means are equal ($\mu_1 = \mu_2 \neq \mu_3$), in which case there is no multiple-testing problem, because there is only one true null hypothesis and thus the opportunity for Type I error is confined to a single comparison; (3) all means are different, in which case Type I error is impossible because there are no true null hypotheses. In other words, in the three-group case, multiple-testing can inflate the FWER only when all null hypotheses are true, so the distinction between weak and strong control of the FWER is moot. Thus, although PLSD is not typically advisable when there are more than three groups, it is valid for controlling the FWER in the three-group case, provided that the assumptions of the omnibus test are met.

Other procedures for pairwise comparisons have been developed, but are beyond the scope of this paper. For example, the Newman–Keuls procedure (Keuls, 1952; Newman, 1939), like PLSD, controls the FWER in the weak sense generally and in the strong sense for three groups. But it has been found to be slightly less powerful than PLSD in the three-group case (Seaman et al., 1991). Another method is Dunnett's (1955) procedure, which is applicable not when conducting "all possible" pairwise comparisons, but rather when comparing multiple treatment groups to a single control group (and not comparing the treatments to each other).

1.3 The Benjamini–Hochberg procedure

The Benjamini–Hochberg procedure (BH; Benjamini and Hochberg, 1995) is a general-purpose multiple-testing procedure that was designed to control the *false discovery rate* (the expected "proportion," loosely speaking, of significances that are Type I errors). Because false discovery rate is a more lenient standard than FWER, BH does not reliably control the FWER in general. However, like PLSD, BH controls the FWER when all null hypotheses are true, and therefore controls the FWER in the three-group case. BH has been shown to be valid for independent tests and for many typically-encountered types of positively dependent tests (Benjamini and Yekutieli, 2001; note that negative dependence is not plausible in typical two-sided testing scenarios).

1.4 The Félix and Menezes study

Previously in this journal, Félix and Menezes (FM; 2018) used Monte Carlo simulations to rank the performance of ten multiple-testing procedures in the context of pairwise mean-comparisons in balanced three-group designs. The pairwise comparisons were Student t-tests using the pooled standard deviation from all three groups. And the three population distributions were either all normal, all logistic, or all Gumbel. Because BH tended to rank highly in terms of both FWER-control and power, FM concluded that "the BH correction was the best overall, that is, it was good in both criteria" (p. 88). However, that conclusion requires some important caveats.

First of all, BH fails to reliably control the FWER (except in the weak sense) when there are more than three groups. Recall that the FWER control BH provides in the three-group case is based on the fact that there can only be multiple true null hypotheses when all means are equal. When there are more than three groups, there can be multiple true null hypotheses even when not all means are equal (e.g., when $\mu_1 = \mu_2 = \mu_3 \neq \mu_4$).

Another caveat is that FM did not compare BH to classical, commonly used FWER-control procedures that were devised explicitly for pairwise comparisons—most notably, HSD and PLSD. Instead, they compared BH to general-purpose multiple-testing procedures (e.g., Bonferroni), most of which were well known to be less powerful than BH (notwithstanding the Li procedure, which can be liberal for dependent tests; Li, 2008). Thus, for pairwise comparisons of three groups, although BH may indeed often be the best choice among the procedures that FM examined, that does not imply that BH is better than the standard procedures that are available for pairwise comparisons.

Interestingly, despite endorsing BH for pairwise comparisons of group means, FM claimed that BH is only valid for independent tests (p. 79). That claim, if it were true, would invalidate BH for pairwise comparisons of group means. However, although there are indeed some types of dependent tests for which BH is invalid, FM's own results suggest that two-sided pairwise comparison of three group means is not such a type (see also Benjamini and Yekutieli, 2001).

One conclusion in the FM study appears to have resulted from a methodological inconsistency in the simulations. Specifically, FM reported that for the logistic distribution, "the empirical power is a lot smaller, since this distribution has heavy tails" (p. 84).

However, as shown in their Table 1, the standard deviations they used for the logistic distribution were larger than the standard deviations they used for the normal distribution by a factor of $\pi/\sqrt{3} \approx 1.8$, and were larger than the standard deviations they used for the Gumbel distribution by a factor of $\sqrt{2} \approx 1.4$. Thus, the dramatically reduced power that FM observed for the logistic distribution essentially reflects the standard deviations that were used—not some inherent characteristic of the logistic distribution's shape, such as its slightly "heavy tails." Note that although FM did match the nominal values of the *scale parameters* across the different distribution types, that does not cause the actual spread of the distributions to be matched in any meaningful way, because the scale parameters were defined differently (i.e., as different linear functions of the standard deviation) for the different distribution types.

There are also some issues with how FM ranked the FWER control of the procedures. First of all, FWER was only examined when all population means were equal, i.e., when all null hypotheses were true. Although some procedures (such as Bonferroni) produce their maximum FWER when all null hypotheses are true, other procedures (such as BH) produce their maximum FWER not when all null hypotheses are true, but rather when only some null hypotheses are true and power is maximal (Finner and Roters, 2001). Note also that FM ranked each procedure's FWER control not by how low the FWER was, but rather by how close the FWER was to .05. That approach produces counterintuitive rankings. For instance, using that system, a controlled FWER of .040 would be considered "worse" than an inflated FWER of .059. That explains why, in certain conditions, FM ranked the Li procedure as number 1 in FWER control—even though it was the one procedure in the study that did not actually control the FWER (see their Figure 1). Moreover, for procedures that did control the FWER, higher FWERs were ranked as "better" than lower FWERs.

Perhaps the reason that FM reported rankings for the estimated FWERs rather than reporting the estimates directly is that the number of simulations—only 10,000 for each combination of parameters—did not provide sufficient precision. Indeed, when using simulations to estimate an incidence rate (such as FWER or power), the standard error for that estimation is inversely proportional to the square root of the number of simulations, as reflected in the following well-known formula: $\widehat{SE} = \sqrt{r(1-r)/n}$, where $\widehat{SE}$ is the estimated standard error, r is the observed incidence rate, and n is the number of simulations (Albert and Rizzo, 2012, p. 309). The corresponding 95% confidence interval may be computed as $r \pm 1.960\widehat{SE}$. For instance, when 10,000 simulations collectively produce an estimated FWER of .050, the $\widehat{SE}$ for that estimation is roughly .002, and the width of the corresponding 95% confidence interval is .009—which is presumably too wide for adequately estimating values beyond two decimal places.

The same principle of precision applies to the power estimates, for which $\widehat{SE}$ can be as high as .005 (maximized thusly when $r = .5$) and for which the corresponding 95% confidence interval can be as wide as .020. In fact, it is evident that the power estimates in the FM study were noisy, because in some cases, different parameter combinations that should have been computationally equivalent nonetheless produced different power rankings for the same procedure. Examples of this inconsistency can be seen in their

Figure 6, which shows that for group size 3 and standard deviation 1, the ranking of the Finner procedure typically changed when the positive/negative sign of the *location parameter* was flipped, i.e., when the directions of the nonzero mean-differences were reversed and the absolute magnitudes of the mean differences were unchanged. Given that the t-tests were two-sided, reversing the directions of the mean differences across the board should not have affected power at all. Note also that power was estimated only when two means were equal, so the all-means-different case was not considered.

Despite the FM study's methodological limitations, its primary empirical finding is sound: BH is more powerful than some other procedures that control the FWER for pairwise comparisons of the three groups. Thus, the FM study's main limitation is simply the lack of consideration given to standard procedures that were designed specifically for pairwise comparisons.

1.5 The present study

The present study followed up on the FM results by conducting Monte Carlo simulations to evaluate the performance of HSD, PLSD, and BH, for pairwise mean-comparisons in balanced three-group designs. This investigation essentially placed BH—the "winning" procedure from the FM study—in competition against classical procedures that were designed for pairwise comparisons. Various group-sizes, population distribution types, and standardized population-mean configurations were used. To avoid the aforementioned methodological limitations of the FM study, the following four steps were taken: (1) standard deviation was fixed, so that power comparisons among distribution types would be meaningful; (2) FWER was estimated not only when all means were equal, but also when only two means were equal (including scenarios in which power was maximal), in order to ensure that each procedure's maximum FWER was produced; (3) power was estimated not only when two means were equal, but also when all means were different; (4) a much larger number of simulations per parameter combination was performed than in the FM study, in order to increase the precision of the estimations.

2 Methods

2.1 Simulation parameters

Simulations were conducted to evaluate the FWER and power of BH, HSD, and PLSD, in the context of pairwise comparisons of three groups. In each simulation, a group of independent observations was randomly sampled from each of three populations. Group size, which was common to all three groups, was set to 5, 10, 15, 20, or 1000 observations. Population distribution type, which was also common to all three groups, was set to normal, logistic, or Gumbel (following FM). The array of population means was set to {0, 0, 0}, {0, 0, 1}, or {0, 1, 2} ("all-means-equal," "two-means-equal," or "all-means-different," respectively). Population standard deviation was fixed at 1 in all cases, so the population means may be considered as standardized.

1,000,000 simulations were performed for each parameter combination, i.e., for each of the 45 unique combinations of group size, population distribution type, and population-mean array. This is 100 times the number of simulations per parameter combination that was used by FM, and thus provides essentially 10 times the precision (recall from Section 1.4 that the standard error and the width of the confidence interval are inversely proportional to the square root of the number of simulations). For instance, given an estimated FWER of .050, the unadjusted 95% confidence interval for the "true" FWER would span from .0456 to .0544 in the FM study, but would span from .0496 to .0504 in the present study—providing tight lower and upper bounds that both round to .050.

Note that parameter manipulations that would be largely redundant with respect to power were not performed in this study. For example, changing the population variance or changing the size of the nonzero population-means would have a similar effect on power as changing the group size. And as noted in Section 1.4, reversing the positive/negative direction of the population mean differences across the board would have no effect whatsoever.

2.2 Pairwise comparisons

For BH and PLSD, the pairwise tests were two-sided Student t-tests using the pooled standard deviation from the three groups (following FM). For HSD, the pairwise tests (which are essentially built into the procedure) are analogous to those t-tests in that they are two-sided, use the pooled standard deviation, and involve the same statistical assumptions (e.g., normality and equal variance).

The familywise alpha level was set to .05 for all procedures. Thus, significance was determined by computing adjusted p-values (using the given procedure) and assessing whether they were lower than .05. Note that although PLSD is not typically described as a p-value adjustment, it may nonetheless be implemented as such by adjusting the p-value for each t-test to $\max\{p_t, p_{om}\}$, where p_t is the raw p-value for the given t-test and p_{om} is the p-value for the omnibus test.

2.3 FWER estimations

FWER was estimated for each procedure in each parameter combination. The estimated FWER was simply the proportion of simulations that produced at least one Type I error (i.e., the proportion of simulations in which at least one pairwise comparison of groups with equal population means was significant).

2.4 Power estimations

Two types of power were estimated for each procedure in each parameter combination (except when all means were equal, in which case power would be meaningless). *Per-pair power* (the mean probability of significance among comparisons for which the corresponding population mean difference is nonzero; Ramsey, 1978) was estimated by taking the comparisons for which the corresponding population means were unequal, and computing the proportion of those comparisons that produced significance. *Any-pair power* (the probability of obtaining significance in at least one pairwise comparison for which the corresponding population mean difference is nonzero; Ramsey, 1978) was estimated as the proportion of simulations that produced significance in at least one pairwise comparison of groups with unequal population means.

2.5 Software

All simulations and estimations were performed using a custom R program that was created using R version 3.3.3 (R Core Team, 2017). That program, which is provided as a supplement to the present article, contains a section entitled "Adjustable Parameters" that allows the user to specify the following 10 inputs: (1) number of simulations, (2) group size, (3) population mean for Group 1, (4) population mean for Group 2, (5) population mean for Group 3, (6) population standard deviation for group 1, (7) population standard deviation for group 2, (8) population standard deviation for group 3, (9) familywise alpha level, and (10) population distribution type (either "normal," "logistic," or "Gumbel"). Specifying "Gumbel" as the population distribution type requires the "evd" package (Stephenson, 2002), which includes the *rgumbel* function used to generate randomly sampled observations from a Gumbel distribution. All other operations in the program use endogenous R commands, such as *rnorm* (to generate randomly sampled observations from a normal distribution), *rlogis* (to generate randomly sampled observations from a logistic distribution), *aov* (to fit the ANOVA model used as the first step in both HSD and PLSD), *pairwise.t.test* (to perform the pairwise t-tests used for both PLSD and BH), *TukeyHSD* (to compute HSD-adjusted p-values from the ANOVA model), and *p.adjust* (to compute BH-adjusted p-values from the raw p-values obtained in the t-tests).

The program is straightforward to use, and readers are invited to use it to explore whatever parameter combinations they are interested in. Because performing a large number of simulations may take considerable processing time (simplicity was favored over speed in the coding), readers who wish to do a large number of simulations are advised to first do a test run using a small number of simulations (e.g., 1000), in order to estimate the processing time per simulation.

3 Results

3.1 FWER

Tables 1, 2, and 3 show the FWERs when population distributions were normal, logistic, and Gumbel, respectively. In each table, the sub-table on the left is for the all-means-equal case, and the sub-table on the right is for the two-means-equal case. $\widehat{SE} \leq .0002$ for each estimation.

For each procedure, the maximum FWER was .050 for each distribution type. Thus, all procedures controlled the FWER in all examined parameter combinations. FWERs for logistic and Gumbel distributions tended to be lower than corresponding FWERs for

Table 1: FWERs for pairwise comparisons of three groups from normal distributions

	Population means: {0, 0, 0}					**Population means: {0, 0, 1}**				
	Group size					Group size				
Method	**5**	**10**	**15**	**20**	**1000**	**5**	**10**	**15**	**20**	**1000**
PLSD	.050	.050	.050	.050	.050	.041	.047	.049	.050	.050
HSD	.050	.050	.050	.050	.050	.020	.020	.020	.019	.019
BH	.044	.046	.046	.046	.047	.031	.036	.040	.043	.050

Table 2: FWERs for pairwise comparisons of three groups from logistic distributions

	Population means: {0, 0, 0}					**Population means: {0, 0, 1}**				
	Group size					Group size				
Method	**5**	**10**	**15**	**20**	**1000**	**5**	**10**	**15**	**20**	**1000**
PLSD	.047	.049	.049	.049	.050	.041	.047	.049	.049	.050
HSD	.047	.049	.049	.049	.050	.019	.019	.019	.019	.019
BH	.041	.044	.045	.045	.046	.030	.036	.040	.043	.050

normal distributions, but such differences were slight and became increasingly negligible as group size increased (in accordance with the *central limit theorem*; see Hogg and Craig, 1965, p. 196).

Note that BH did not exhibit its maximum FWER when all means were equal, but rather when two means were equal and power was maximal (i.e., when group size was very large, though the same high power could have been achieved for smaller group sizes by making μ_3 very large). That is because in the BH algorithm, obtaining a low p-value in a given comparison can allow other comparisons to be tested more leniently, meaning that high power in comparisons of groups that truly differ in the population can increase

Table 3: FWERs for pairwise comparisons of three groups from Gumbel distributions

	Population means: {0, 0, 0}					**Population means: {0, 0, 1}**				
	Group size					Group size				
Method	**5**	**10**	**15**	**20**	**1000**	**5**	**10**	**15**	**20**	**1000**
PLSD	.046	.047	.048	.048	.050	.037	.045	.048	.049	.050
HSD	.045	.047	.048	.048	.050	.019	.019	.019	.019	.019
BH	.040	.042	.044	.044	.046	.028	.034	.038	.042	.050

the probability of Type I error in other comparisons. As noted by Finner and Roters (2001), PLSD is somewhat similar to BH in that regard: When a single population mean is different and power is arbitrarily high, the omnibus test in PLSD is essentially guaranteed to be significant, thereby allowing all pairwise comparisons to be conducted (without adjustment) essentially 100% of the time and thus maximizing the opportunity for Type I error.

3.2 Power

Figures 1A and 1B show power versus group-size for the two-means-equal case and all-means-different case, respectively, when population distributions were normal. Results are not shown for group size 1000, because both types of power were always 1 in that case. Per-pair power exhibited a clear hierarchy: PLSD was more powerful than BH (though to an increasingly negligible extent as power increased when all means were different), and BH was more powerful than HSD. Any-pair power also exhibited a clear pattern, though differences between procedures were small: PLSD was marginally more powerful than HSD when two means were equal, and was essentially indistinguishable from HSD when all means were different, whereas BH was the least powerful procedure for both mean configurations by a small margin (and of course the power of all three procedures converged as power approached 1). Altogether, PLSD emerged as the clear winner with regard to power, in that it essentially performed as well or better than the other procedures in every examined parameter combination and by both definitions of power.

Using Gumbel or logistic distributions instead of normal distributions did not alter the power hierarchies described in the preceding paragraph. In fact, power estimates for the Gumbel and logistic distributions were typically only marginally different from the corresponding power estimates under normality. Pooling across all procedures and parameter combinations (excluding group size 1000, for which power was always 1), the mean difference in power between the logistic distribution and the normal distribution was .007 for per-pair power and .004 for any-pair power (favoring the logistic distribution in both cases). And the mean difference in power between the Gumbel distribution and the normal distribution was .012 for per-pair power and .008 for any-pair power (favoring the Gumbel distribution in both cases). These results confirm that, as noted in Section 1.4, the dramatically reduced power FM observed for the logistic distribution reflects a methodological inconsistency in their simulations, rather than the non-normality itself. That said, because the logistic and Gumbel distributions are not radically different from the normal distribution in shape, the present results should not be taken to imply that pairwise parametric testing is highly robust to non-normality in general. More substantial departures from normality can have more substantial impact on FWERs and power for parametric tests, especially when sample sizes are small (Cribbie and Keselman, 2003).

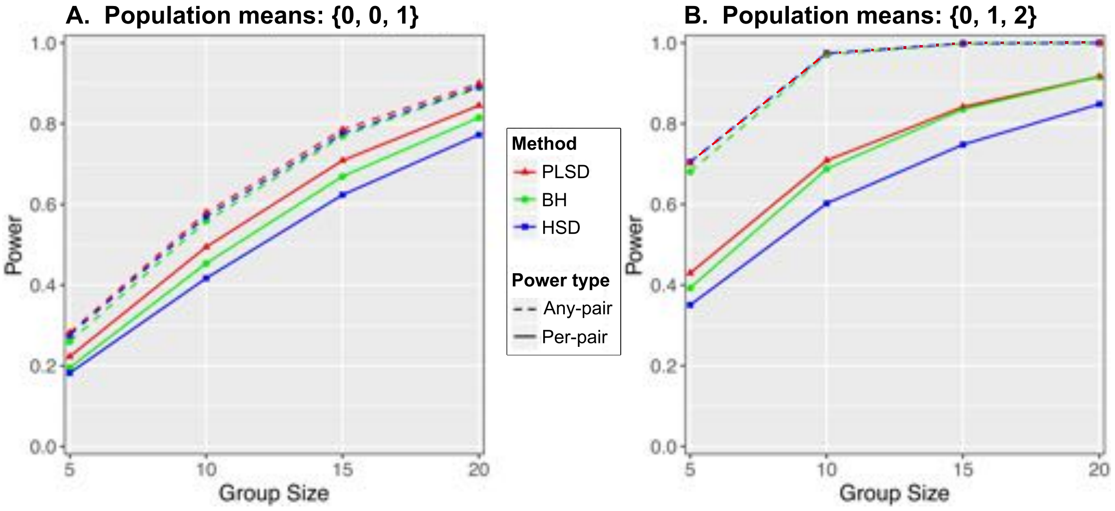

Figure 1: Power of Fisher's protected least significant difference tests (PLSD), Tukey's honest significant difference tests (HSD), and Benjamini–Hochberg adjusted Student t-tests (BH), for pairwise comparisons of three groups from normally distributed populations with standard deviation 1. $\widehat{SE} \leq .0005$ for each point estimation.

4 Discussion

For pairwise comparisons of three group means, FM recommended BH-adjusted Student t-tests. However, the present study's results suggest that PLSD is a preferable method in that context. Indeed, PLSD was consistently either as powerful as, or more powerful than, both HSD and BH—despite the fact that all three procedures had the same maximum FWER (which was equal to the designated familywise alpha level). This held true for both per-pair power and any-pair power, and held true regardless of whether population distributions were normal, logistic, or Gumbel.

On the other hand, HSD offers a notable feature that PLSD and BH do not: simultaneous confidence intervals. Given that researchers should typically be interested not only in whether means are different, but also in how different the means are, simultaneous confidence intervals are often valuable (Phillips et al., 2013). If using PLSD or BH, one could still report confidence intervals (unadjusted, or perhaps HSD-adjusted), but then the confidence intervals might be incongruent with the significance test results. Thus, heuristically speaking, it appears reasonable in the three-group case to recommend HSD when simultaneous confidence intervals are required, and to recommend PLSD otherwise (a heuristic that is consistent with Hancock and Klockars, 1996).

A caveat is that both PLSD (in its classical form) and HSD assume that the population distributions have equal variance. The present study did not examine procedures designed for unequal variances (Keselman et al., 1999; Ramsey et al., 2011; Ramsey and Ramsey, 2009). However, simulations by Keselman et al. (1999) suggest that in the three-group case, PLSD remains more powerful than BH when adapted for unequal variances (i.e., when using test statistics based on the unpooled variances). Note also that the logic of PLSD remains valid if one substitutes nonparametric tests that do not assume normality (e.g., a Kruskal-Wallis omnibus test followed by Wilcoxon rank-sum tests, rather than an ANOVA followed by t-tests).

It is interesting that authors often seem reluctant to recommend PLSD for the three-group case, even if they acknowledge the validity of the approach (e.g., Olejnik and Hess, 1997; Tamhane, 2009, p. 133; Zwick, 1986). Perhaps that reluctance is a reaction to the fact that many misguided researchers have relied on PLSD when comparing more than three groups, under a false sense of security that the FWER was still "protected" from inflation (Keselman et al., 1998; Zwick, 1986). Nonetheless, given that three-group designs are common in experimental research, it is valuable to inform researchers about specialized tools that are optimal for the three-group case (Levin et al., 1994)—even if those tools are not suitable for other cases. If there is concern that recommending valid use of PLSD in the three-group case could inadvertently encourage invalid use of PLSD in other cases, then one could instead recommend a procedure such as Hayter's (1986), which is equivalent to PLSD in the three-group case yet remains valid for larger numbers of groups (see also Richter and McCann, 2012; Shaffer, 1986). That said, there are often several valid statistical tools to choose from for a given problem, and authors should be cautious about declaring any one approach to be the "best."

References

Albert, J. and Rizzo, M. (2012). *R by example.* Springer.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Cribbie, R. A. and Keselman, H. J. (2003). The effects of nonnormality on parametric, nonparametric, and model comparison approaches to pairwise comparisons. *Educational and Psychological Measurement*, 63(4):615–635.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Félix, V. B. and Menezes, A. F. B. (2018). Comparisons of ten corrections methods for t-test in multiple comparisons via Monte Carlo study. *Electronic Journal of Applied Statistical Analysis*, 11(1):74–91.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal*, 43(8):985–1005.

Fisher, R. A. (1935). *The design of experiments.* Oliver and Boyd.

Frane, A. V. (2015a). Are per-family Type I error rates relevant in social and behavioral science?. *Journal of Modern Applied Statistical Methods*, 14(1):12–23.

Frane, A. V. (2015b). Power and type I error control for univariate comparisons in multivariate two-group designs. *Multivariate Behavioral Research*, 50(2):233–247.

Hancock, G. R. and 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(3):269–306.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, 81(396):1000–1004.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparison procedures*. John Wiley and Sons.

Hogg, R. V. and Craig, A. T. (1965). *Introduction to mathematical statistics*. Macmillan.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386.

Keselman, H. J., Cribbie, R., and Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise/comparisonwise Type I error control. *Psychological Methods*, 4(1):58–69.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., and Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3):350–386.

Keuls, M. (1952). The use of Studentized range in connection with an analysis of variance. *Euphytica*, 1(2):112–122.

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

Kramer, C. Y. (1956). Extensions of multiple range tests to group means with unequal number of replications. *Biometrics*, 12(3):307–310.

Levin, J. R., Serlin, R. C., and Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin*, 115(1):153–159.

Li, D. (2008). A two-step rejection procedure for testing multiple hypotheses. *Journal of Statistical Planning and Inference*, 138(6):1521–1527.

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(1–2):20–30.

Olejnik, S. and Hess, B. (1997). Top ten reasons why most omnibus ANOVA F tests should be abandoned. *Journal of Vocational Educational Research*, 22(4):219–232.

Phillips, A., Fletcher, C., Atkinson, G., Channon, E., Douiri, A., Jaki, T., Maca, J., Morgan, D., Roger, J. H., and Terrill, P. (2013). Multiplicity: Discussion points from the statisticians in the Pharmaceutical Industry Multiplicity Expert Group. *Pharmaceutical Statistics*, 12(5):255–259.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 73(363):479–485.

Ramsey, P. H., Barrera, K., Hachimine–Semprebom, P., and Li, C.-C. (2011). Pairwise comparisons of means under realistic nonnormality, unequal variances, outliers and equal sample sizes. *Journal of Statistical Computation and Simulation*, 81(2):125–135.

Ramsey, P. H. and Ramsey, P. P. (2008). Power of pairwise comparisons in the equal variance and unequal sample size case. *British Journal of Mathematical and Statistical Psychology*, 61(1):115–131.

Ramsey, P. H. and Ramsey, P. P. (2009). Power and Type I errors for pairwise comparisons of means in the unequal variances case. *British Journal of Mathematical and Statistical Psychology*, 62(2):263–281.

R Core Team. (2017). R: A language and environment for statistical computing. https://www.R-project.org/

Richter, S. J. and McCann, M. H. (2012). Using the Tukey–Kramer omnibus test in the Hayter–Fisher procedure. *British Journal of Mathematical and Statistical Psychology*, 65(3):499–510.

Seaman, M. A., Levin, J. R., and Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, 110(3):577–586.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831.

Stephenson, A. G. (2002). evd: extreme value distributions. *R News*, 2(2):31–32.

Tamhane, A. C. (2009). *Statistical analysis of designed experiments: Theory and applications.* Wiley.

Tukey, J. W. (1953). The problem of multiple comparisons. In H. I. Braun (Ed.), *The collected works of John W. Tukey, volume VIII multiple comparisons: 1948–1983.* Wiley.

Zwick, R. (1986). Testing pairwise contrasts in one-way analysis of variance designs. *Psychoneuroendocrinology*, 11(3):253–276.