Pairwise Comparisons of Means when Distributions are Normal Mixtures with Unequal Variance

Mary M. Whiteside, PhD, Mark E. Eakin, PhD
The University of Texas at Arlington
Box 19437
Arlington, TX 76019

Abstract

This paper extends the authors' earlier work, on the power and robustness of multiple comparison procedures when assumptions of normality and equal variance do not hold, to underlying distributions that are normal mixtures, violating both assumptions. Normal mixtures can represent an incorrectly specified design that results in a one-way layout rather than a two way design with a normally distributed response. We examine two approaches: ANOVA F followed by Tukey-Kramer comparisons with the Hayter-Fisher adjustment in degrees of freedom (HFF) and Mehrotra's modification of the Brown-Forsythe F test (changing numerator degrees of freedom as well as the Brown-Forsythe modification of the denominator degrees of freedom) followed by Games-Howell multiple comparisons which adjust Tukey-Kramer comparisons for unequal variances along with the Hayter-Fisher adjustment in Tukey-Kramer degrees of freedom (EHF).

Key words: Power, At least one pair, All pairs, Simultaneous confidence intervals, Robustness,

1. Introduction

Multiple comparison procedures were introduced in the 1940's and 1950's by David Duncan, S.N. Roy, Henry Scheffe, and John Tukey, although ideas appear much earlier in the works of Fisher, Gossett,
and others (Harter 1980, Rao and Swarupchand, 2009). However, conflicts about the importance of controlling family

wide Type 1 error rates with pairwise comparison procedures were still appearing in the literature (especially medical research methodology) 50 years later (Rothman, 1990). The recent explosion of research in genetics and the human genome and brain mapping with functional magnetic resonance imaging (fMRI) have revived an interest in simultaneous procedures such as pairwise comparisons and have motivated the development of relatively new metrics in addition to family wide error rates (FWER) for comparing competing procedures. These new metrics include all pairs (conjunctive) power, any pair (disjunctive) power, and the false discovery rate (FDR) (Bretz, et al 2010).
ANOVA F followed by Tukey-Kramer pairwise comparisons of all means (HSF) is a frequently recommended procedure for analysis of variance in texts (Kutner et al 2004) and software (Hintze, 2012). HSF assumes normal distributions with

equal variance. A recent paper by Ramsey et al (2011) compares the performance of HSF to five other procedures with various non-normal distributions and unequal variances. They conclude that there is not one “best” procedure but that HFF (a significant ANOVA F followed by the Tukey-Kramer test with the Hayter-Fisher adjustment (degrees of freedom k for the Studentized range distribution replaced by (k-1)) avoids problems of complexity and is both robust with respect to alpha and relatively powerful on average. Our paper compares the HFF of Ramsey (2011) with Mehrotra’s modification of the Brown-Forsythe F test (changing numerator degrees of freedom as well as the Brown-Forsythe modification of the denominator degrees of freedom for the F test) followed by Games-Howell multiple comparisons which adjust Tukey-Kramer comparisons for unequal variances along with the Hayter-Fisher adjustment in Tukey-Kramer degrees of freedom (EHF).

2. Design of the study

Our simulation is for a one-way layout with k = 4 levels. The four variances are {1, c, c^2, c^3} where c=1, 1.6, 2, as in Ramsey (2011) and similar to Dunnett (1980). The design of the study is as follows:

normal mixture - 7 levels
wt.=1, c=1
wt.= .5, .2, .058824 x c=1.6, 2
sample size - 3 levels
n=5, 10, 15
Cohen effect size – 6 levels
0, .5, 1, 1.5, 2, 2.5
configuration of means – 2 levels
OEM -m, 0, 0, 0
one extreme mean
TEM -m, 0, 0, m
Two extreme means.

This results in 252 factor level combinations: 6 under H_0: $\square_1=\square_2=\square_3=\square_4$ with normality and equal variance assumptions satisfied, 30 under H_1 with normality and equal variance assumptions satisfied, 36 under H_0 with assumptions violated and 180 under H_1 with assumptions violated. For each factor level combination we simulated 500,000 runs.

The following metrics were used to compare the HFF and EHF methodologies for the four cases described above:

COVERAGE = average proportion of the six simultaneous 95% confidence interval estimates containing the true pairwise difference in means
SIGNIFICANCE OF OMNIBUS = proportion of runs where the initial F test rejected the null of all means equal
ANY PAIR POWER = proportion of runs where the procedure detected at least one true difference in means
ALL PAIRS POWER = proportion of runs where the method detected all truly different group means

FAMILY WIDE ERROR RATE (FWER) = proportion of times that the methodology makes at least one Type I error
AVEARGE TYPE I ERROR RATE = average proportion of Type I errors given all comparisons
FALSE DISCOVERY RATE (FDR) = average proportion of Type I errors given all significant differences

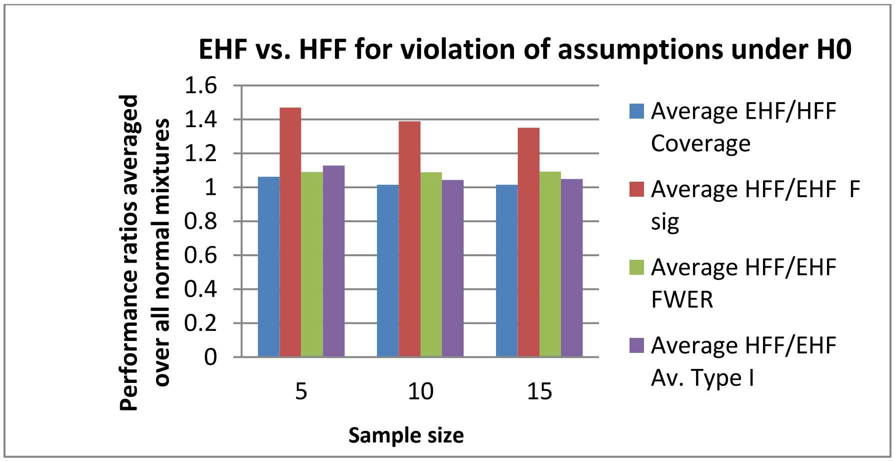
3. Results and conclusions

In order to make some overall comparisons, we have constructed ratios of these metrics so that EHF out performs HFF when the ratio is greater than one and conversely HFF out performs EHF when the ratio is less than one. Figures 1 depicts these results for the third of the four cases previously described.

For the first case, where there are no differences in means and both assumptions are valid, EHF outperforms HFF on average (across factor levels other than n) for n=5 with all four metrics and never underperforms HFF. For the second case, where there are differences in means and both assumptions are valid, EHF outperforms HFF on average for n=5 as measured by FWER and AVERAGE TYPE I ERROR RATE. For the other metrics and sample sizes the two approaches are similar or HFF outperforms EFF. For the third case, where there are no differences in means and both assumptions are violated, EHF outperforms HFF on average (across factor levels other than n) for all n with all four metrics and never underperforms HFF. For the fourth case, where there are differences in means and both assumptions are violated, EHF underperforms HFF with all metrics and sample sizes with the exception of average ALL PAIRS POWER for n=15.

The initial conclusion from these results is that EHF does generally provide more protection against Type 1 errors under H_0 for the cases when all means are equal but not for the cases when both Type 1 and Type 2 errors are possible The FDR is consistently worse with EHF. Future work will be to examine these results without averaging across factor level combinations.

Figure 1:



REFERENCES

Bretz, F., Hothorn, T., and Westfall, P.; *Multiple Comparisons Using R*. Chapman and Hall, London, (2010).

Brown, M. B. and Forsythe, A.B., "The ANOVA and multiple comparisons for data with heterogeneous variances," *Biometrics*, **30** (December 1974), 719-724.

Dunnett, C. W., "Pairwise multiple comparisons in the homogeneous variance, unequal sample, size case," *Journal of the American Statistical Association*, **75** (1980) 789-795.

Games, P. A. and Howell, J. F., "Pairwise multiple comparison procedures with unequal *N*'s and/or variances: a Monte Carlo study," *Journal of Educational and Behavioral Statistics* **1** (1976) 113-125.

Harter, H. L., "Early history of multiple comparison tests," *Handbook of Statistics*, ed., PR Krishnaiah, North Holland, Amsterdam **1**, (1980), 617-22.

Hintze, J., NCSS 9. NCSS, LLC. Kaysville, UT, (2013), www.ncss.com.

Kutner, M.; Nachtsheim, C.; Neter, J. and Li, W., *Applied Linear Statistical Models,* 5th ed., McGraw-Hill/Irwin, (2004).

Rao, C. V., and Swarupchand, U., "Multiple comparison procedures – a note and a bibliography," *Journal of Statistics*, **16** (2009), 66 – 109.

Rothman, K.J., "No Adjustments are needed for multiple comparisons," *Epidemiology*, **1** (1990), 43-46.