The Performance of Robust Procedures for Multiple Comparison Tests under Heteroscedasticity in Psychological Research

by

Linsey LIU

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Abstract

Multiple comparison test (MCT) strategies are widely used in psychological research following an omnibus F test (aka analysis of variance; ANOVA). Traditional methods such as Bonferroni, Tukey's Honestly Significant Difference, and Dunnett's test aim to control the familywise error rate (FWER) but rely on assumption of homogeneity (i.e., variances of residual scores are the same across the levels of an independent variable), which are often violated in practice. To address these violations, robust alternatives—such as those incorporating sandwich estimators or the Plug-In procedure—have been proposed. However, due to differences (e.g., *k* setting, variance ratio (VR) setting, the sample size setting) in prior simulation settings, it remains unclear which procedures perform best under realistic conditions involving heteroscedasticity.

This study systematically evaluated the robustness of MCT strategies, including classical MCT procedures (Bonferroni's test, Tukey's test and Dunnett's test) and robust procedures (sandwich estimator and plug-in procedure) via Monte Carlo simulations under 51 manipulated conditions, including 12 balanced conditions with the same sample size (3 levels of group size * 3 levels of VR + 3 homogeneity conditions) as control group, 19 unbalanced conditions with 3 different group sizes (6 levels of combinations between group sizes & variances * 3 levels of VR + 1 condition where VR=1) , 10 unbalanced conditions with 1 extremely small group (3 levels of combinations between group sizes & variances * 3 levels of VR + 1 condition where VR=1), and 10 unbalanced conditions with 1 extremely large group (3 levels of combinations between group sizes & variances * 3 levels of VR + 1 condition where VR=1). Both classical and robust MCT methods were examined across four

key metrics: Type I error rate, confidence interval (CI) exclusion criterion, width of CI, and statistical power.

Results showed that classical methods (Tukey, Dunnett, and Bonferroni) performed well in balanced conditions but exhibited inflated Type I error rate and reduced power under variance heteroscedasticity or unequal sample sizes. In contrast, robust procedures maintained more stable power and better control of Type I error rate across varied conditions. The CI results further revealed that robust methods provided more flexible and accurate adjustments for interval width associated with a better coverage rate of the true parameter value, particularly in complex and unbalanced designs.

Among robust strategies, Tukey-HC2, Tukey-HC3, and Dunnett-PI consistently demonstrated the best trade-off between power and control of Type I error rate control. Tukey combined with Heteroscedasticity-Consistent (HC) estimators minimized Type I error rate, while Dunnett paired with the PI procedure maximized power. Games-Howell effectively limited false positives but at the cost of lower power, making it more suitable when flexibility is prioritized.

Overall, the findings underscore the importance of selecting MCT procedures that are both statistically powerful and robust to assumption violations. This study offers practical recommendations for psychological researchers, highlighting the advantages of robust methods in enhancing the accuracy and reliability of post hoc inference.

Keywords: Multiple Comparison Tests, Robust Procedures, Sandwich Estimator, Plug-In Procedure, Heteroscedasticity, Monte-Carlo Simulation

Acknowledgement

I would like to thank my advisor, Dr. Li, and my committee members, Dr. Giuliano and Dr. Wang. This thesis would not have been possible without their help, support, and patience. Dr. Li gave me the idea for this project and helped me use my programming skills to make it happen. Dr. Giuliano and Dr. Wang gave helpful feedback on my proposal, pointed out things I had missed, and helped improve the experiment design. I am very thankful for their support during the whole process.

I also want to thank my best friend, Kayson, who kindly shared helpful advice about using multi-core processing during the experiment. Because there was a large amount of data, the program would have taken more than ten days to run. With Kayson's help, the run time was reduced to 21 hours, which made it much easier for me to fix problems and adjust the experiment.

Finally, I would like to thank my family for their patience, encouragement, and unconditional support—for tolerating my stress and for reminding me that there is life beyond academia.

Contents

List of Tables *1*

List of Figures *2*

Chapter I: Introduction *7*

Chapter II: Literature Review *10*

2.1 Classical Multiple Comparison Test Techniques **10**

2.2 Assumptions for Multiple Comparison Tests **11**

2.3 Assumptions Violations in Real-World Research **13**

2.4 Robust Procedures for MCTs **14**

2.4.1 Games-Howell Method **14**

2.4.2 Sandwich Estimator **16**

2.4.3 Plug-In Procedure **18**

2.4.4 Max-t test: A new framework **20**

2.5 Comparisons for Robust Procedures **22**

2.6 Research Gap **26**

2.7 Goal of This Study **29**

Chapter III: Methods *31*

Chapter IV: Results *39*

4.1 Type I error rate **39**

4.2 CI Exclusion Rate **54**

4.3 Confidence interval width **63**

4.4 Power **67**

Chapter V: Discussion *77*

5.1 The Effect of Sample Size, VR and Their Pairing **77**

5.2 Comprehensive Performance of all MCT Methods **78**

5.3 Comparison with Prior Studies **81**

5.4 Summary and Implications **84**

Chapter VI: Conclusion ***86***

Reference ***88***

List of Tables

Table 3.1

Simulation Manipulations for 3 Level MCTs……32

Table 3.2

The Correlation between Sample Sizes and Variance under A Condition ……34

Table 3.3

The Correlation between Sample Sizes and Variance under B & C Condition ……35

Table 3.4

Group Means Under the Assumption of a High Effect Size (Cohen's d = .80) ……36

List of Figures

Figure 4.1

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes Between G1-G2 Contrasts (Balanced Group) ……41

Figure 4.2

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G3 Contrasts (Balanced Group) ……42

Figure 4.3

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2-G3 Contrasts (Balanced Group) ……43

Figure 4.4

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G2 Contrasts (Unbalanced Group & 3 Different Sizes) ……45

Figure 4.5

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G3 Contrasts (Unbalanced Group & 3 Different Sizes) ……46

Figure 4.6

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2-G3 Contrasts (Unbalanced Group & 3 Different Sizes) ……47

Figure 4.7

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G2 Contrasts (Unbalanced Group & 1 Smaller Group) ……49

Figure 4.8

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G3 (Unbalanced Group & 1 Smaller Group) ……50

Figure 4.9

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2-G3 (Unbalanced Group & 1 Smaller Group) ……51

Figure 4.10

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G2 Contrasts (Unbalanced Group & 1 Larger Group) ……52

Figure 4.11

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1-G3 Contrasts (Unbalanced Group & 1 Larger Group) ……53

Figure 4.12

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2-G3 Contrasts (Unbalanced Group & 1 Larger Group) ……54

Figure 4.13

The Differences of Frequencies between CI exclusion rate and P Value (n=30) ……56

Figure 4.14

The Differences of Frequencies between CI exclusion rate and P Value (n=10) ……56

Figure 4.15

The Differences of Frequencies between CI exclusion rate and P Value (n=50) ……56

Figure 4.16

The Differences of Frequencies between CL and P Value under A1……57

Figure 4.17

The Differences of Frequencies between CL and P Value under A2……58

Figure 4.18

The Differences of Frequencies between CL and P Value under A3……58

Figure 4.19

The Differences of Frequencies between CL and P Value under A4……58

Figure 4.20

The Differences of Frequencies between CL and P Value under A5……59

Figure 4.21

The Differences of Frequencies between CL and P Value under A6……59

Figure 4.22

The Differences of Frequencies between CL and P Value under B1……60

Figure 4.23

The Differences of Frequencies between CL and P Value under B2……60

Figure 4.24

The Differences of Frequencies between CL and P Value under B3……60

Figure 4.25

The Differences of Frequencies between CL and P Value under C1……61

Figure 4.26

The Differences of Frequencies between CL and P Value under C2……61

Figure 4.27

The Differences of Frequencies between CL and P Value under C3……62

Figure 4.28

The Relative CI Width under Balanced-Group Condition (n=30, 10, 50) ……64

Figure 4.29

The Relative CI Width under Unbalanced-Group Condition (A1-A6 Condition) ……65

Figure 4.30

The Relative CI Width under Unbalanced-Group Condition (B1-B3 Condition) ……66

Figure 4.31

The Relative CI Width under Unbalanced-Group Condition (C1-C3 Condition) ……67

Figure 4.32

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Balanced Group) ……68

Figure 4.33

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Balanced Group) ……69

Figure 4.34

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 3 Different Sizes) ……71

Figure 4.35

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 3 Different Sizes) ……72

Figure 4.36

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Smaller Group) ……73

Figure 4.37

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Smaller Group) ……74

Figure 4.38

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Larger Group) ……75

Figure 4.39

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Larger Group) ……76

Chapter I: Introduction

According to a review of hundreds of empirical psychological studies published in all issues of the four major Canadian psychology journals between 2013 and 2017 (Alyssa & Harlow, 2017), between-subjects analysis of variance (ANOVA) is still the most widely employed (over 25%) quantitative method in practice. If the null hypothesis of the ANOVA that supports an overall significant mean difference between groups (when the numbers of groups > 3) is rejected, a subsequent step of conducting multiple comparisons testing (MCT) is recommended (Nanda, Mohapatra, & Mahapatra, 2021; Benjamini & Braun, 2002; Neyman & Pearson, 1928). For most psychological studies that use ANOVA, MCT is widely employed by researchers (Sauder, & DeMars, 2019) to further examine all possible pairwise group comparisons and to specifically identify the differences between two treatments or levels of the grouping variable. For comparing between two groups, an independent-samples *t* test should not be directly used because of the inflation of the Type I error rate after repeating independent-samples *t* tests without adjusting for the individual Type I error rate per comparison. A term, familywise error rate (FWER), is defined as the probability that a Type I error (wrongly rejecting the null hypothesis and concluding the result is significant) occurs at least once in a set of hypothesis tests in a study. When researchers conduct separate *t* tests without adjusting for the individual Type I error rate per comparison, the familywise Type I error rate will be inflated and become larger than the Type I error rate specified in each of the individual *t* tests (Ryan, Thomas, 1959).

Some strategies have been proposed to control the probability of the overall false positive or familywise Type I error rate (e.g., Nanda, Mohapatra, & Mahapatra, 2021;

Armstrong, 2014; Dunnett, 1980). These comparisons have different focuses (e.g., adjustments based on the sampling distribution and/or on inequality) and have been incorporated into several conventional MCTs, such as Bonferroni's adjustment (Dunn, & Olive, 1961), Tukey's Honestly Significant Difference (HSD) test (Tukey, 1949), and Dunnett's test (Dunnett, 1964). However, before these strategies can be applied to real-world research, there are some important prerequisites or statistical data assumptions that need to be met. The scores in a dataset need to be normally distributed, are independently observed and the variances of the scores are identical across the levels of the grouping variables. All of the widely used strategies for MCTs mentioned above were developed based on the assumptions of normality and homogeneity, which are often violated in real-world experimental datasets, especially in psychological studies (Alyssa & Harlow, 2017; Field & Wilcox, 2017; Lix, Keselman, & Keselman, 1996). If the assumptions cannot be met, the usefulness of those tests will become less-informative and questionable given that the Type I error rate can no longer be controlled correctly and lead to misleading statistical inferences.

To cope with assumption violations, some mathematical procedures based on classical MCT methods have been proposed and developed. Although their robustness has been tested under different situations, it is hard to determine which procedure is better in varied realistic data situations for MCTs, which are based on entirely different experimental settings in a variety of fields apart from psychological studies. There are no clear guidelines for researchers to choose the most robust procedure that is appropriate in different complex real-world data situations.

Thus, in this study, we focus on examining the performance of several multiple

comparisons tests and robust statistical procedures that have the potential to address assumption violations in psychological research. To evaluate the robustness of these procedures, we use a Monte Carlo simulation, a computer-based experiment that simulate data and repeatedly apply the test procedures with known characteristics (Siepe et al., 2024) considering factors in psychological studies such as group size, variance ratio, and heteroscedasticity, which have been proven to influence test results separately in existing studies across disciplines. The remainder of this thesis is structured as follows: **Chapter II** reviews assumption violations, existing multiple comparison methods, comparisons between methods and their limitations. **Chapter III** describes the simulation design and methodology. **Chapter IV** illustrates the results, followed by a discussion in **Chapter V** and conclusion in **Chapter VI**.

Chapter II: Literature Review

2.1 Classical Multiple Comparison Test Techniques

The aim of most MCT techniques is to control the familywise error rate to be lower than 0.05, if a researcher decides to control the overall Type I error rate as 0.05 in the entire study. One popular technique is Bonferroni's adjustment to *t* test (Armstrong, 2014; Perneger, 1998; Neyman & Pearson, 1928). Bonferroni's *t* test adjusts for the Type I error rate (α) via a formula: α' (the familywise Type I error rate) divided by the number of comparisons (c). Next, researchers check the critical *t* value by α. Taking an independent variable with 3 groups or levels as an example. If the familywise error rate is set to 0.05, a more stringent threshold for the Type I error rate per comparison is required. This can be achieved by dividing the familywise error rate by the number of comparisons (e.g., 0.05 / 3 = 0.0167), which is then used as the significance level for each individual t-test. However, this method is imperfect because the estimated familywise error rate is not exactly equal to 0.05, but slightly smaller. As shown in the calculation below: $\text{FWER} = 1 - (1 - 0.0167)^3 \approx 0.0498$. Bonferroni's adjustment is more conservative and, accordingly, less powerful among the strategies because it sets a stringent error rate per comparison (Armstrong, 2014; Shi, Pavey, Carter, 2012). As one adaptation of the *t test*, Bonferroni's test needs to meet three assumptions: independence, normality and homoscedasticity.

Another common and popular test, Tukey's Honestly Significant Difference (HSD), controls the familywise error rate by recalculating the confidence level for all between-group contrasts (Nanda et al., 2021; Benjamini & Braun, 2002; Keselman & Rogan, 1977). A critical q value is used in this technique to compare the means of groups in pairs. Tukey's

HSD is recommended when every pairwise comparison needs to be conducted, and its significance results are less conservative compared to Bonferroni's *t* test (Keselman & Rogan, 1977). The strategy that compares the means of treatment groups with that of a control group is Dunnett's adjustment (Lee & Lee, 2018; Ludbrook, 1991; Dunnett, 1980; Dunnett, 1955). Using a modified *t*-distribution, this procedure tests only whether an experimental treatment yields significantly higher or lower results than the control group, without comparing the differences between treatment groups. Regarding significance, Dunnett's test is relatively powerful and more likely to detect treatment effects, though it is not as conservative in preventing familywise error (Lee & Lee, 2018; Kim, 2015). However, it is less likely to inflate the familywise error rate compared to Tukey's HSD and other approaches, as it performs only specific contrasts between treatment groups and the control group, thereby effectively reducing the number of comparisons (Hasler, 2014; Tamhane & Logan, 2004; Dunnett, 1980).

2.2 Assumptions for Multiple Comparison Tests

In general, the above-mentioned MCT methods are based on adjustments to the original *t* test for comparing two group means and are commonly applied as follow-up procedures after ANOVA (Sauder, & DeMars, 2019). As a result, they inherit the fundamental assumptions of the *t* test and *F* test:

a) the observations must be independent;

b) the data should follow a normal distribution;

c) the variances across groups must be equal (i.e., homogeneity of variances).

The first assumption is that the tests being conducted are independent, meaning that the outcome of one test does not influence the results of any others. In statistical applications, if there is a correlation between observations due to multi-stage sampling (e.g., students nested within classrooms and schools), adjustments such as Bonferroni's correction may become overly conservative, leading to an increased Type II error rate (Armstrong, 2014).

The second assumption is homogeneity of variances, which requires that the variance within each group being compared is approximately equal. Homoscedasticity is essential for the validity of parametric tests such as ANOVA, as it ensures the accuracy of pooled variance estimates. Violations of the homogeneity assumption not only increase the Type I error rate but also significantly reduce statistical power (Wang, Rodríguez, Chen, et al., 2017), thereby weakening the test's ability to detect true differences between group means.

The third assumption is that the data within each group should be normally distributed. If this assumption is violated, *p* values and confidence intervals may be inaccurately estimated, resulting in invalid conclusions from ANOVA and subsequent multiple comparison tests with significant power loss (Lantz, 2013; Wilcox, 1995). To quantify the impact of three assumptions against the robustness of statistical tests, Knief and Forstmeier (2021) compared the *p*-value and power results when three assumptions were violated respectively based on a linear regression model foundation. Results found that if homogeneity was met while non-independent samples were also fixed by fitting an appropriate random effect structure, the non-normality had little effect on the significance tests (Knief, & Forstmeier, 2021). Since ANOVA can be mathematically represented as a form of linear model (Nelson, & Zaichkowsky, 1979), the result from this study shows that the violation of normality may be

less problematic and leads us to focus more on another violation: heteroscedasticity.

2.3 Assumptions Violations in Real-World Research

However, when conducting *F* tests and subsequent multiple comparison tests (MCTs) in real-world psychological studies, the underlying assumptions are rarely met (Alyssa & Harlow, 2017; Field & Wilcox, 2017; Lix, Keselman, & Keselman, 1996). A landmark study by Micceri (1989), which examined over 440 datasets from published psychological and related studies, found that fewer than 7% of them even satisfied the assumption of normality. Violations of homogeneity were also widespread. In many cases, the variance ratio (i.e., the ratio of the largest variance to the smallest) reached as high as 16, whereas homogeneity assumes this ratio should be close to 1. In a review of 10 studies from a leading clinical psychology journal, Grissom (2010) reported that all datasets exhibited heteroscedasticity, with VRs exceeding 3.2. Building on Micceri's findings, Ruscio and Roche (2012) analyzed 455 newly published psychological studies and introduced a novel metric—standardized variance heteroscedasticity (SVH)—to assess violations of the homogeneity assumption. Their results showed that variance ratios in these studies ranged from 1 to as high as 20,264.36, with nearly one-quarter of the studies having a variance ratio greater than 3. Furthermore, few studies employed adapted procedures to account for assumption violations. A review of publications spanning over 40 years (Lix et al., 1996) revealed that 75% of the studies applied traditional ANOVA on heteroscedastic datasets, and 46% ignored the non-normality of their data. As can be inferred, a substantial number of psychological studies likely violated the homogeneity assumption while still employing traditional *F* tests and MCT

procedures to detect group differences—potentially leading to inflated false positive rates.

This issue cannot be fully addressed by nonparametric tests either (Hollander & Wolfe, 1999). Although nonparametric methods do not require normality, they still assume that all groups have the same distribution shape—an implicit form of the homogeneity assumption.

2.4 Robust Procedures for MCTs

To address this problem, statistics researchers have conducted simulations under conditions of unequal variances to develop robust adaptations or techniques (Hothorn & Hasler, 2023; Midway, Robertson, Flinn, & Kaller, 2020; Lee & Lee, 2018; Hasler, 2014; Herberich, Sikorski, & Hothorn, 2010). In these studies, factors such as sample size, balanced or unbalanced designs, number of groups, and variance structures were systematically manipulated to simulate datasets that closely resemble real-world scenarios. The Type I error rate (using a significance level of 0.05) and statistical power were the most common metrics used to evaluate the robustness and effectiveness of the test techniques. Midway et al. (2020) identified one approach, known as the Waller-Duncan *k*-test, which can address heteroscedasticity by adjusting for variance differences, although it has not been widely adopted. Notably, this method does not rely on the same metrics as other normal MCT procedures (*p* value) to determine significance (Waller & Duncan, 1969), which makes it relatively difficult for statistics researchers to directly compare it with other procedures (Midway et al., 2020).

2.4.1 Games-Howell Method

Based on Tukey's HSD test, Games-Howell test was developed to cope with

heteroscedasticity and unequal sample sizes (Games, Howell, 1976). It is more robust than Tukey's HSD except for really small samples (fewer than 5 samples) under heteroscedasticity, but still needs to meet the assumption of normality and independence. Both the Games-Howell and Tukey's HSD tests utilize studentized range statistic (q-distribution) to calculate significance. Compared with Tukey's test, Games-Howell test uses variances of each group and sample sizes of each group, instead of mean square within (MS_{within}), to calculate the standard error without reliance on homogeneity assumptions, which can be mathematically written as:

$$SE_{ij} = \sqrt{\frac{S_i^2}{n_i} + \frac{S_j^2}{n_j}},$$

where S_i^2 is the standard deviation, n_i is the group sample size. Besides, unlike the Tukey's HSD, which directly uses $N - k$ as the degree of freedom, GH uses a Welch correction to dynamically adjust the degree of freedom according to the change of sample variance and sample size, weakening the impact of unequal variance and unbalanced sample size. Thus, the *df* for each comparison will be different, which can be calculated by:

$$df' = \frac{(S_i^2/n_i + S_j^2/n_j)^2}{\frac{(S_i^2/n_i)^2}{n_i - 1} + \frac{(S_j^2/n_j)^2}{n_j - 1}}$$

Compared with classical MCT methods, the GH method is particularly good at handling Type I error rate and unequal groups or variances. It consistently narrows the width of confidence intervals, which enhances precision, and helps maintain control over the FWER. Besides, another key benefit is its robustness to violations of the normality assumption (Day & Quinn, 1989). The GH is relatively conservative. However, when the sample size is small, it can be overly permissive (Toothaker, 1991). Besides, while the GH

provides a narrower confidence interval and higher power, there'll be a risk of occasionally increasing the FWER (Tamhane, 1979; Rafter et al., 2002; De Muth, 2006). Proven to be more robust than some normal MCT methods (e.g. Tukey's HSD and Bonferroni's test), it still requires further comparison with other robust procedures.

2.4.2 Sandwich Estimator

Some statistics researchers have focused on parameter-based approaches to develop robust estimators that adjust for unequal variances (Hothorn & Hasler, 2023; Hasler, 2016; Hothorn, 2008; Freedman, 2006; Huber, 1967). The sandwich estimator, also known as the robust covariance matrix estimator, was developed to compute a robust variance by constructing a parameter vector and reshaping the covariance matrix (Freedman, 2006). Sandwich estimators are widely applied to various types of heteroscedastic data, such as cross-sectional data—where it is referred to as the Heteroscedasticity Consistent (HC) estimator—and time-series data—where it is known as the Heteroscedasticity and Autocorrelation Consistent (HAC) estimator (Zeileis, 2004; Zeileis, 2006). In the context of multiple comparison tests, the HC estimator is particularly useful as it corrects biased and inconsistent variance estimates among treatment groups tested simultaneously.

The mathematical formulation of the sandwich estimator is outlined as follows. In a parameter estimation with null-hypothesis test, θ is denoted as the unknown parameter, a $1 \times p$ dimension vector for a population whose observations are X_i ($i = 1, 2, \ldots, n$), and observed values are Y_i ($i = 1, 2, \ldots, n$), while $\hat{\theta}$ is the estimate for it. Our goal is to infer the true value of $\hat{\theta}$, which is denotes as θ_0, to help estimate what the population looks like from the sample. For a sample with observed values from Y_1 to Y_n, let's assume $fi(y|\theta)$ is the function of

probability density. Its likelihood $\mathrm{L}(\theta) = \prod_{i=1}^{n} fi(y|\,\theta)$. Then the log likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \log\left(fi(y_i|\theta)\right)$$

According to the Fisher Information Matrix which is defined as the expectation for the negative second partial derivative of the log likelihood function in order to characterize the loss, we could get a symmetric $p \times p$ dimension matrix,

$$-E_\theta\, l''(\theta) = -E_\theta\left(\sum_{i=1}^{n} log[fi(y_i|\theta)]\right)''$$

while $-E_\theta\, l''(\theta) = E_\theta\left(l'(\theta)^T \times l'(\theta)\right)$, which is larger than 0.

T stands for transposition. The inverse of this matrix can be used to calculate the covariance matrix associated with maximum-likelihood estimates (e.g. Petz, 2002; Abt & Welch, 1998). To make a MLE which is the foundation for a sandwich estimator, considering that $l(\theta)$ is a quadratic, its maximum can be calculated by:

$l'(\theta) = 0$. Expand $l(\theta)$ in a Taylor series around the true value θ_0, we could infer that:

$$\begin{aligned} l'(\theta) &= [\, l(\theta_0) + l'(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T l''(\theta_0)(\theta - \theta_0)^2 + \cdots]' \\ &\approx [l(\theta_0) + l'(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T l''(\theta_0)(\theta - \theta_0)^2]' \\ &= l'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T l''(\theta_0) \times 2 \\ &= l'(\theta_0) + (\theta - \theta_0)^T l''(\theta_0) = 0 \end{aligned}$$

Thus,

$$(\hat{\theta} - \theta_0)^T = -l'(\theta_0)[l''(\theta_0)]^{-1}$$

From the multidimensional central limit theorem (Li, Tang, Charon & Priebe, 2020; Durieu & Tusche, 2014; Roberts & Tweedie, 1996), an independent and identically distributed (i.i.d.) vector θ shows a convergent distribution of:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, s(\hat{\theta}))$$

, where $s(\hat{\theta})$ is a covariance matrix for $\hat{\theta}$. From the idea of sandwich estimator, the θ_0 is directly inferred from $\hat{\theta}$, so

$$s(\hat{\theta}) = cov_{\theta_0}\hat{\theta} = [-l''(\theta_0)]^{-1}[cov_{\theta_0}l'(\theta_0)][-l''(\theta_0)]^{-1}$$

$$= [-l''(\hat{\theta})]^{-1}[cov_{\hat{\theta}}l'(\hat{\theta})][-l''(\hat{\theta})]^{-1}$$

In this equation, $[-l''(\theta_0)]^{-1}$ looks like the bread for the sandwich, holding a piece of 'meat' between them. And the 'meat' could be calculated by a covariance transformation (Freedman, 2006):

$$cov_{\hat{\theta}}l'(\hat{\theta}) = \sum_{i=1}^{n}(\log fi(y_i|\hat{\theta}))^T \log fi(y_i|\hat{\theta})$$

The sandwich estimator is aimed at providing a robust SE, which can be used to replace the covariance estimate that is based on correct assumptions in classical procedures for linear models (Zeileis, 2006), especially of small samples. Herberich et al., (2010) first examined it in MCT, and verified its robustness related to a classical Tukey test when the homoscedasticity assumption was violated. When utilizing it as an asymptotically unbiased σ^2, modified MCT also needs to adjust the distribution and degrees of freedom for it (see Hasler, 2016). As a result, the individual H0 test for each contrast shares a joint distribution and the same degree of freedom, comparing the parameter θ with the same critical value. After being applied in statistical methods across different subjects, sandwich estimator (with a family of HC, HC1, HC2 and HC3) now is available in R packages called 'multcomp' (Hothorn, Bretz, Westfall, Heiberger, Schuetzenmeister, 2014) & 'sandwich' (Lumley & Zeileis, 2014; Zeileis, 2006; Zeileis, 2004).

2.4.3 Plug-In Procedure

Another promising estimator for MCT is the plug-in (PI) procedure (Hasler, 2013;

Hasler & Hothorn, 2008; Dunnett, 1980; Tamhane, 1979). It also uses an estimated variance for test. However, different from sandwich estimator that only focus on adjusting the σ^2 for one joint distribution, PI procedure aims to build several multivariate *t* distributions for the contrasts, each of which has a separate degree of freedom. The idea from Games and Howell (1976), which is the inspiration for PI procedure, was that a multiple comparison test hardly had a perfect joint distribution, thus separate comparison-specific distribution would be a better approach (Hasler, 2013). Plugging the estimated σ_i^2 into the correlation of *t* statistic calculation, this method didn't adopt a pooled variance for a joint *t* test. As a result, for each contrast from 1 to *q*, the different S_i^2 builds its own test statistic from a *q*-variate distribution with correlation matrix *R**, where the individual σ_i^2 played an important role. Since PI procedure can incur both conservative and liberal results when the difference of unequal variances and sample allocation vary, several studies were conducted to detect its characteristics. Hasler (2013) compared PI procedure with other three MCT plans: HOM, GH and HTL procedure. The first one used the original unadjusted *t* test statistic, assuming that data are ideally homogeneity. The second one is developed by Games and Howell (1976), now known as the predecessor of PI procedure. It also used separate recalculated *t* distributions and compared them with their distinct quantile for the contrasts, but this lost the step to further adjust the *t* and its *df* by correlation matrix *R* with σ_i^2. The last one used another way to adjust *R* for unbalanced groups, taking the average of correlation with control group while using the average of *df*. Researchers considered four different scenarios for a 3-group post hoc test using these three procedures: (a) balanced groups with gradually increasing standard deviations (SD), where the last group had the largest SD; (b) two

balanced groups with different SDs, and one small group sharing the smaller SD with one of the balanced groups; (c) two balanced groups with the same SD, and one small group with an extremely larger SD; and (d) three balanced groups with the same SD. After 100,000 simulations, results showed that the familywise error rate (FWER) of the HOM procedure was notably volatile except in scenario (d), which met the ideal assumptions. In contrast, the FWERs of the GH and HTL procedures were either conservative or liberal across the different scenarios. Meanwhile, the PI procedure consistently maintained the nominal α level of significance. Thus, the PI procedure is highly recommended by statisticians. Similar conclusions can be found in previous studies (Mondal, Sattler, Kumar, 2023; Tamhane, 2023; Pallmann & Hothorn, 2016).

2.4.4 Max-t test: A new framework

Hothorn (2008) utilized simultaneous inference procedures combined with correlated parameter estimates to test individual null hypotheses (H0) for each comparison among k groups simultaneously. In this procedure, researchers only need to assume that the parameter estimates follow asymptotic normality (inferred from equations), while the associated covariance matrix can be consistently estimated. Since ANOVA is technically a linear model, this procedure can unify group contrasts (i.e., $\mu_2 - \mu_1, \mu_3 - \mu_1, \ldots, \mu_k - \mu_1$) into a vector θ and infer its distribution from error terms, fitting well within this framework. Because the estimates do not conflict with subsequent comparison procedures, this method can be embedded with many classical multiple comparison tests, such as Tukey's test and Dunnett's test, and combined with sandwich estimators to relax the classical assumptions of normality and homoscedasticity.

Based on this framework, several recent studies have examined its robustness across different situations via simulations and real-world datasets (Herberich, Sikorski, & Hothorn, 2010). Herberich et al. (2010) generated datasets in which most assumptions were violated, including non-normal data distribution (right skewness), unequal variances, and unbalanced sample sizes. The number of groups was set to 4, resulting in 6 contrasts following ANOVA. For the analysis, they combined the sandwich estimator (HC3) with this framework (termed '*max-t*') and compared it with the Tukey-Kramer test (also known as Tukey HSD). Total sample sizes (*N*) of 60, 120, 180, and 240 were considered in each trial to simulate small, moderate, and large real-world sample sizes. Results showed that for datasets with equal variance and balanced group sizes, the significance levels of both methods approximated the nominal α-level of 0.05, although the false positive rate of *max-t* was slightly higher than that of Tukey HSD. When variances were unequal, specifically when smaller groups had relatively smaller variances, Tukey HSD was more conservative than the *max-t* test regardless of sample size. However, when a larger group had a small sample size, the Tukey HSD test exhibited a severely inflated familywise error rate, whereas the max-*t* test remained relatively robust and close to the nominal 0.05 level. Moreover, this robustness improved as sample size increased. Other relevant studies have confirmed these characteristics, concluding that the *max-t* test combined with the sandwich estimator is robust under unbalanced sample sizes and heteroscedasticity, especially when sample sizes and variances are negatively paired (Pallmann & Hothorn, 2016; Konietschke, Bösiger, Brunner, & Hothorn, 2013; Hothorn & Brunner, 2012). However, it should be noted that this method is not recommended when the total sample size is small (e.g., 60 for 4 groups).

2.5 Comparisons for Robust Procedures

Combining the aforementioned robust methods, several researchers have compared existing robust procedures for multiple comparison tests (MCTs) to identify their suitable application scenarios (Herberich, Sikorski, & Hothorn, 2010; Hasler & Mario, 2014; Hothorn & Mario, 2023). Herberich, Sikorski, and Hothorn (2010) conducted a comparison between the sandwich estimator (HC3) and the classical MCT method, Tukey's HSD. Focusing on post hoc contrasts following a one-way ANOVA with four levels ($k = 4$), they set the sample sizes of the four groups as an arithmetic sequence defined by $n_k = n_1 + 0.2 \times k \times n_1$, with n_1 taking values 10, 20, 30, or 40. Five variance scenarios were considered: one homogenous variance condition and four heteroscedastic sequences. Two unequal variance sequences were specified as (3, 5, 7, 9) and (0.14, 0.18, 0.29, 0.35), respectively, each positively paired with the sample size sequence. Additionally, reversed versions of these variance sequences were included, producing a negative pairing with the sample sizes. After 1,000 simulation runs, the results demonstrated that the sandwich estimator consistently outperformed Tukey's HSD in all heteroscedastic scenarios, exhibiting lower familywise error rates and higher statistical power.

Hasler and Mario (2014) compared HC3 and PI procedure for contrasts between groups ($k = 3, 4, 5$). The sample size was unbalanced and extremely small. The smallest group only contained 2 observations, while n_k followed an arithmetic sequence of $n_k = n_{k-1} + 2$. There were three variances settings for these groups. a. variances were homoscedastic; b. smallest sample had the smallest SD; c. smallest sample had the largest SD. Since PI procedure and

HC3 can all be embedded on classical MCT methods, this research covered 5 methods for comparison: Dunnett test, Tukey HSD test, William's test, Changepoint test and Average test. The only one metric to be compared was familywise error rate under different tests. From the result, a. the number of treatments (k value) had a stronger influence on HC3 than PI. The FWE for PI was almost the same across different k values, while that for HC3 would be higher when k value increased. This gives evidence that for groups with extremely small sample sizes, HC3 is more liberal when treatments increases. b. For samples with homoscedasticity, both PI and HC3 performed worse than classical methods, while PI for Dunnett test was relatively close to the α level. c. For samples whose sizes paired positively with variances, the performance of robust procedures depended on the tests for MCTs. To be specific, when using Dunnett or Tukey tests, PI procedure was robust and consistent at the α level, while HC3 was more liberal only if $n_1 = 2$ (extremely small) and gradually become robust when $n_1 \geqslant 3$. When using Williams, Changepoint and Average tests for MCTs, HC3 was particularly conservative for contrasts between 3 groups ($k = 3$) and more robust than PI procedure no matter the sample size.

c. When sample sizes were negatively paired with variances, the results differed notably. The HC3 procedure exhibited considerable liberality across methods under this condition, particularly when the smallest group size n_1 was very small (e.g., $n_1 = 2$ or 3). Conversely, the PI procedure shared this liberality only for very small n_1, but it demonstrated robustness as n_1 increased. Furthermore, with increasing group sample sizes, the HC3 method's robustness markedly improved, especially in the Changepoint test scenario.

In summary, this study examined the interaction between robust statistical procedures

and sample size-variance pairings. It validated the PI procedure's performance under heteroscedasticity, with an emphasis on extremely small sample sizes. Additionally, the study applied these findings to a biological dose-response dataset with small samples. The results indicated that using the HC3 estimator could erroneously classify a dangerous dose of 75 mg/kg as safe, highlighting risks of false conclusions in small-sample contexts. However, such extremely small sample sizes are uncommon in published psychological research, where ANOVA with very small *n* is rarely preferred or reported formally. Therefore, for practical psychological research, greater attention should be paid to more typical sample size and variance scenarios when selecting robust procedures.

Hothorn and Mario (2023) further extended this line of research by conducting simulations comparing multiple robust methods under heteroscedasticity, adding a Bonferroni-adjusted Welch t-test for comparison with the PI procedure and the sandwich estimator. Notably, the Bonferroni-Welch test is best suited for small numbers of groups (*k*) because it ignores correlations between marginal tests. Simulations were conducted under both the null hypothesis (*H0*, no group mean differences) and the alternative hypothesis (*H1*, presence of mean differences). This dual approach allowed for assessment of both false positive rates (familywise error rate) and false negative rates (power loss). Consequently, the study's evaluation metrics encompassed both familywise error rate and statistical power, providing a comprehensive assessment of each method's performance.

Focusing on the MCT between 4 treatment groups, this study considered two types of sample size *n*: small (n_1 = 6 or 9, n_k = 6 or 5) and moderate ($n_1 = n_k = 20$). The classical MCT methods adopted in this study were Dunnett's test, where all treatment groups were made

contrasts with the control group (n_1). For the variance setting, the variances for moderate groups were all homoscedastic. In contrast, a significantly larger variance (s = 4) would be given for one group in turn for small groups and others would be the same (s = 1). After 5000 simulations for H0 and 2000 simulations for H1, results showed that: a. For tests conducted for small sample size under H0, the original Dunnett's test is unacceptably liberal with inadequate *power* estimations; When variances were unequal, HC3 procedure was found more conservative than the Dunnett's test, but more liberal than the PI procedure and Bonferroni-Welch test; b. For tests conducted for small sample size with heteroscedasticity, the original Dunnett's test will cause not only power loss due to variance increasing in exactly the considered group (e.g. comparing n_2 and the control group while n_2 had the largest variance), but also power loss due to variance increasing in a different group (e.g. comparing n_2 and the control group while n_4 had the largest variance). For balanced small sizes under H1, the HC3 procedure was more powerful than the other two procedures and would cause less power loss due to variance increasing in the considered group. And when it comes to unbalanced design, HC3 was also more powerful than the other two robust procedures. c. If the sample sizes were moderate (n = 20), HC3 were found to exactly control the familywise error rate (better than the PI procedure) under heteroscedastic situations for H0, while the Bonferroni-Welch is a bit conservative. Meanwhile, the results under heteroscedastic situations for H1 showed that HC3 brought less power loss than the other two robust procedures without causing power loss due to a larger variance from a different group. Combining all the simulation results, the Bonferroni-Welch-test and PI procedure were recommended for small sample sizes, whereas the sandwich estimator was recommended for

moderate or higher sample sizes.

To summarize these studies, there is no universal procedure that performs best for all situations. The most important insight we can learn is that different sample sizes have different preferences for procedures. For small or tiny small sample sizes, Mario (2014) verified the PI procedure performed better than HC3, especially when the variances of groups were large. While Hothorn & Mario (2023) concluded that the Bonferroni-Welch-test was better than the PI procedure regarding the power loss. For moderate or medium sample sizes, the HC3 procedure was considered more powerful than Tukey's test and original Dunnett's test, but the discussion on heteroscedasticity of medium-size samples was rare.

2.6 Research Gap

Based on the review of the previous studies, it remains challenging for researchers to determine which robust procedure is most suitable for specific real-world scenarios. This difficulty arises because prior research has predominantly focused on isolated factors, such as the number of groups (e.g., Hasler & Mario, 2014), different hypotheses (e.g., Hothorn & Mario, 2023), or sample size relationships (e.g., Herberich, Sikorski, & Hothorn, 2010), thereby lacking a unified framework for comparison. The heteroscedasticity in study designs, particularly in the number of groups (k), sample sizes (n), and variance settings (SD), impedes the derivation of comprehensive conclusions.

Firstly, the variability in the independent variables across studies is substantial. For example, findings from MCTs conducted on three-group designs cannot be directly compared with those involving five groups, as the properties of robust procedures can shift with the

number of groups. Additionally, although all the aforementioned studies aimed to investigate heteroscedasticity effects, their methods to impose unequal variances differed markedly. One study employed extremely small variance values (e.g., 0.14), another combined equal variances with a single extreme variance, and yet another did not specify variance values at all. Similarly, the range of sample sizes explored varied considerably, offering limited insight into the effective operational bounds of the procedures. Furthermore, many studies did not include promising or commonly used alternative robust procedures, restricting meaningful cross-method comparisons.

Moreover, these studies often overlooked critical factors such as variance ratio (VR), sample size ratio, and the interplay between sample sizes and variances. The variance ratio—defined as the ratio of the largest variance to the smallest—is a key metric for assessing heteroscedasticity (Field & Wilcox, 2017; Ruscio & Roche, 2012; Grissom, 2000), yet it has been rarely addressed in robust MCT literature. Notably, Blanca et al. (2018) demonstrated that VR values exceeding 1.5 threaten the robustness of the F test in unbalanced designs, whereas lower VRs have minimal impact. Without explicit consideration and quantification of VR, it is difficult to formulate generalizable conclusions. Most prior robust MCT studies set variance differences as fixed increments rather than ratios, and some arbitrarily selected variance values spanning wide and irregular ranges. Given the fundamental role of VR in classical F tests, incorporating it systematically is essential when evaluating and comparing robust MCT procedures.

Besides, the patterns for sample sizes in previous studies were always set in a simple way, (equal or monotonic increase). The sizes across studies varied largely, making it difficult

to compare. Multiple studies on sample sizes found that equal group sizes would make ANOVA more robust when affected by heteroscedasticity (Blanca, et al., 2018; Monder, 2010; Lee & Ahn, 2003). However, it's hard to tell the exact where the line between equality and inequality is. Hothorn and Mario (2023) considered a potential situation for real-world studies, that one or two groups may be outstanding after the data cleaning while others shared the same size. Due to the complex situations and the influence of sizes on robustness, the setting of sample sizes should be considered. Blanca et al. (2018) utilized a new metric for sample size, coefficient of sample size variation (Δn), to measure the amount of inequality in sample sizes. It is calculated by dividing the SD of group size by its mean. According to their definition, a range of [0.141, 0.178] is for a low Δn, a range of [0.316, 0.334] is for a medium Δn, and a range of [0.491, 0.521] is for a large Δn. Thus, our study plans to take not only the sample size values but also the variation into experimental design.

Meanwhile, the relationship between sample sizes and variances needs attention. Early studies on classical *F*-test found that the *F*-test tended to be conservative when the pairing was positive, and it tended to be liberal when the pairing was negative (Blanca et al., 2018; Glass, Peckham & Sanders, 1972). Based on findings of Hasler and Mario (2014), the FWER of three procedures under hetero I situation (where the pairing of sample sizes and variances are positive) is quite different from the error rate under hetero II situation (where the pairing of sample sizes and variances are negative). For extremely small sample sizes, HC3 was a bad procedure for hetero II situation, but a robust tool for hetero I situation. We could calculate the pairing by the correlation between group size and variance, range between [-1,1]. In addition to the monotonic positive and negative pairing, a group with medium size in

a real study is likely to have the smallest/largest variance. Thus, the effect of pairing should be explored thoroughly, especially after introducing the measure of sample size deviation.

Lastly, for the dependent variables, FWER was the only one metric that were always used in previous studies to measure the performance of procedures, while power was seldom used. But for a real-world study, there are more metrics we need to concern. According to Counsell and Lisa (2017) who calculated and reported important metrics from published psychological research, effect size and *p* values were the most used metrics (91.4%, 92.7%, respectively). Confidence interval and standard error would be included for some cases (10.6%, 24.6%, respectively). To examine the procedures as comprehensive as possible, it's necessary to cover more metrics in further study, especially the effect size.

2.7 Goal of This Study

This study aims to compare and promote robust procedures for multiple comparison tests (MCTs), with the goal of identifying the most appropriate methods across various heteroscedastic conditions. Building on previous findings regarding the robustness of the PI procedure and the sandwich estimator, the current research focuses on more realistic experimental designs typical of psychological studies, excluding extremely small group sizes ($n \leqslant 5$), which are seldom considered reliable for formal psychological analysis. The study systematically varies key factors such as the number of groups, variance ratios, sample size imbalances, and the pairing patterns between variances and group sizes to reflect practical research scenarios. Through extensive simulation, this study aims to derive generalizable conclusions regarding the optimal scope of application for each robust procedure under

heteroscedasticity.

Chapter III: Methods

To comprehensively investigate multiple parameters affecting data distributions under various conditions typical of psychological research, this study employs Monte Carlo simulation, a computational technique that utilizes random sampling and statistical methods to approximate numerical solutions to complex problems (Morris et al., 2019). It generates large datasets by drawing pseudo-random samples—often from a specified distribution such as the normal distribution—based on predefined parameters including mean, sample size, and standard deviation. In this study, Monte Carlo simulation is used to create datasets reflecting realistic psychological study conditions, such as unbalanced sample sizes and unequal variances, which are typically beyond researchers' control. The aim is to compare different multiple comparison test (MCT) methods, robust procedures, and their combinations to determine which techniques yield the lowest familywise error rate (FWER) while maintaining adequate statistical power.

Leveraging R Studio's mature computational environment, which supports intensive simulations, enables the generation and analysis of all possible datasets under the designed conditions, thus providing a robust framework for evaluating the performance of various procedures (Burton, Altman, Royston, & Holder, 2006; Angelis & Young, 1998). To specifically address violations of the homogeneity of variance assumption, the rnorm() function in R is used to generate normally distributed data. Given that the precision and accuracy of evaluating Monte Carlo simulation results improves with the number of repetitions, each simulation scenario was replicated 5000 times.

Based on the literature review, the main parameters identified as critical—yet

inconsistently addressed across previous studies—are summarized as follows: **a. the number of groups**. In this study, we went through all combinations of data distribution and parameters setting under one common condition of k = 3. **b. the ratio of variances.** The manipulations cover the ideal case of equal variance to the extreme case where VR equals to 16. **c. the group sample size**. For each group of treatment, the small sample size (n = 10), moderate sample size (n = 30), and large sample size (n = 50) were combined and put into manipulations. **d. the pairing between variance and group size**. Apart from monotonically positive and monotonically negative, other types of pairing were also taken into consideration. For example, the smallest n has an intermediate variance, while the largest n has a smallest variance.

Overall, the variables are manipulated as follows:

Table 3.1

Simulation Manipulations for 3 Level MCTs

design	n1	n2	n3	N	n size	v1	v2	v3	VR	pairing
balanced	30	30	30	90	mode-rate size	1	1	1	1	/
	30	30	30	90		1	1.25	1.5	1.5	
	30	30	30	90		1	2	4	4	
	30	30	30	90		1	4	16	16	
	10	10	10	30	small size	1	1	1	1	/
	10	10	10	30		1	1.25	1.5	1.5	
	10	10	10	30		1	2	4	4	
	10	10	10	30		1	4	16	16	
	50	50	50	150	large size	1	1	1	1	/
	50	50	50	150		1	1.25	1.5	1.5	
	50	50	50	150		1	2	4	4	
	50	50	50	150		1	4	16	16	
unbalan	10	30	50	150		1	1	1	1	/

ced	10	30	50	90		1	1.25	1.5	1.5	min n + min SD; max n + max SD
	10	30	50	90	A1	1	2	4	4	
	10	30	50	90		1	4	16	16	
	10	30	50	90		1	1.5	1.25	1.5	min n + min SD; middle n + max SD
	10	30	50	90	A2	1	4	2	4	
	10	30	50	90		1	16	4	16	
	10	30	50	90		1.25	1	1.5	1.5	middle n + min SD; max n + max SD
	10	30	50	90	A3	2	1	4	4	
	10	30	50	90		4	1	16	16	
	10	30	50	90		1.25	1.5	1	1.5	max n + min SD; middle n + max SD
	10	30	50	90	A4	2	4	1	4	
	10	30	50	90		4	16	1	16	
	10	30	50	90		1.5	1	1.25	1.5	middle n + min SD; min n + max SD
	10	30	50	90	A5	4	1	2	4	
	10	30	50	90		16	1	4	16	
	10	30	50	90		1.5	1.25	1	1.5	max n + min SD; min n + max SD
	10	30	50	90	A6	4	2	1	4	
	10	30	50	90		16	4	1	16	
unbalanced	10	30	30	70	B0	1	1	1	1	/
	10	30	30	70		1	1.25	1.5	1.5	min n has min SD
	10	30	30	70	B1	1	2	4	4	
	10	30	30	70		1	4	16	16	
	10	30	30	70		1.25	1	1.5	1.5	min n has middle SD
	10	30	30	70	B2	2	1	4	4	
	10	30	30	70		4	1	16	16	
	10	30	30	70		1.5	1	1.25	1.5	min n has max SD
	10	30	30	70	B3	4	1	2	4	
	10	30	30	70		16	1	4	16	
	50	30	30	110	C0	1	1	1	1	/
	50	30	30	110		1	1.25	1.5	1.5	max n has min SD
	50	30	30	110	C1	1	2	4	4	
	50	30	30	110		1	4	16	16	

50	30	30	110		1.25	1	1.5	1.5	max n has middle SD
50	30	30	110	C2	2	1	4	4	
50	30	30	110		4	1	16	16	
50	30	30	110		1.5	1	1.25	1.5	max n has max SD
50	30	30	110	C3	4	1	2	4	
50	30	30	110		16	1	4	16	

Note: For all A conditions (A1 to A6), the 3 group sizes were different. The difference between A1 to A6 was the sample-size-variance pairing.

The detailed design are shown as follows. Under unequal sample size (A) conditions, most of the possible scenarios of variance heteroscedasticity have been simulated through permutation and combination. The correlation between sample size and variance is calculated as an evaluation metric under the corresponding conditions, as shown in **Table 3.2**.

Table 3.2

The Correlation between Sample Sizes and Variance under A Condition

Name	Sample Size Ratio	Order of Variance	Correlation
A1	10: 30: 50	1: 2: 4	0.982
A2	10: 30: 50	1: 4: 2	0.327
A3	10: 30: 50	2: 1: 4	0.655
A4	10: 30: 50	2: 4: 1	-0.327
A5	10: 30: 50	4: 1: 2	-0.655
A6	10: 30: 50	4: 2: 1	-0.982

Note. In this table, take VR=4 as an example to illustrate the correlation.

Taking VR = 4 as an example, it can be seen that in A1, the ratio of sample size to variance is approximately 1.0, while in A6, the ratio is approximately -1.0. Compared to A2 and A4, the correlation between sample size and variance in A3 and A5 is relatively larger, which may be due to the bias introduced by combining the highest variance with the smallest/largest groups. Thus, the correlation ranking for the six scenarios is as follows:

$$-1 < A6 < A5 < A4 < 0 < A2 < A3 < A1 < 1$$

Specifically, in A1, the smallest group has the smallest variance, and the largest group has the largest variance; in A2, the smallest group has the smallest variance, while the largest group has a moderate variance; in A3, the smallest group has a moderate variance, and the largest group has the largest variance; in A4, the smallest group has a moderate variance, and the largest group has the smallest variance; in A5, the smallest group has the largest variance, and the largest group has a moderate variance; in A6, the smallest group has the largest variance, and the largest group has the smallest variance. Through this approach, we simulated a comprehensive range of possible variance heteroscedasticity scenarios.

The same procedure happens for B & C conditions, where there are 2 equal groups and 1 extremely larger/smaller group. As shown in **Table 3.3**, the three combinations represent groups with different sample sizes, having the smallest, medium, and largest variances. For the B conditions, the correlation between sample size and variance is negative only when the smallest group has the largest variance. For the C conditions, the correlation is positive only when the largest group has the largest variance.

Table 3.3

The Correlation between Sample Sizes and Variance under B & C Condition

Name	Sample Size Ratio	Order of Variance	Correlation
B1	10: 30: 30	1: 2: 4	0.756
B2	10: 30: 30	2: 1: 4	0.189
B3	10: 30: 30	4: 1: 2	-0.945
C1	50: 30: 30	1: 2: 4	-0.756
C2	50: 30: 30	2: 1: 4	-0.189
C3	50: 30: 30	4: 1: 2	0.945

Note. In this table, take VR=4 as an example to illustrate the correlation.

To ensure reliable results, each combination of the conditions described above was simulated with 5,000 replications. The empirical Type I error rate (at a significance level of

0.05), statistical power, Confidence Interval (CI) width, and CI exclusion rate were recorded as evaluation metrics. For the metrics related to Type I error rate, CI exclusion rate, and confidence interval width, the simulations were conducted under the null hypothesis *H0*, where all group means are equal to zero. Conversely, for the evaluation of power, the simulations were performed under the alternative hypothesis *H1*, in which specified mean differences between groups exist.

Specifically, the mean differences between Group 1 and Group 2, as well as those between Group 2 and Group 3, were set to ensure the alternative hypothesis was met. In this design, mean differences were expressed in terms of effect size to mitigate the confounding influence of varying variances across groups. This approach allows for an accurate assessment of the multiple comparison methods' performance across different variance conditions while maintaining a relatively consistent level of statistical power for each group comparison.

Table 3.4

Group Means Under the Assumption of a High Effect Size (Cohen's d = 0.80)

VR	μ **(1)**	μ_{min} **(2)**	μ_{min} **(3)**	σ_{pooled} **(C12)**	σ_{pooled} **(C13)**	***Cohen's d***
1: 1: 1	0	0.8	0.8	1	1	0.8
1: 1.25: 1.5	0	0.906	1.020	1.132	1.275	0.8
1: 2: 4	0	1.265	2.332	1.581	2.915	0.8
1: 4: 16	0	2.332	9.069	2.915	11.336	0.8

Note: VR = variance ratio, μ (1) = mean of group 1, μ (2) = mean of group 2, μ (3) = mean of group 3, σ_{pooled} (C12) = the pooled SD between group 1 and group 2, σ_{pooled} (C13) = the pooled SD between group 1 and group 3, *Cohen's d* = effect size.

As shown in **Table 3.4**, statistical power was adjusted according to the predefined effect

size, with the required mean difference for an effect size of 0.8 (considered a large effect) being calculated accordingly. The primary focus of group comparisons was on Group 1 versus Group 2, where the mean of Group 1 was fixed at zero and the mean of Group 2 was adjusted according to the variance ratio. In contrast, false significance was assessed by comparing Group 1 with Group 3, both having means set to zero. This design enables simultaneous evaluation of power and false positive rates under various experimental conditions.

It is important to highlight that the power measurement design used in this study (i.e., Group 1 = Group 3 < Group 2) more accurately reflects real-world research scenarios, where not all group comparisons reveal significant differences. By simulating false significance between groups with no true mean differences, this approach reveals the susceptibility of multiple comparison methods to false discoveries, thus indicating their stability and robustness. In other words, this design provides a comprehensive framework for assessing parameter adjustments in various multiple comparison procedures, identifying which methods reliably maintain accuracy when the alternative hypothesis holds, while avoiding inflated error rates when the null hypothesis is true.

Regarding the multiple comparison tests (MCTs) and robust procedures examined in this study, both classical methods—Bonferroni's test, Dunnett's test, and Tukey's HSD—and robust procedures—including the Games-Howell test, the PI procedure, and the sandwich estimator family—were evaluated. Additionally, combinations of classical tests with robust estimators (e.g., Bonferroni's test + PI procedure, Dunnett's test + PI procedure) were also considered. Within the sandwich estimator family, HC0, HC1, HC2, and HC3 variants were

included in the simulation, whereas HC4, HC4m, and HC5 were excluded due to their specific design for handling extreme values and high leverage points.

In summary, this study compares the performance of all these 14 MCT methods based on 5000 replicated tests via Monte-Carlo Simulation to obtain as accurate and reliable result as we can within our computing power. The 14 MCT methods are: Bonferroni's test, Dunnett's test, Tukey's HSD, Games-Howell test, Dunnett's test + HC0, Dunnett's test + HC1, Dunnett's test + HC2, Dunnett's test + HC3, Dunnett's test + PI, Tukey's HSD + HC0, Tukey's HSD + HC1, Tukey's HSD + HC2, Tukey's HSD + HC3, Tukey's HSD + PI. Except for methods in the Dunnett's family, which produce 2 contrasts between the control group and two treatment groups, all other methods produced 3 pairwise contrasts: group 1 to group 2, group 1 to group 3, and group 2 to group 3. The evaluation metrics for all these contrasts include the *p* value, CI exclusion rate, confidence interval width, power, and effect size. In order to ensure the reproducibility of the research findings, the seed for data simulation is set as 76.

Chapter IV: Results

The results of the 5,000 Monte Carlo simulations are presented in this section, organized into four subsections: Type I error rate, confidence interval width, CI exclusion rate, and statistical power. The Type I error rate and CI exclusion rate were calculated based on the proportion of false positives observed across the 5,000 simulation iterations. Confidence interval width was computed as the average interval length over all simulations. Statistical power was determined by the proportion of correctly detected significant effects within the 5,000 replicates. Consequently, four distinct result sets were obtained, each comprising a 51 × 36 matrix representing combinations of manipulated conditions and MCT methods. In this chapter, the primary focus is on overarching patterns and general trends. Detailed comparisons and nuanced performance analyses of individual robust methods and their combinations are provided in **Chapter V**.

4.1 Type I error rate

The Type I error rates are presented below using faceted plots. Each plot displays comparisons across different variance ratios (VR), sample sizes, and multiple comparison methods. Separate plots were generated for the contrasts between Group 1 and Group 2 (G1–G2), Group 1 and Group 3 (G1–G3), and Group 2 and Group 3 (G2–G3), with additional stratification based on the manipulated experimental conditions.

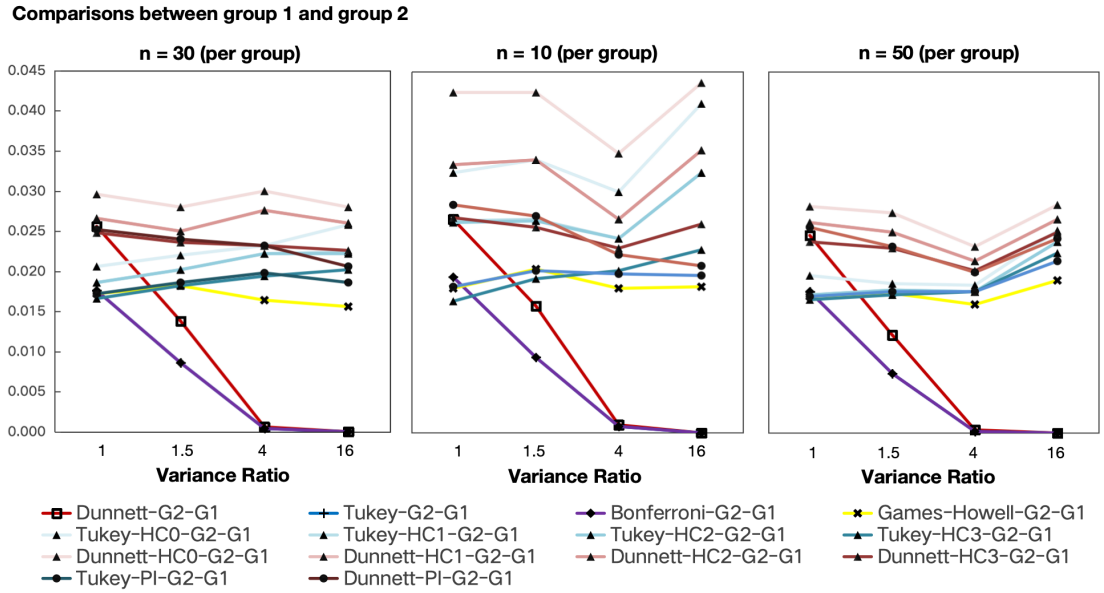
Results under balanced group conditions are illustrated from **Figures 4.1** to **4.3**. Since the group means were set to 0 in the simulation, i.e., the null hypothesis (H0) is satisfied and there are no significant differences between the means, a significant *p* value indicates a false

positive, which should ideally be controlled below 0.05. If it exceeds this threshold, it suggests an excessive false positive rate. We first went through each figure, then concluded the performance of 14 MCT methods combining all contrasts from **Figure 4.1** to **Figure 4.3**.

As shown in **Figure 4.1**, the false positive rates for all 14 methods comparing G1 and G2 contrasts remained below 0.045, which is within the acceptable Type I error threshold of 0.05. Classical MCT methods, including Dunnett's test, Bonferroni correction, and Tukey's HSD, exhibited a consistent decreasing trend in false positive rates as the variance ratio increased across all sample size conditions. In contrast, robust procedures displayed a more variable pattern, with fluctuations observed as VR increased. Notably, when the sample size was extremely small (n = 10), all MCT methods showed the poorest performance, with Type I error rates ranging from 0 to 0.045. However, when the sample size increased, the performance of MCT methods turned more conservative, with error rates falling within the narrower range of 0 to 0.030.

Figure 4.1

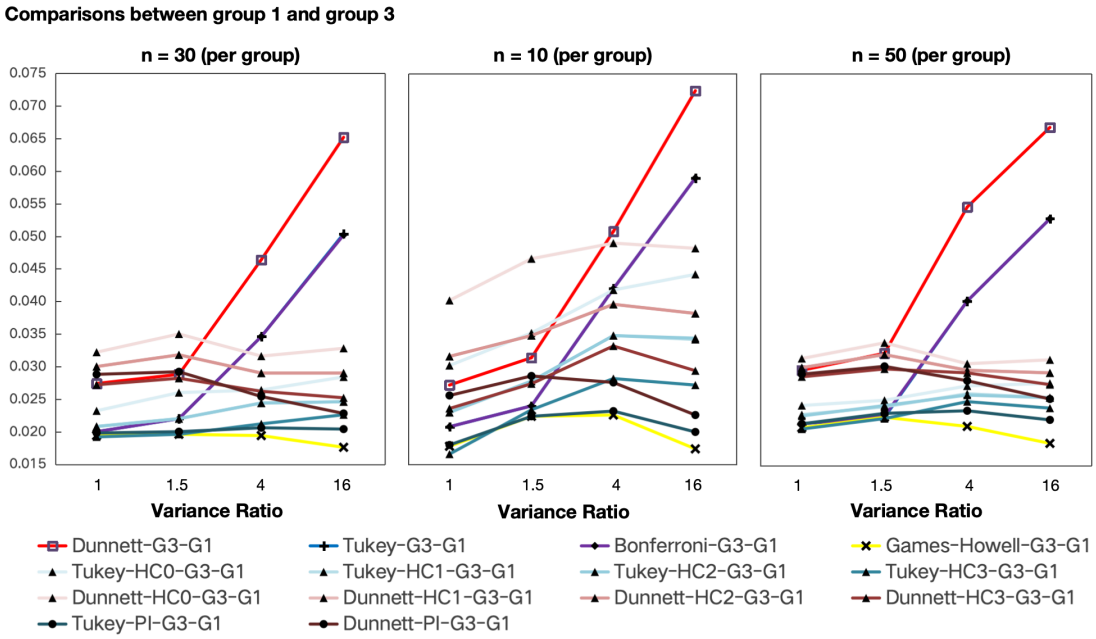
The FWER Results of Different MCT Methods under Different VRs and Sample Sizes Between G1 -G2 Contrasts (Balanced Group)



As shown **Figure 4.2**, the false positive rates for classical and robust methods exhibited distinct patterns in the G1–G3 contrasts. Specifically, the classical methods showed a marked increase in false positive rates as the variance ratio (VR) rose from 1 to 16, surpassing the nominal 0.05 significance threshold, with rates reaching as high as 0.075. In contrast, all robust MCT methods effectively controlled the false positive rates, maintaining values below the 0.05 threshold, thus adhering to the acceptable limit for Type I error. Notably, when the sample size was extremely small (n = 10), the performance of robust methods demonstrated greater variability and fluctuation compared to scenarios with medium (n = 30) and large (n = 50) sample sizes, where the error rates were more stable.

Figure 4.2

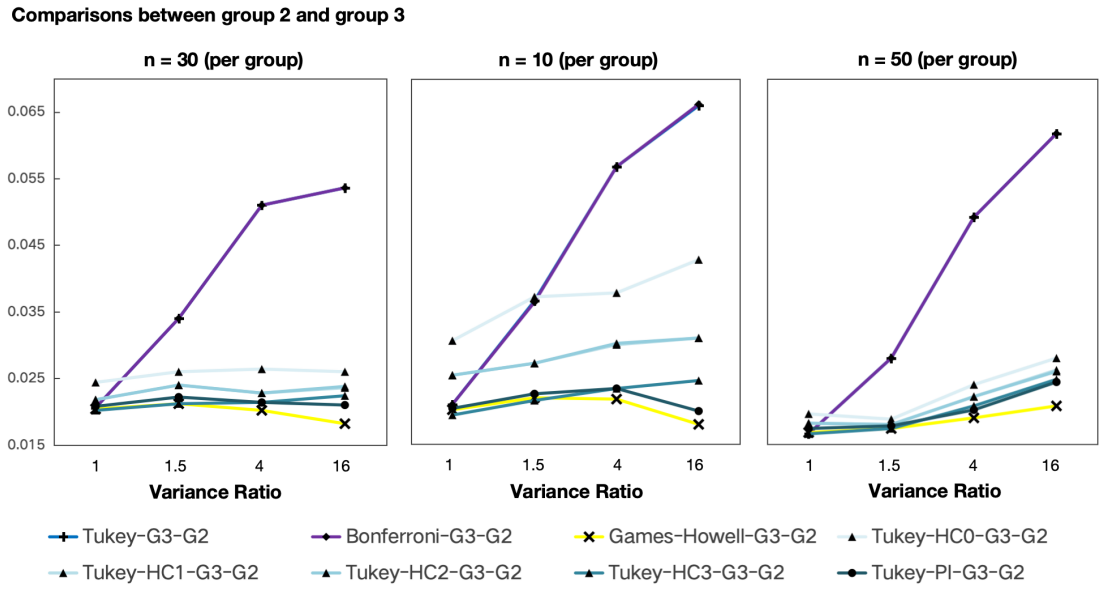
The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G3 Contrasts (Balanced Group)



As shown **Figure 4.3**, the familywise error rate (FWER) results for the G2–G3 contrasts revealed distinct patterns between classical and robust methods. For the classical Tukey and Bonferroni tests, false positive rates increased notably as the variance ratio (VR) escalated from 1 to 16, surpassing the nominal 0.05 threshold, albeit with a lower peak around 0.065. Conversely, all robust MCT methods maintained strong control over false positive rates, keeping them below the 0.05 significance level, thereby meeting the acceptable criteria for Type I error control. When sample sizes were extremely small (n = 10), the robust methods exhibited greater variability in performance compared to the more stable results observed under moderate (n = 30) and large (n = 50) sample sizes. Under these latter conditions, FWER values remained consistently between 0.015 and 0.030.

Figure 4.3

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2 – G3 Contrasts (Balanced Group)



In summary, all 14 methods maintained good robustness across the G1–G2 contrast, keeping the false positive rate below 0.05. However, it was worth noting that for the classical MCT methods, this relatively ideal performance was not maintained in the G1–G3 and G2–G3 contrasts. As the VR increased, the false positive rates of the three classical MCT methods surged, approaching or exceeding the acceptable threshold of 0.05 when VR equaled 4, with the highest value even surpassing 0.07. Increasing the sample size did not mitigate the impact of variance heteroscedasticity on these classical methods.

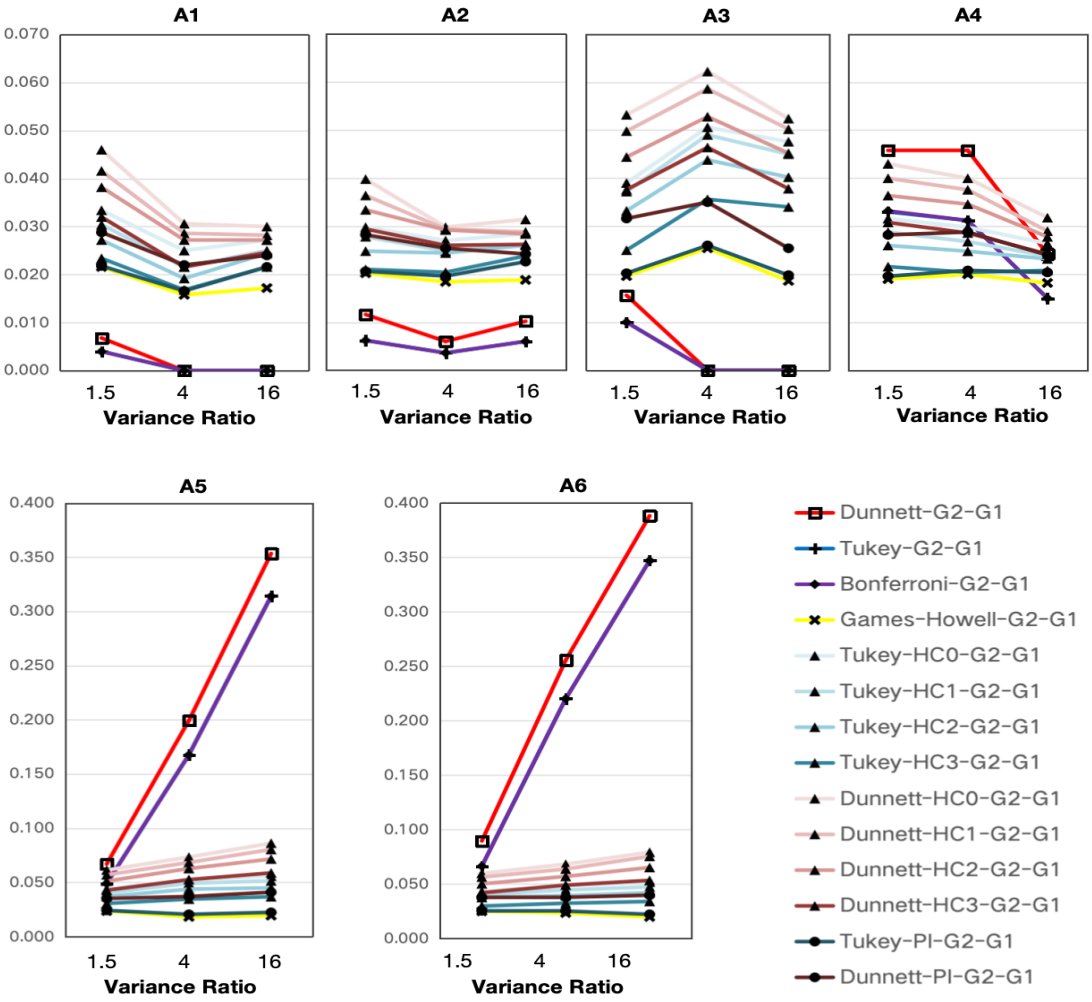
Across the three contrasts, most methods (such as the HC family, PI procedure, and Games-Howell) were not strongly affected by heteroscedasticity and automatically adjusted their false positive probabilities at VR = 4 or VR = 16. Among these methods, the Games-Howell method exhibited the smallest fluctuations, indicating greater sensitivity to variance imbalance. For both the HC estimator and the PI procedure, the combination of HC-series/PI

with Tukey consistently demonstrated better robustness than the combination of HC-series with Dunnett. As for the performance within the HC family, HC3 performed the best, with a relatively robust Type I error rate, especially when combined with Tukey. The results under unbalanced group conditions (3 different sample sizes) are illustrated from **Figure 4.4** to **Figure 4.6**. From A1 to A6 condition, the sample sizes of the three group were: G1: n=10, G2: n=30, G3: n=50. Among these conditions, the combinations of group sample size and variance were different, as shown in **Chapter III**. We first went through each contrast figure before comprehensively comparing performance across contrasts.

As shown **Figure 4.4**, the FWER results of classical methods (Tukey, Dunnett, and Bonferroni) and robust methods exhibited distinct patterns in the G1–G2 contrasts, particularly under conditions A5 and A6. While Dunnett, Tukey, and Bonferroni tests maintained acceptable false positive rates under conditions A1 through A4, their rates became unacceptably high (reaching up to 0.400) under A5 and A6, where small groups were paired with large SD. In contrast, the robust MCT methods demonstrated more stable performance, with false positive rates generally controlled below the 0.05 threshold across varying variance ratios. However, some Dunnett-based robust procedures under conditions A3, A5, and A6 slightly exceeded this level regardless of the VR settings. Overall, increasing the VR did not lead to inflation of false positive rates when applying robust procedures.

Figure 4.4

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G2 Contrasts (Unbalanced Group & 3 Different Sizes)

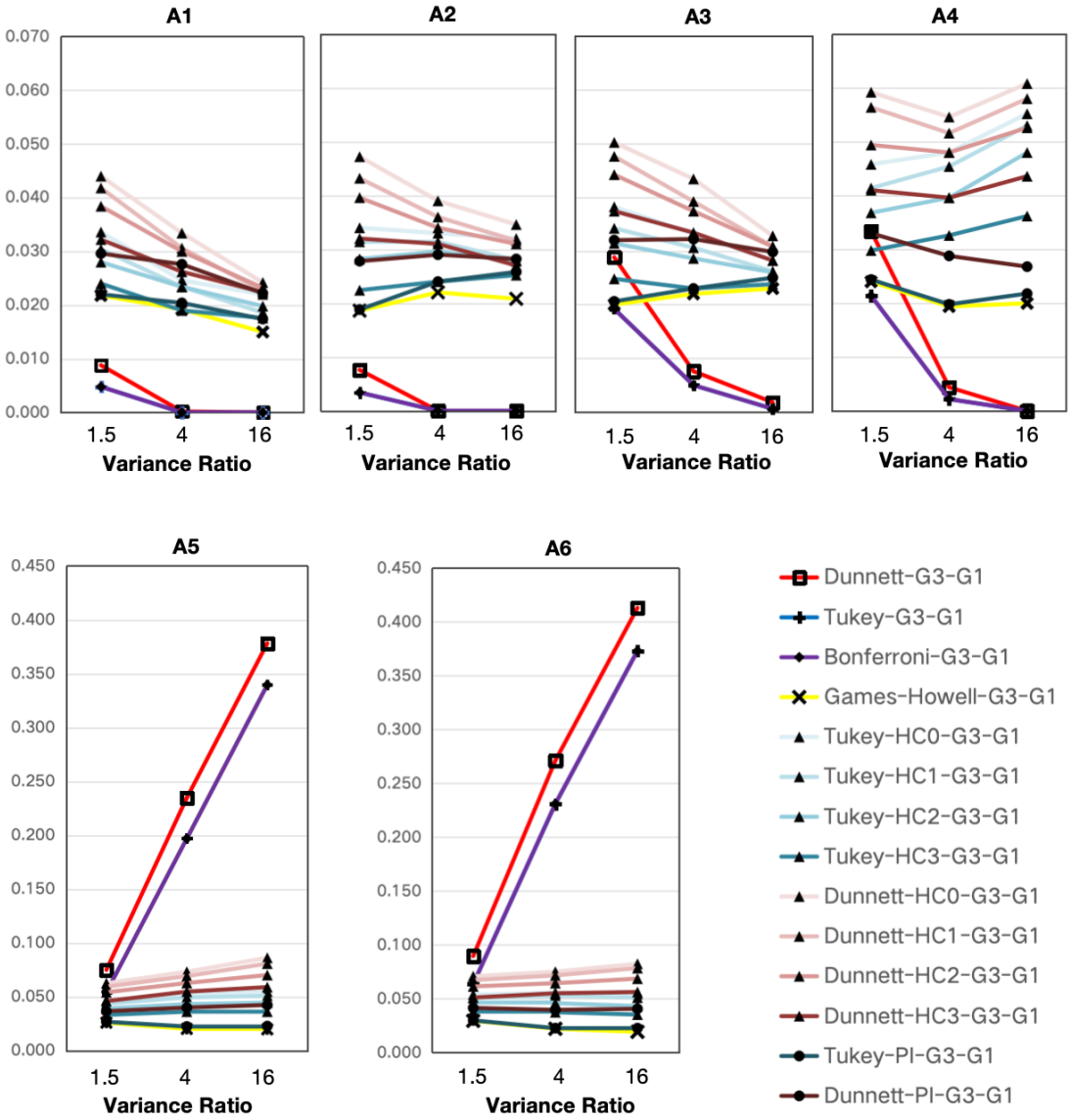


As shown in **Figure 4.5**, the FWER results of classical methods (Tukey, Dunnett, and Bonferroni) and robust methods displayed different patterns in the G1–G3 contrasts. For Dunnett, Tukey, and Bonferroni tests, false positive rates were very low under conditions A1 to A4, outperforming all robust procedures, but became unacceptably high (exceeding 0.400) under A5 and A6, where size-variance pairing was negative. Robust MCT methods maintained false positive rates below 0.050 under A1 to A3, showing a decreasing trend as VR changed. However, Dunnett-related methods exceeded 0.05 under A4, A5, and A6

regardless of VR.

Figure 4.5

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G3 Contrasts (Unbalanced Group & 3 Different Sizes)

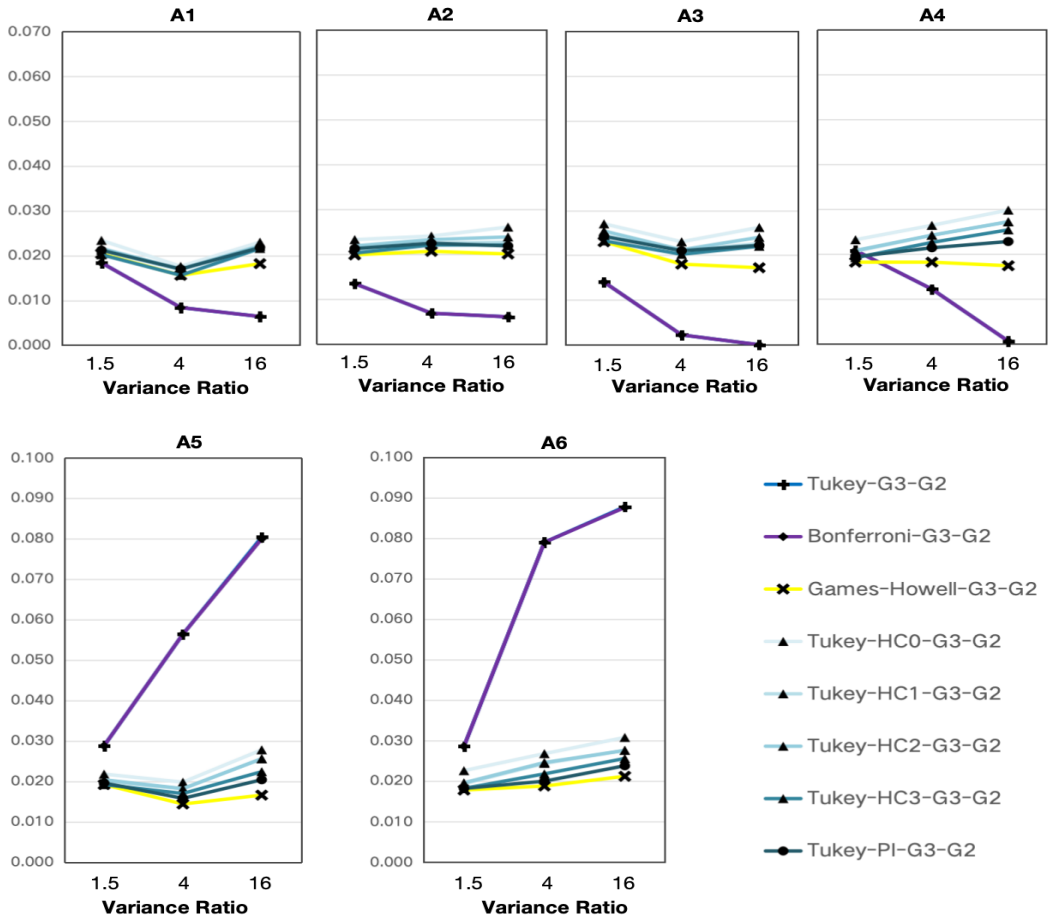


As shown **Figure 4.6**, the FWER results of two classical methods (Tukey and Bonferroni) and robust methods exhibited different patterns for the G2–G3 contrasts. For Tukey and Bonferroni tests, false positive rates were relatively low under conditions A1 to A4 compared to robust procedures but became unacceptably high—reaching 0.080 and 0.090—

as VR increased under A5 and A6, where the size-variance pairing was negative. In contrast, all robust MCT methods maintained controllable false positive rates below 0.030 across conditions A1 to A6, demonstrating greater stability.

Figure 4.6

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2 -G3 Contrasts (Unbalanced Group & 3 Different Sizes)



In summary, the three classical MCT methods tended to be much more conservative across A1 to A4 condition and shifted to liberal under A5 and A6 conditions, showing a poor performance when facing various situations of sample-size-variance pairing. As to the robust procedures, the Dunnett-HC family showed an increase of approximately 0.02–0.03. Since these combinations already exceeded 0.05 and became liberal when VR = 1.5 (except for

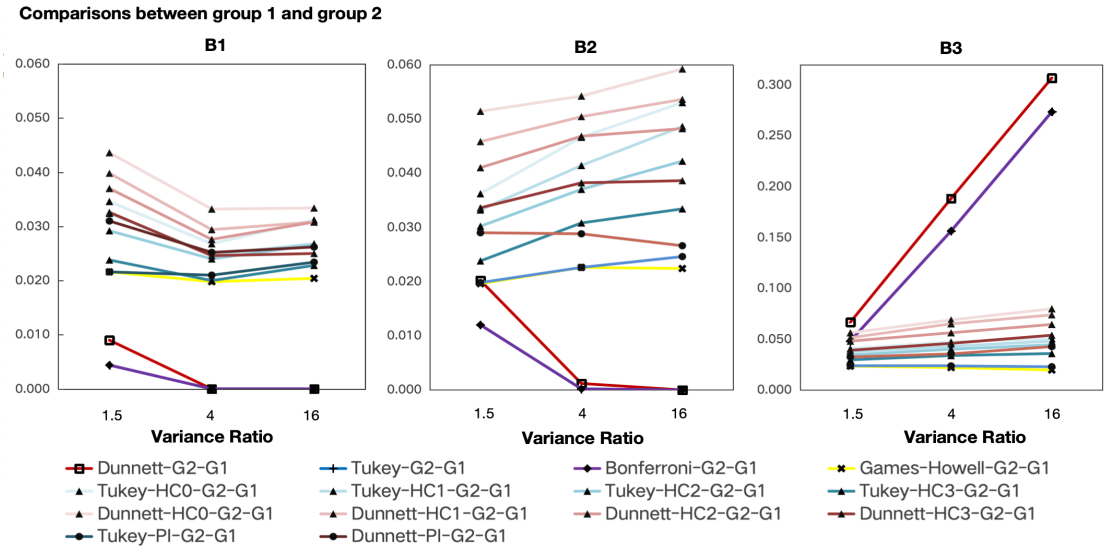
Dunnett-HC3), they exhibited higher FWER. In contrast, Tukey-HC, Tukey-PI, and Dunnett-PI were robust and controlled the increase within a smaller range, demonstrating stronger control, while the Games-Howell method remained consistently robust and stable around 0.025.

Results under unbalanced group conditions (two balanced groups with one extremely larger or smaller group) are shown in **Figures 4.7** to **4.12**. Conditions B1 to B3 had group sizes of G1: n=10, G2: n=30, G3: n=30, with G1 being the smallest group. Conditions C1 to C3 had group sizes of G1: n=50, G2: n=30, G3: n=30, with G1 being the largest group. All combinations of group sample size and variance were examined within these conditions. We first went through each contrast figure, then analyze the performance of all MCT methods across contrasts.

As shown **Figure 4.7**, the FWER results of classical methods (Tukey, Dunnett, and Bonferroni) and robust methods exhibited different patterns for the G1–G2 contrasts. For Dunnett, Tukey, and Bonferroni tests, false positive rates were very low under conditions B1 and B2, outperforming all robust procedures, but rose to unacceptable levels—over 0.200—under B3, where the smallest group had the largest SD. Robust MCT methods maintained false positive rates below 0.050 under B1, with a decreasing trend as VR increased. However, some Dunnett-related methods exceeded 0.05 under B2, and all did so under B3 as VR increased.

Figure 4.7

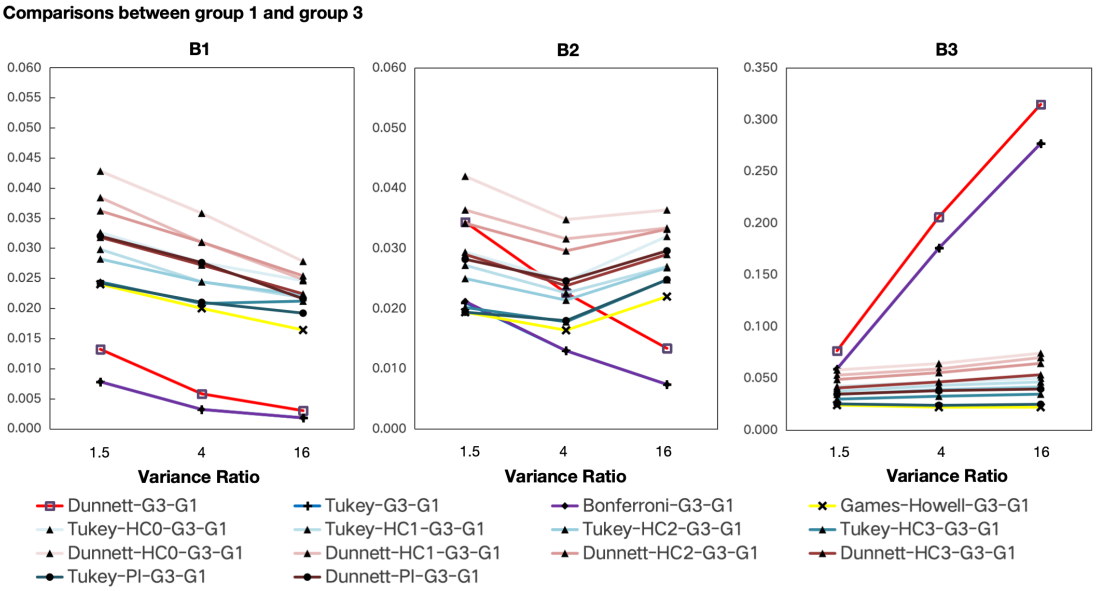
The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G2 (μ=0) Contrasts (Unbalanced Group & 1 Smaller Group: n_{G1}=10, n_{G2}=n_{G3}=30)



As shown **Figure 4.8**, the FWER results of classical methods (Tukey, Dunnett, and Bonferroni) and robust methods showed different patterns for the G1–G2 contrasts. For Dunnett, Tukey, and Bonferroni tests, false positive rates were very low under conditions B1 and B2, outperforming all robust procedures, but rose to unacceptable levels—over 0.200—under B3, where the smallest group had the largest SD. Robust MCT methods maintained false positive rates below 0.050 under B1, with a decreasing trend as VR increased. However, some Dunnett-related methods exceeded 0.05 under B2, and all did so under B3 as VR increased.

Figure 4.8

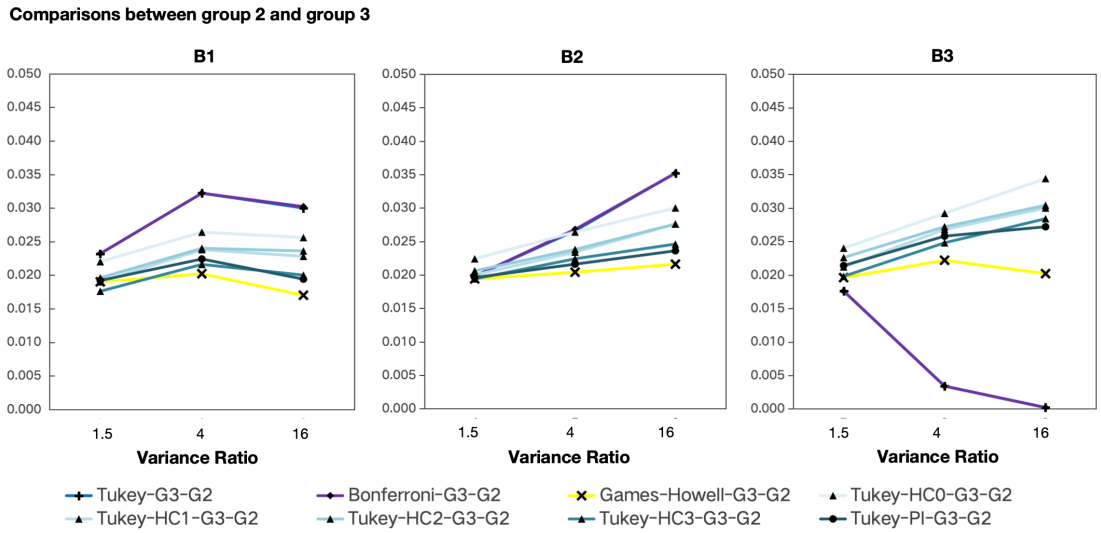
The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G3 (Unbalanced Group & 1 Smaller Group: n_{G1}=10, n_{G2}=n_{G3}=30)



As shown **Figure 4.9**, the FWER results of classical methods (Tukey and Bonferroni) and robust methods showed similar patterns for the G2–G3 contrasts under conditions B1 and B2, but they differed under condition B3. For Tukey and Bonferroni tests, false positive rates ranged from 0.015 to 0.035 under B1 and B2, comparable to robust procedures and within acceptable FWER limits, but dropped sharply as VR increased under B3. Robust MCT methods maintained controllable false positive rates below 0.050 across B1, B2, and B3, with a slight increase as VR grew.

Figure 4.9

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2 -G3 (Unbalanced Group & 1 Smaller Group: n_{G1}=10, n_{G2}=n_{G3}=30)

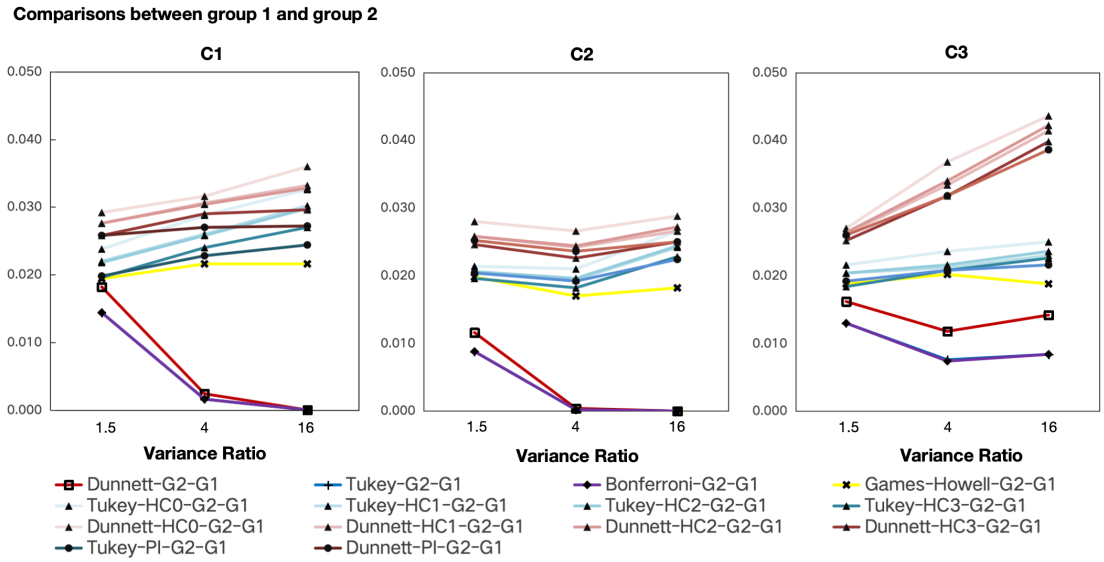


The following figures illustrate performance in a more realistic scenario where two groups are equal in size and one group is larger. All groups meet or exceed the minimum sample size standard (n=30). This allows us to examine whether heteroscedasticity still affects the results in such conditions.

As we can see from **Figure 4.10**, the false positive rates of all 14 methods for the G1–G2 contrasts remained below 0.050, within the acceptable range. Classical MCT methods (Dunnett, Bonferroni, and Tukey) showed a decreasing trend as VR increased across all sample size conditions, while robust procedures exhibited stable or slightly increasing patterns. Dunnett-related methods (red lines) and Tukey-related methods (blue lines) followed similar patterns under C1 and C2 conditions, but diverged under C3, where Dunnett-related methods increased with VR, unlike the stable trend seen in Tukey-related methods.

Figure 4.10

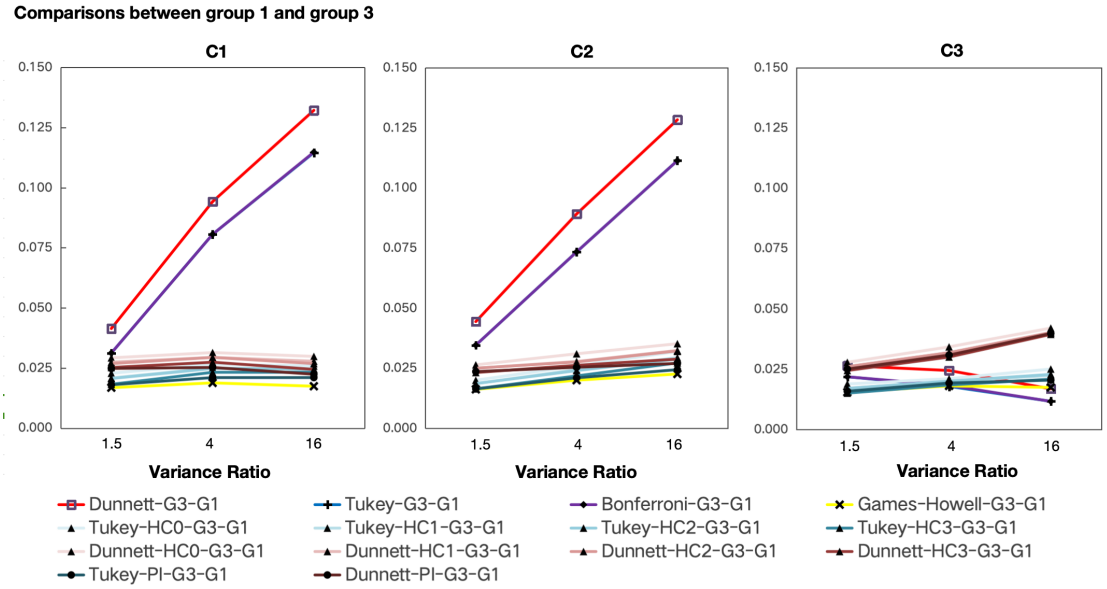
The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G2 (μ=0) Contrasts (Unbalanced Group & 1 Larger Group: n_{G1}=50, n_{G2}=n_{G3}=30)



As shown **Figure 4.11**, the FWER results of classical methods (Tukey, Dunnett, and Bonferroni) and robust methods showed different patterns between the G1–G3 contrasts under C1 and C2 conditions, and a similar pattern under the C3 condition. Specifically, for classical MCT tests, the false positive rates increased significantly under conditions C1 and C2, reaching a peak of over 0.130, which is clearly unacceptable. However, the false positive rates of robust MCT methods under these two conditions remained between 0.015 and 0.025, showing a consistent and reliable trend as VR changed.

Figure 4.11

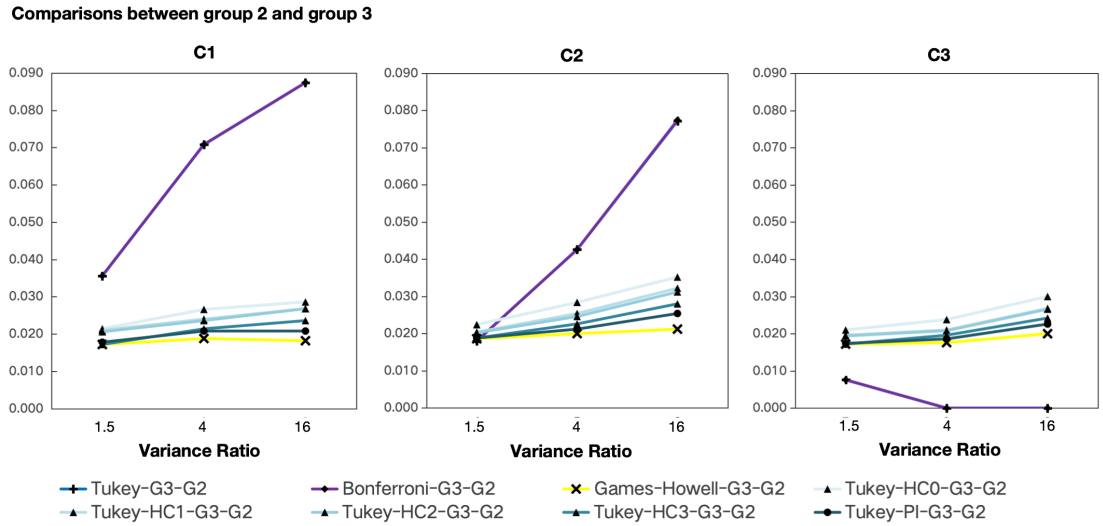
The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G1 -G3 Contrasts (Unbalanced Group & 1 Larger Group: $n_{G1}=50$, $n_{G2}=n_{G3}=30$)



As shown **Figure 4.12**, the FWER results of classical methods (Tukey, Dunnett, and Bonferroni) and robust methods showed different patterns between the G1–G3 contrasts. For the two classical MCT tests, Tukey and Bonferroni, the false positive rates increased significantly under conditions C1 and C2, reaching an unacceptable peak of about 0.090, while the rates decreased to 0 under condition C3. Meanwhile, the false positive rates of robust MCT methods remained stable between 0.015 and 0.030 across all three conditions, regardless of how VR changed.

Figure 4.12

The FWER Results of Different MCT Methods under Different VRs and Sample Sizes between G2 -G3 Contrasts (Unbalanced Group & 1 Larger Group: $n_{G1}=50$, $n_{G2}=n_{G3}=30$)



To summarize, when faced with unbalanced situations involving a small or large group, the classical methods became extremely liberal, with a much higher Type I error rate (the highest reaching up to 0.315), or extremely conservative down to 0, which was unacceptable for significance testing. However, most of the other methods remained robust, with good false positive rate control, consistently keeping the rate between 0.015 and 0.035, except for the Dunnett-HC series. The Games-Howell method exhibited the strongest conservative pattern, while the PI series and the Tukey-HC series were relatively robust.

4.2 CI Exclusion Rate

This section evaluates the occurrence of false positives in 5000 simulations based on whether the confidence interval (CI) includes the true mean (0) and compares it to the false positive rate based on *p* values from Chapter 4.1. For the three groups with the same mean of 0, all mean differences are supposed to be 0 for 3 contrasts. If the CI of mean difference

contains 0, it indicates that the H0 is successfully accepted. Otherwise, if the CI excludes 0, it indicates that the inferred interval is biased, leading to a false positive conclusion. When analyzing the CI exclusion results, we use the *p* value results from Chapter 4.1 as a benchmark and observe the main differences.

The main difference is shown below in bar charts. A positive bar indicates that the CI exclusion rate yielded more false positive results; a negative bar indicates that the *p* value yielded more false positives; and a bar equal to zero means both calculation methods produced the same results. The results are presented in **Figures 4.13 to 4.27.** Specifically, **Figures 4.13 to 4.15** show differences under three balanced conditions (n = 30, 10, 50). **Figures 4.16 to 4.21** show differences under unbalanced conditions with groups of three different sizes (A1–A6). **Figures 4.22 to 4.24** show differences under unbalanced conditions with one smaller group (B1–B3). **Figures 4.25 to 4.27** show differences under unbalanced conditions with one larger group (C1–C3). We went through the four experimental conditions, followed by a detailed evaluation of the CI exclusion performance of the 14 multiple comparison test (MCT) methods across conditions.

First, as seen in **Figures 4.13 to 4.15**, Dunnett's test and robust Dunnett-related tests exhibited bars of zero length, indicating these methods yielded identical results for both *p value* and CI exclusion rate calculations. In contrast, a cluster of bars was observed for the Games-Howell test, indicating a gap between the two calculation methods. The differences in frequencies ranged from –0.0002 to 0.0008, with the highest bar corresponding to the Games-Howell test, while other differences were small and around 0.0002.

Figure 4.13

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value When n=30

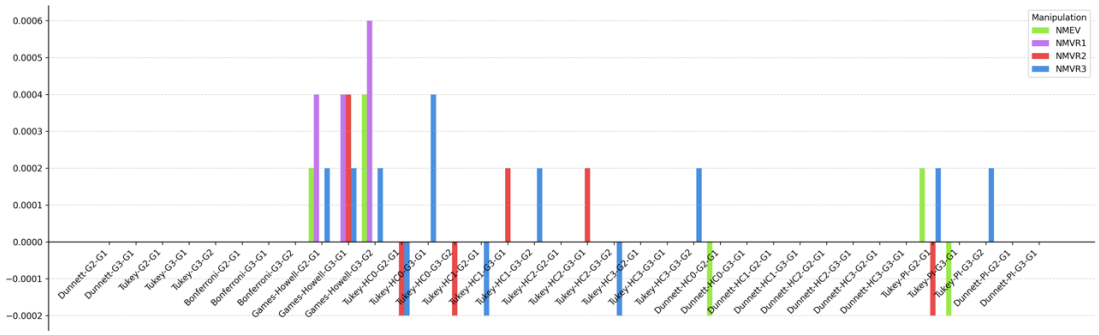


Figure 4.14

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value When n=10

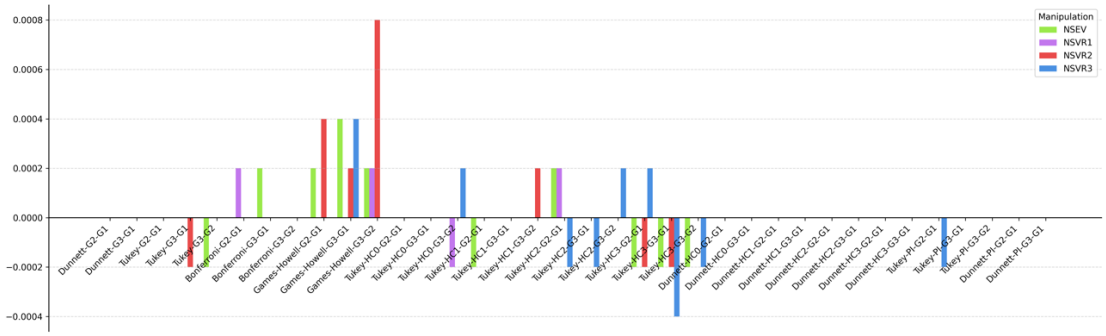
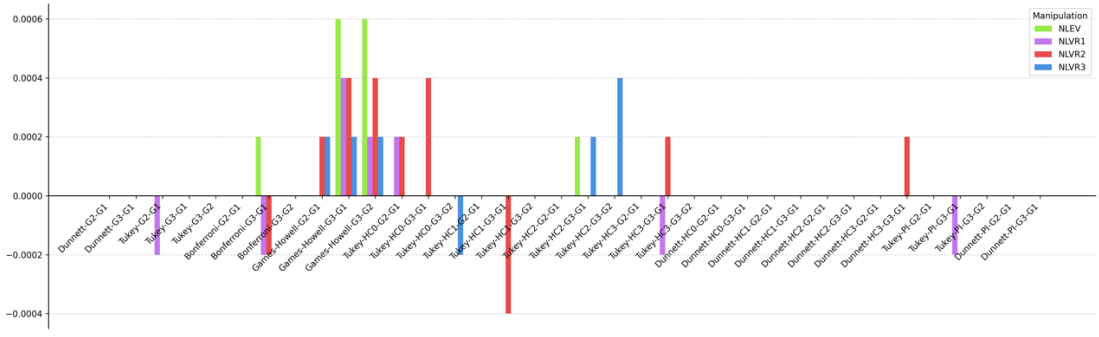


Figure 4.15

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value When n=50



Secondly, as we can see from **Figure 4.16** to **Figure 4.21**, Dunnett's test and robust Dunnett-related tests showed zero-length bars under conditions A1 and A4, indicating these methods captured the same false positive rates from both *p* value and CI exclusion rate calculations. A cluster of bars was also observed for the Games-Howell test, with the highest bars of 0.0008 under condition A1 and 0.0010 under condition A3. Apart from the Games-Howell test, the differences in frequencies ranged from –0.0004 to 0.0004. Notably, under condition A6 (smallest group with largest SD vs. largest group with smallest SD), larger gaps appeared between the false positive frequencies derived from CI exclusion rates and *p* values.

Figure 4.16

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under A1

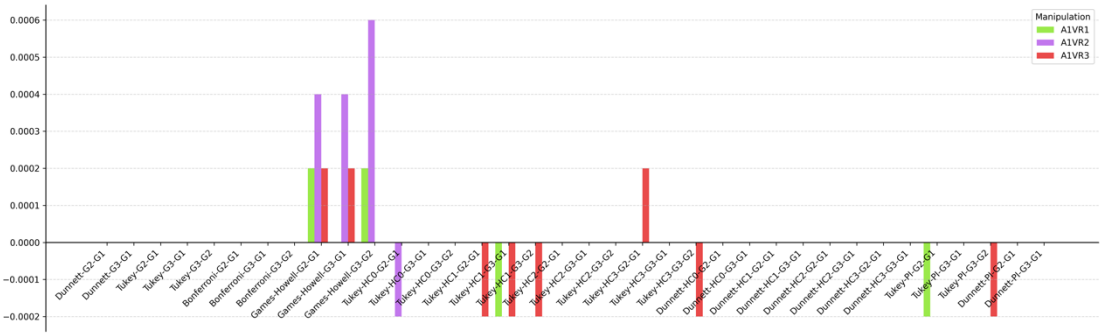


Figure 4.17

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under A2

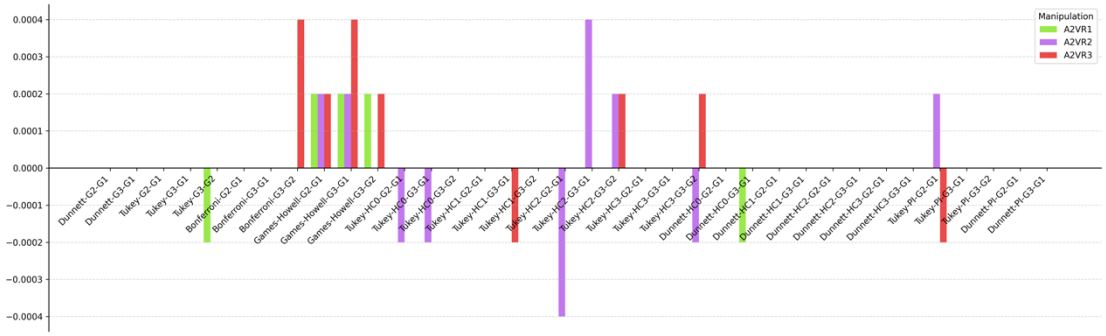


Figure 4.18

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under A3

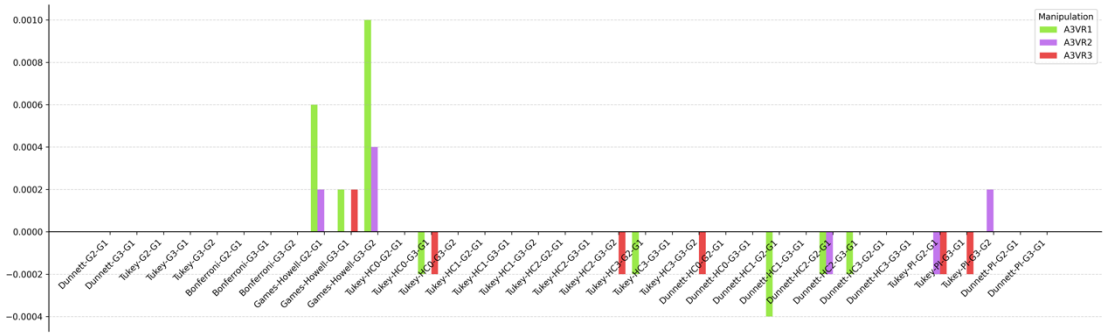


Figure 4.19

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under A4

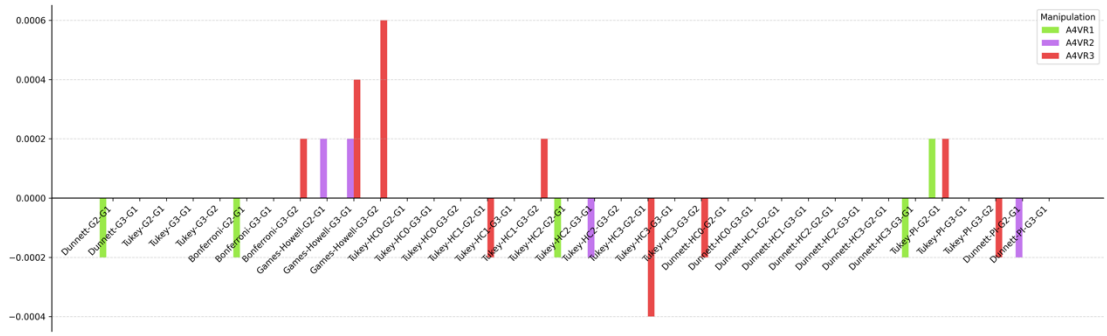


Figure 4.20

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under A5

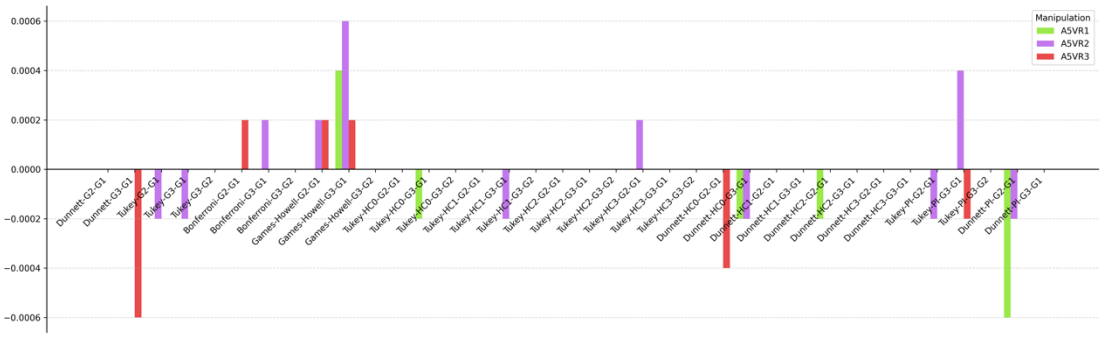
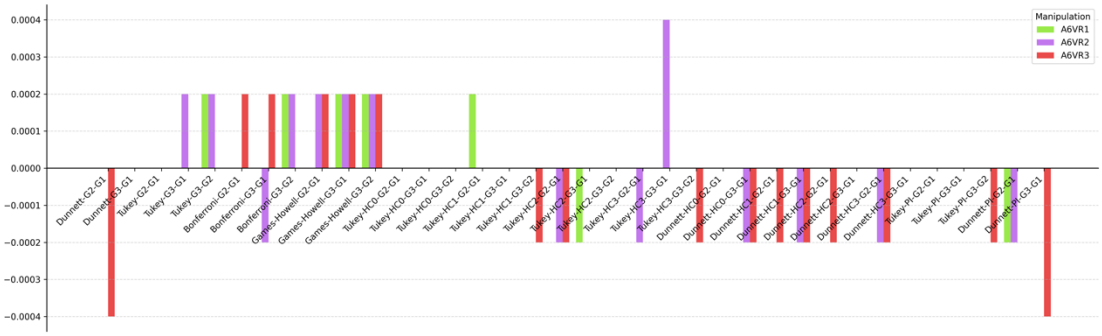


Figure 4.21

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under A6



As we can see from **Figure 4.22** to **Figure 4.24**, Dunnett's test displayed a zero-length bar, indicating that this method yielded the same results from both *p* value and CI exclusion rate calculations. A cluster of bars was observed for the Games-Howell test, reflecting the gap between the two calculation methods. The differences in frequencies ranged from –0.0004 to 0.0006, with the highest bar of 0.0006 belonging to the Games-Howell test, while other differences were smaller, around 0.0004.

Figure 4.22

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under B1

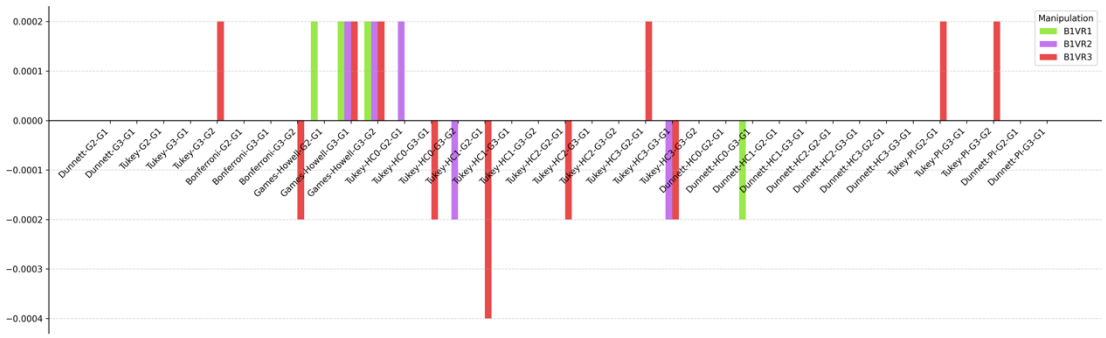


Figure 4.23

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under B2

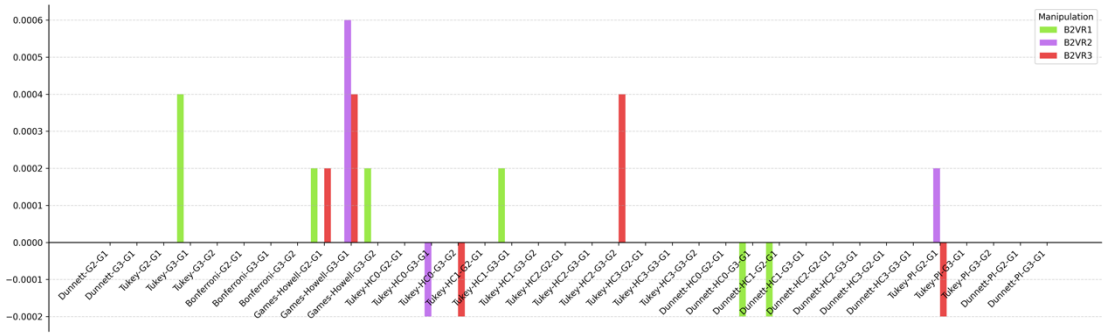
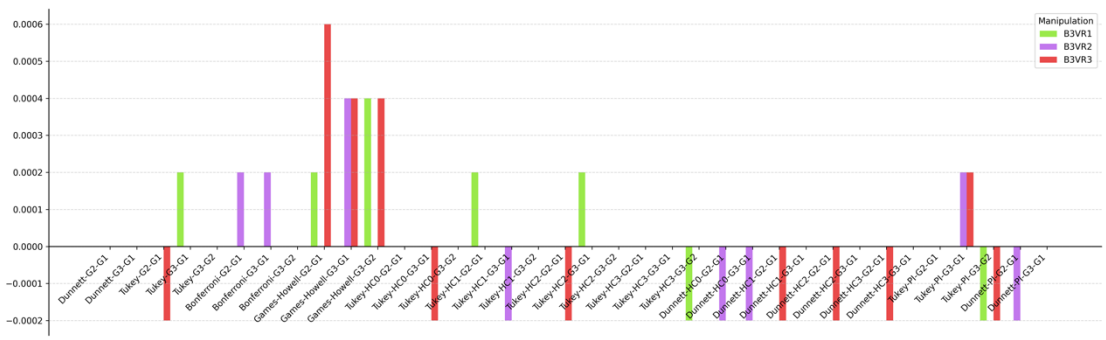


Figure 4.24

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under B3



As we can see from **Figure 4.25** to **Figure 4.27**, Dunnett's test displayed a zero-length bar, indicating that it produced the same results from both *p* value and CI exclusion rate calculations. Clusters of bars were observed for the Games-Howell test (across C1 to C3) and Tukey-HC1 (under C2), with a peak difference of 0.0008. Aside from these two clusters, the differences in frequencies ranged from –0.0002 to 0.0004.

Figure 4.25

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under C1

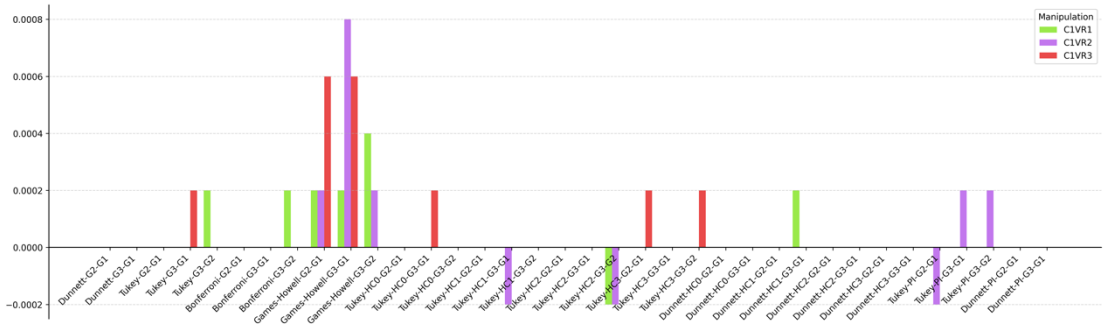


Figure 4.26

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under C2

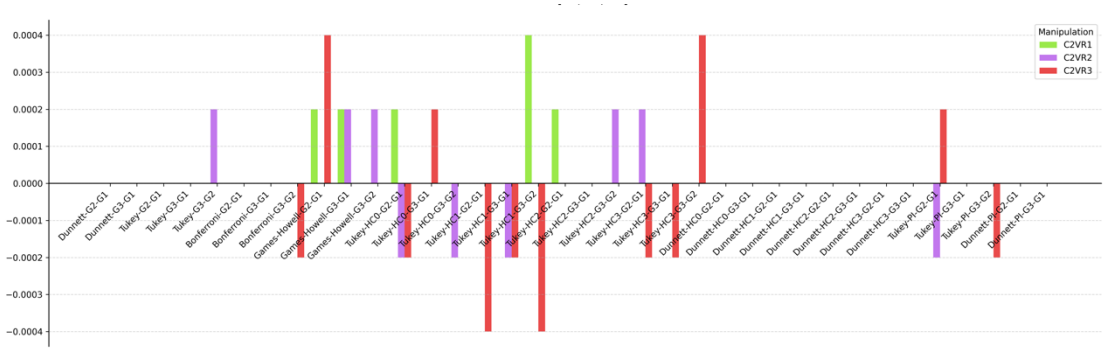
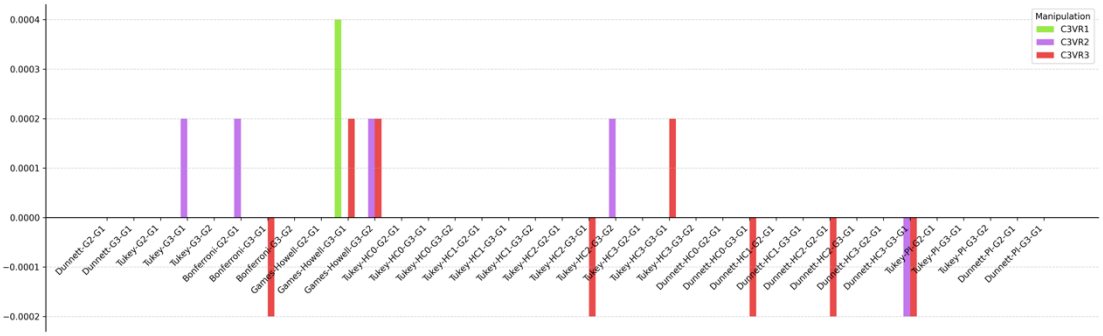


Figure 4.27

The Differences of Frequencies of Inconsistent Decisions between CI exclusion rate and P Value under C3



When interpreting statistical significance, it was also important to consider the stability of the confidence interval for a comprehensive judgment. Overall, the two evaluation standards aligned closely but showed some deviation under specific methods and conditions. Under all conditions, the Games-Howell method showed the largest deviation from the p-value results. In a total of 36 inter-group comparisons under balanced situations, 75.0% exhibited under-adjustment. The misalignment rate was 70.4% under A1 to A6 conditions, and 64.9% under B1 to C3 conditions.

As for other robust procedures, misalignment was less frequent than that of the Games-Howell method. Under the balanced conditions, the Tukey-HC series showed more frequent misalignment than the Dunnett-HC series (25.7% vs. 2.1%), while the Tukey-PI series exhibited a similar pattern compared to the Dunnett-PI series (19.4% vs. 0.0%). Under unbalanced conditions, for the Tukey-HC series methods, the difference between CI and p-value significance levels was widespread (misalignment rate: 17.6%), which was larger than that of the Dunnett-HC series (12.5%). The Tukey-PI series also showed more frequent CI adjustment than the Dunnett-PI series (misalignment rate: 27.8% vs. 16.7%).

Among the three classical methods, Dunnett's method had the smallest difference in false positive rates between the CI and p-value (misalignment rate: 8.3%). Tukey showed a higher proportion of misalignment (rate: 11.1%), but the differences were all around 0.0002, which was very minor. The misalignment of Bonferroni's method appeared greater (misalignment rate: 18.5%), but the differences were also generally around 0.0002.

4.3 Confidence interval width

Considering that the confidence interval width varied as VR increased under different manipulated conditions, the relative interval length was calculated by dividing the absolute interval length by the overall average interval length for the corresponding manipulation condition. Thus, if the relative confidence width is greater than 1, it means the confidence width is larger than average; if it is less than 1, the confidence width is shorter than average. Several heat maps are used to illustrate the width results. Relative widths larger than average are colored in red, while smaller widths are colored in blue. The results are shown in **Figures 4.28 to 4.31**.

Figure 4.28 presents the results under all balanced-group conditions. It is clear that as VR increased across all sample size types, the blue and red colors intensified, indicating that deviations from the average grew larger. The confidence widths of the three classical methods remained stable and close to the average, while the Games-Howell test showed the largest adjustments in magnitude.

Figure 4.28

The Relative CI Width under Balanced-Group Conditions (n=30, 10, 50)

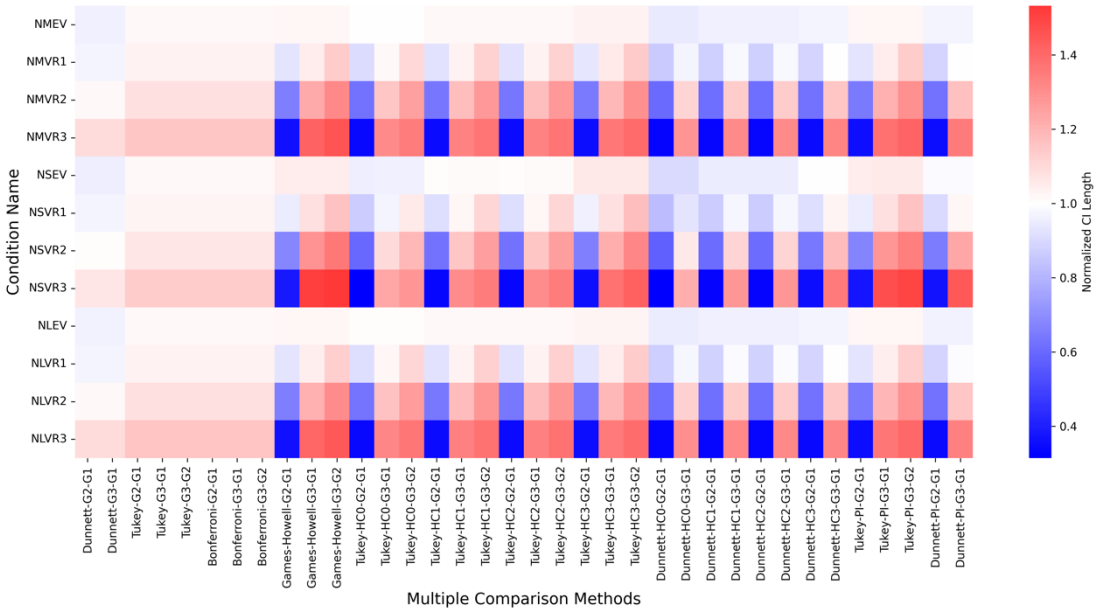


Figure 4.29 shows the result under A1-A6 conditions. As seen in the heatmap, the widths under A1 to A4 shared a similar pattern, while those under A5 and A6 exhibited a different pattern. Similar to the previous figure, as VR increased, the blue and red colors intensified, indicating that deviations from the average became larger.

Under conditions A1 to A4, the classical methods adjusted the confidence intervals by enlarging the widths to capture more variance, whereas the robust procedures slightly enlarged or shrunk the intervals for the contrasts. Under conditions A5 and A6, the classical methods adjusted the confidence intervals by shrinking the widths to obtain relatively accurate ranges. Dunnett-related robust procedures enlarged the intervals to address heteroscedasticity, while the Games-Howell test and Tukey-related robust procedures adjusted the G1-G2 and G1-G3 contrasts differently.

Figure 4.29

The Relative CI Width under Unbalanced-Group Conditions (A1-A6 Condition)

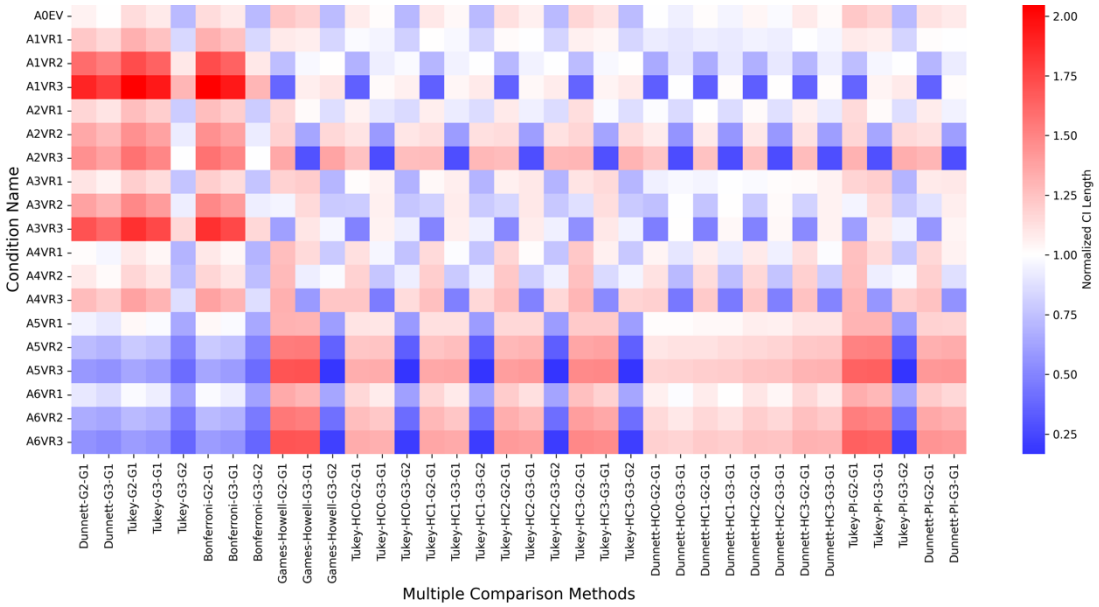


Figure 4.30 shows the result under B1-B3 conditions. As shown in the heatmap, the widths under B1 and B2 conditions shared a similar pattern, while those under the B3 condition showed a different pattern. Under B1 and B2, the classical methods adjusted the confidence intervals by largely enlarging the widths, whereas the robust procedures either slightly enlarged or sharply shrunk the intervals for the contrasts. Under B3, the classical methods adjusted the confidence intervals by shrinking the widths; Dunnett-related robust procedures enlarged the intervals to address heteroscedasticity, while the Games-Howell test and Tukey-related robust procedures widened the intervals for the G1-G2 and G1-G3 contrasts but narrowed them for the G2-G3 contrasts.

Figure 4.30

The Relative CI Width under Unbalanced-Group Conditions (B1-B3 Condition)

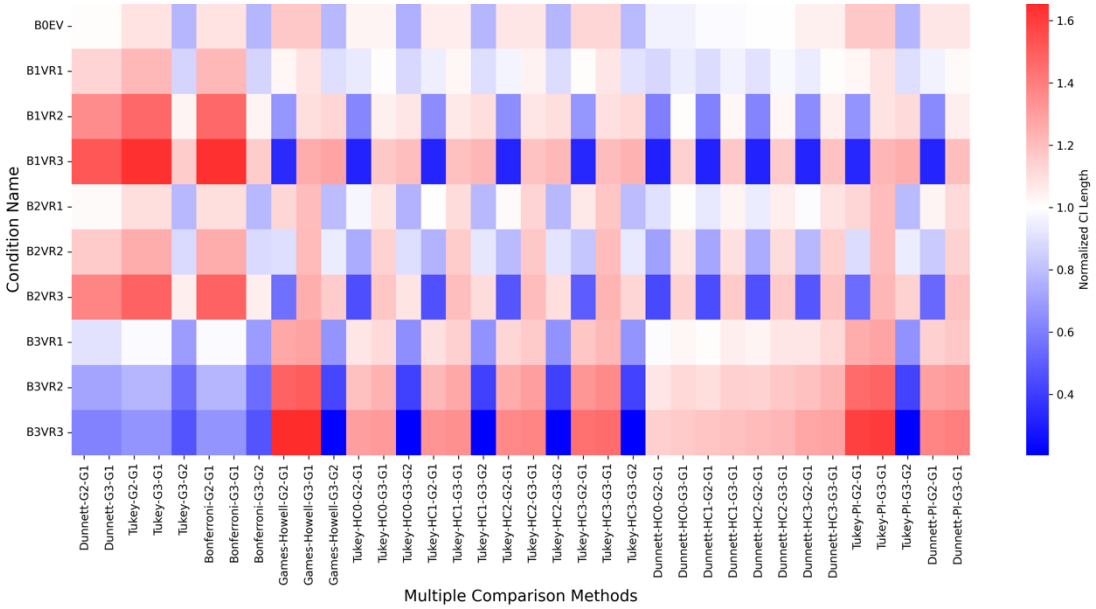
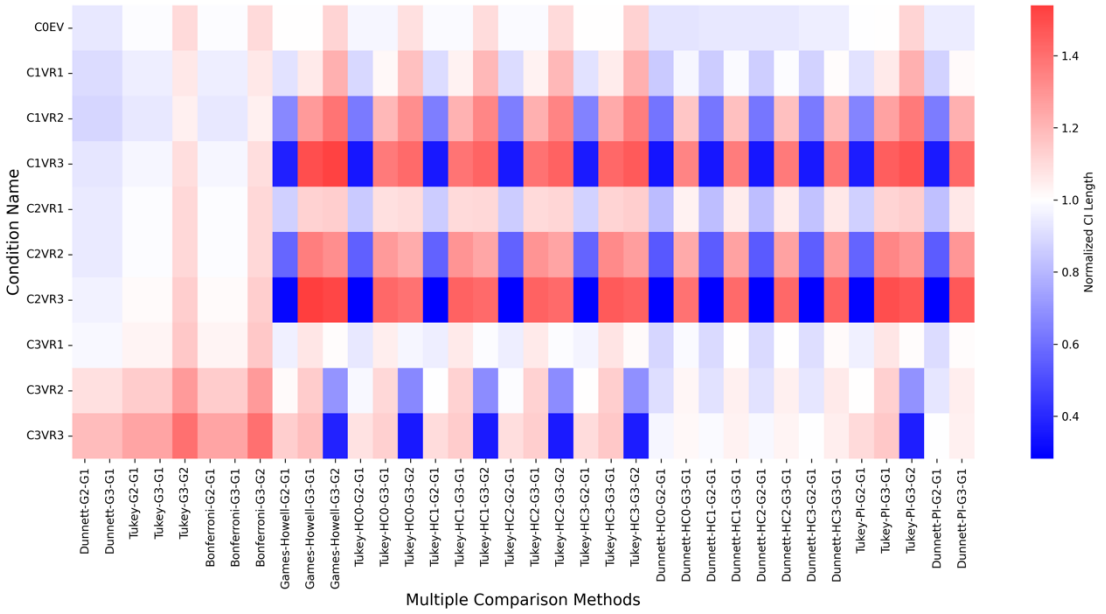


Figure 4.31 shows the result under C1-C3 conditions. For the three classical MCT methods, the adjustments of confidence intervals showed a similar and stable pattern, remaining close to the average width. For the robust procedures, the widths under C1 and C2 conditions shared a similar pattern, while the widths under the C3 condition showed a different pattern. Under C1 and C2, all robust procedures narrowed the intervals for the G1-G2 contrasts and widened the intervals for the G1-G3 and G2-G3 contrasts. Under the C3 condition, Dunnett-related robust procedures maintained widths close to the average across contrasts, whereas the Games-Howell test and Tukey-related robust procedures slightly widened the intervals for the G1-G2 and G1-G3 contrasts but greatly narrowed the intervals for the G2-G3 contrasts.

Figure 4.31

The Relative CI Width under Unbalanced-Group Conditions (C1-C3 Condition)



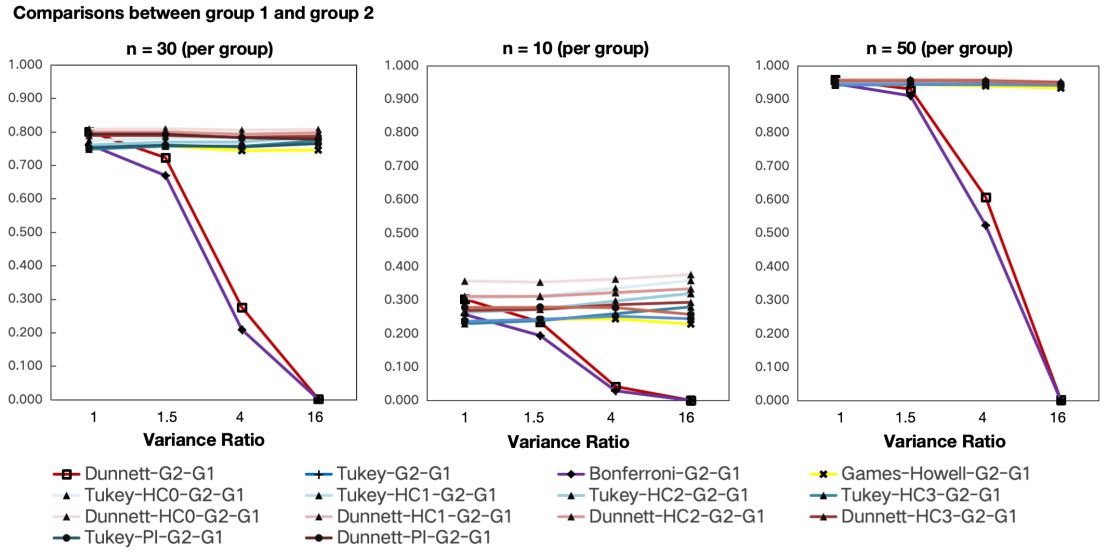
4.4 Power

The power results are presented below using faceted plots. In this study, to simultaneously assess the probability of correctly accepting H1 (power) and incorrectly rejecting H0 (FWER) under complex variance and sample settings, the means of the three groups are initially set as M1 = M3 $\neq$ M2, with the difference between M1 and M2 calculated based on variance and power to achieve a balance across different designs. Therefore, the plot for the comparison of G1-G2 reflects the magnitude of power (the probability of correctly detecting a significant difference in 5000 simulations), while the plot for the comparison of G1-G3 reflects the probability of false significance under designs that are supposed to be significant in other group comparisons. Each plot compares VR, sample size, and different methods. Results under balanced group conditions are illustrated in **Figures 4.32 and 4.33**.

As shown in **Figure 4.32**, for the classic methods, regardless of sample size, as VR increases, the power of the three methods sharply decreases, dropping to zero when VR = 16, indicating poor resistance to variance heteroscedasticity, which cannot be compensated for by increasing sample size. In contrast, the power of robust methods remains relatively stable, with minimal impact from variance heteroscedasticity, showing good adaptability to heteroscedasticity. However, sample size also had a large influence on power. Smaller sample sizes (e.g., n=10) corresponded to lower power values, ranging between 0.2 and 0.4. A power above 0.6 is generally considered medium. Therefore, none of these methods met this criterion with extremely small sample sizes.

Figure 4.32

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Balanced Group)



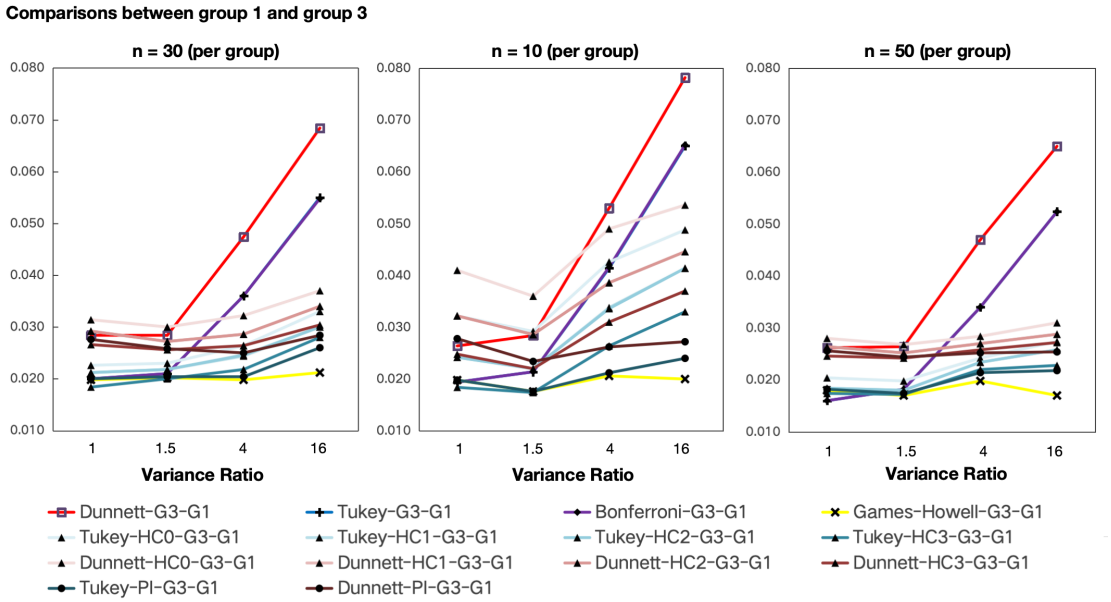
Comparing different robust methods, the test power at n = 30 and VR = 1.5 was as follows: Dunnett-HC0 (0.809), Dunnett-HC1 (0.800), Dunnett-HC2 (0.800), Dunnett-PI (0.792), Dunnett-HC3 (0.789), Tukey-HC0 (0.781), Tukey-HC1 (0.769), Tukey-HC2 (0.769),

Tukey-PI (0.760), Games-Howell (0.758), Tukey-HC3 (0.757).

In **Figure 4.33**, the false positive results of classical and robust methods exhibited different patterns. For classical methods, the false positive rates increased significantly as VR rose from 1 to 16, slightly exceeding the 0.05 threshold, with the highest rate reaching 0.080. In contrast, all robust MCT methods maintained false positive rates well below 0.050, staying within the acceptable range for wrongly rejecting H0. When the sample size was extremely small (n=10), the performance of robust methods showed greater fluctuation compared to middle (n=30) and large (n=50) sample sizes.

Figure 4.33

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Balanced Group)



Considering the combination of the HC family and Dunnett and Tukey, we found that while the Dunnett-HC series had the highest power, its false positive probability for H0 was unacceptable (exceeding 0.05 under various experimental conditions), and relatively speaking, Tukey-HC2 and Tukey-HC3 were better choices.

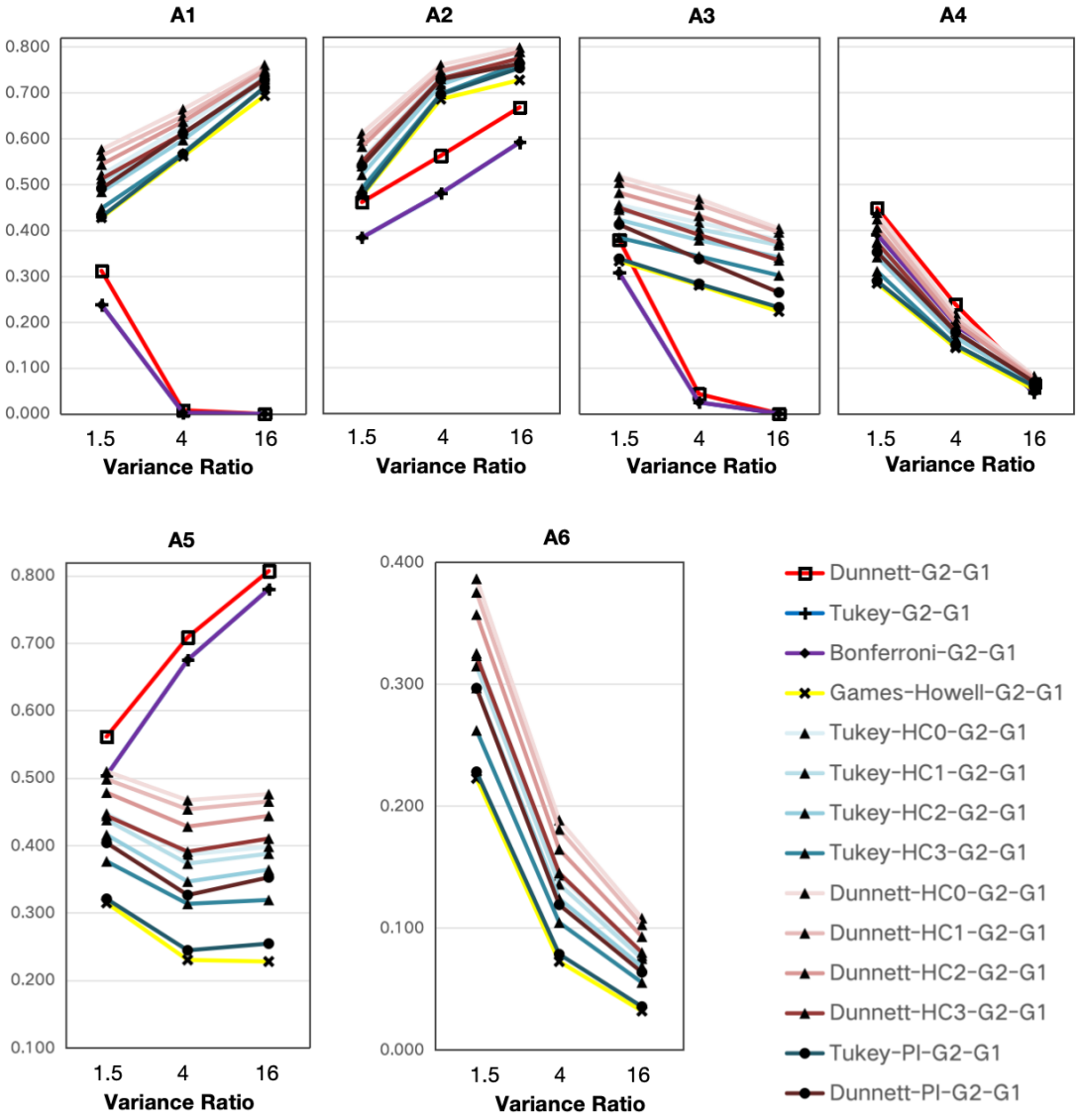
When comparing the combination of PI with Dunnett and Tukey, although Dunnett-PI had a slightly higher false positive probability than Tukey-PI when H0 is true, it remained within an acceptable range (below 0.05 in most experimental conditions) while also achieving higher power, so its overall performance was better than Tukey-PI.

As shown in **Figure 4.34**, the power of classical and robust methods varied considerably. For classical methods, power decreased to zero as VR increased from 1.5 to 16 under A1 and A3 conditions, increased to over 0.800 under A5 condition, and showed a similar pattern to robust procedures under other conditions. For robust procedures, power increased with VR under A1 and A2 conditions but decreased under A3 to A6 conditions. In most cases, the power of robust procedures remained below 0.6, indicating they failed to achieve medium power in complex situations. All methods performed worst under A6, where the pairing between sample size and variance was most negatively aligned.

Among all robust methods, Dunnett-HC methods demonstrated the best statistical power, outperforming the Tukey-HC series, Dunnett-PI, Tukey-PI, and Games-Howell methods. The statistical power of the three classic methods was also the lowest, with Dunnett being slightly higher than Tukey and Bonferroni. Overall, the methods' powers were ranked from high to low as follows: Dunnett-HC, Tukey-HC, Dunnett-PI, Tukey-PI, Games-Howell, Dunnett, Tukey/Bonferroni.

Figure 4.34

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 3 Different Sizes)

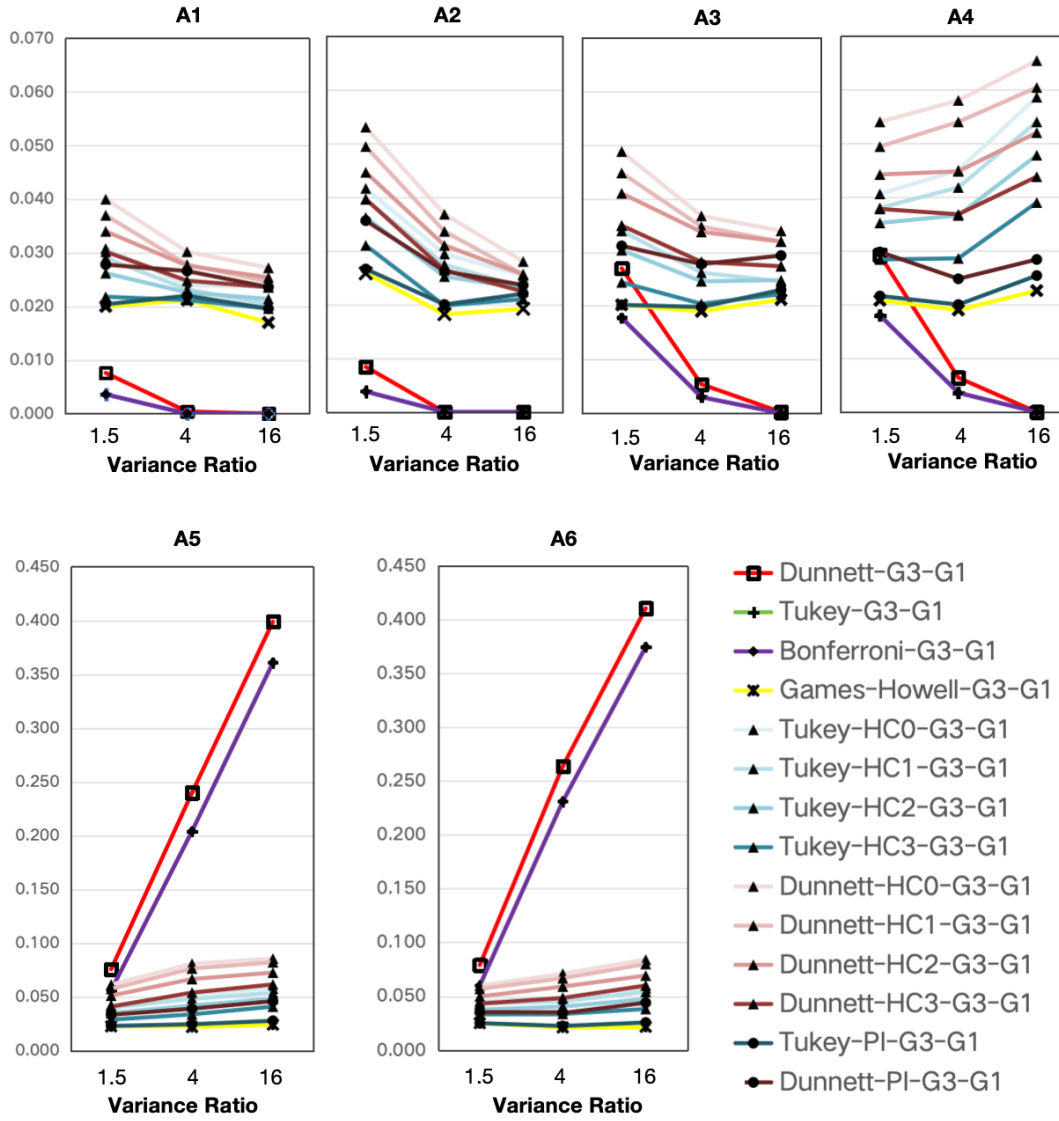


As shown in **Figure 4.35**, the false positive results of classical and robust methods exhibited different patterns. For classical methods, the false positive rates decreased significantly to zero as VR increased under A1 to A4 conditions but rose sharply to over 0.400 under A5 and A6 conditions. The performance of robust MCT procedures was much more stable, fluctuating within a range of 0.030 across all conditions except A4. The FWER

of all these procedures was controlled below 0.050 under A1 and A3 conditions, while some Dunnett-related robust procedures exceeded 0.050 under A2, A4, A5, and A6 conditions.

Figure 4.35

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 3 Different Sizes)



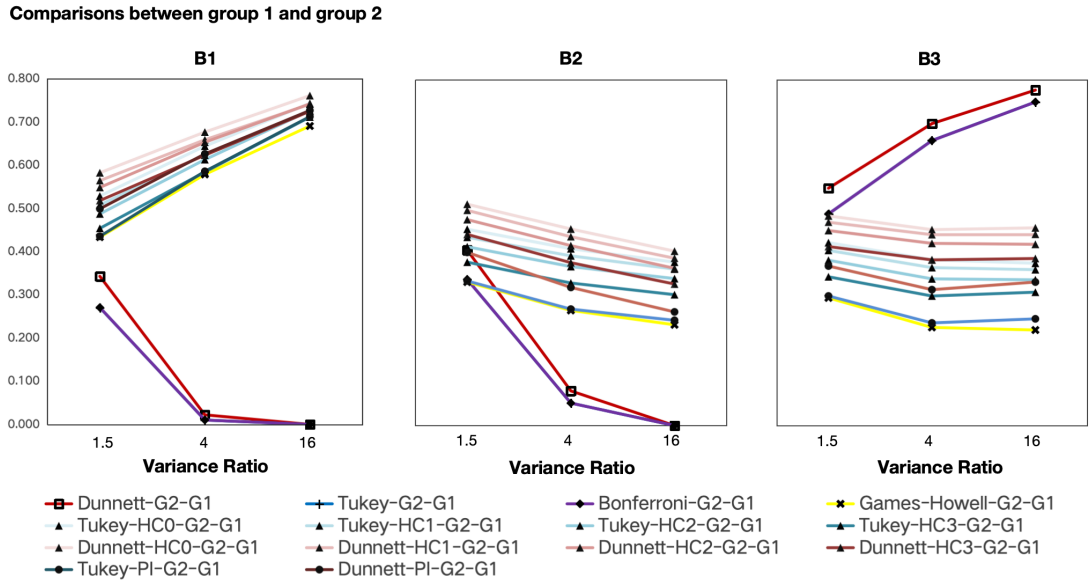
Although some methods can resist the impact of heteroscedasticity in terms of false positives, from the perspective of test power, robust methods were significantly affected and cannot be effectively controlled by the methods. For the HC family, the overall trend from

HC0 to HC3 was that as the HC value increases, the power decreases. The power of Dunnett-HC2 was comparable to Tukey-HC0, and the power of Dunnett-HC3 was comparable to Tukey-HC1. For PI, similarly, it performed better when combined with Dunnett, with Dunnett-PI having power equivalent to Tukey-HC1, and Tukey-PI having power equivalent to Tukey-HC3. Bonferroni and Tukey neither showed outstanding test power nor effective false positive control in complex situations, and should not be considered.

As shown in **Figure 4.36**, the power of classical and robust methods displayed different patterns. For classical methods, the power decreased to zero when VR sharply decreased to 0 under B1 and B2 conditions, but increased to nearly 0.800 under the B3 condition. For robust procedures, power increased with VR under B1, decreased under B2, and remained stable under B3 conditions.

Figure 4.36

The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Smaller Group)



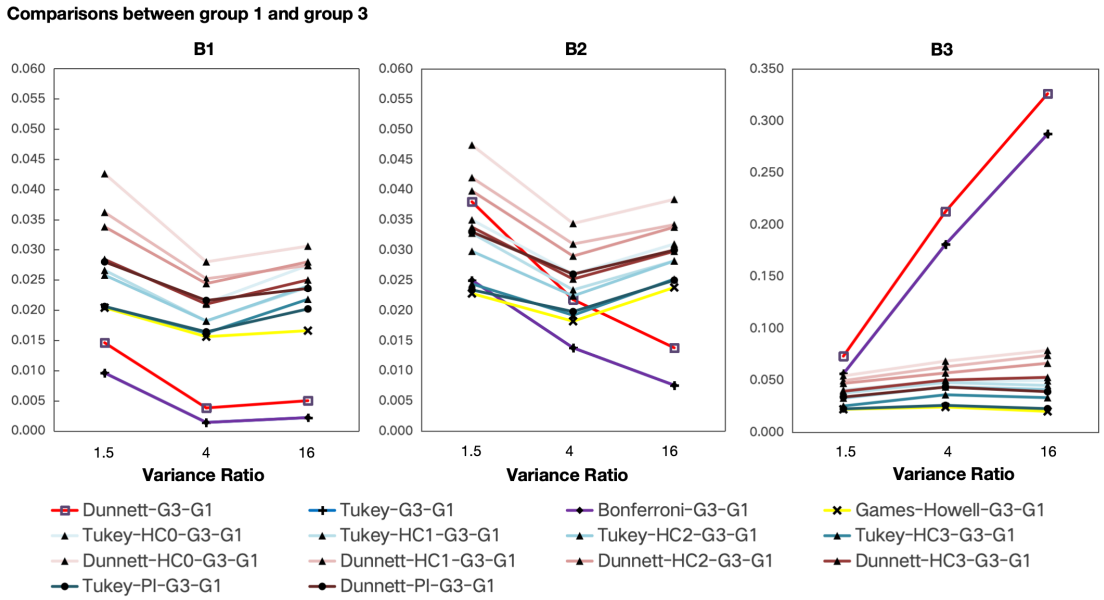
Overall, most power results did not reach the medium power threshold of 0.600, except

for the power under B1 when VR equaled 4 and 16, highlighting the significant negative impact of comparing an extremely small group (n=10) with a normal-sized group (n=30).

As shown in **Figure 4.37**, the false positive results of classical and robust methods exhibited different patterns. For classical methods, the false positive rates decreased as VR increased under B1 and B2 conditions but rose sharply to over 0.300 under the B3 condition. In contrast, the robust MCT procedures demonstrated much more stable performance, controlling the FWER below the 0.050 criterion across B1 to B3 conditions. The exception was some Dunnett-related robust procedures under the B3 condition, where the FWER slightly exceeded 0.050 when VR reached 16.

Figure 4.37

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Smaller Group)

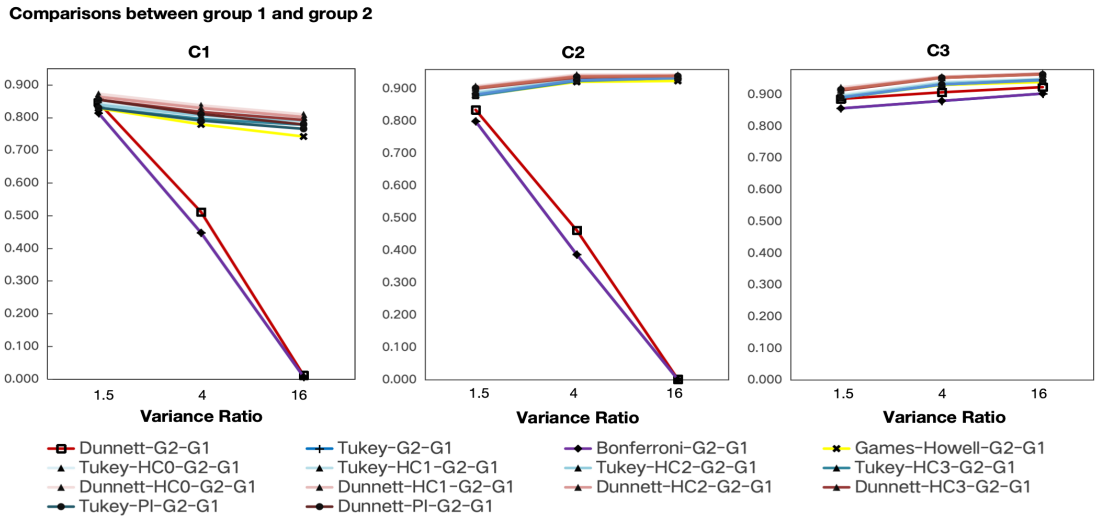


When comparing larger groups (n = 50 and 30), the conditions improved significantly. As shown in Figure 4.38, the power results of all robust procedures followed a consistent pattern, maintaining power levels at or above 0.800 regardless of the VR. In contrast, the

power of classical MCT methods dropped sharply to 0 as VR increased under C1 and C2 conditions, while under the C3 condition, they exhibited a pattern similar to that of the robust procedures. A power value above 0.800 is generally considered strong; thus, all robust methods met this criterion under these conditions.

Figure 4.38

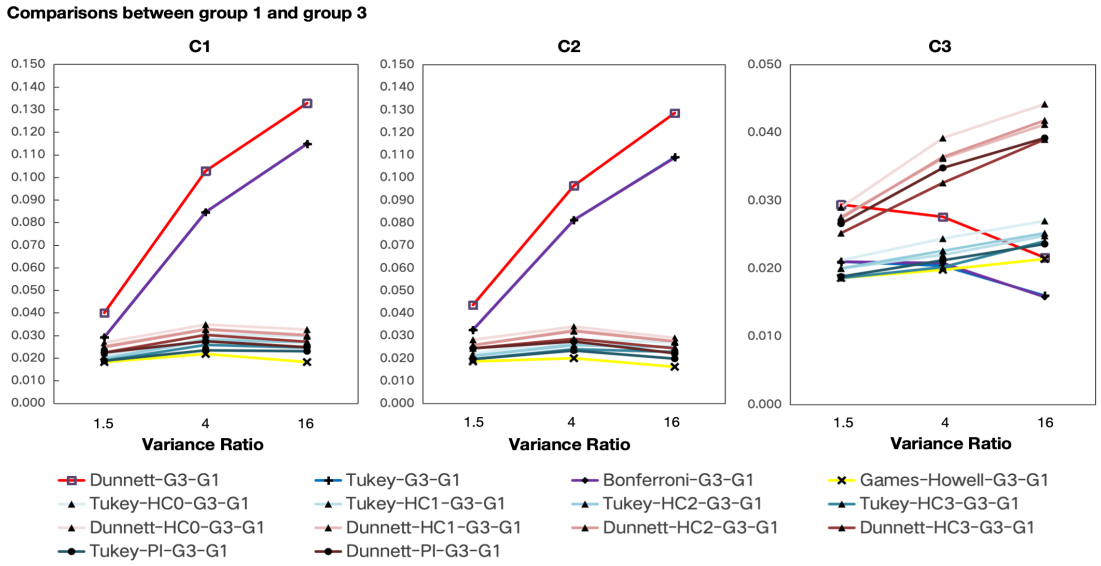
The Power Results (G1-G2 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Larger Group)



As shown in **Figure 4.39**, the FWER results for classical and robust methods exhibited distinct patterns. For the classical methods, the false positive rates increased substantially with higher VR under C1 and C2 conditions, whereas under the C3 condition, the rates decreased, ranging between 0.020 and 0.030. In contrast, the performance of robust MCT procedures was considerably more stable, maintaining FWER values below the 0.050 threshold across all conditions (C1 to C3).

Figure 4.39

The FWER Results (G1-G3 Contrasts) of Different MCT Methods under Different VRs and Sample Sizes (Unbalanced Group & 1 Larger Group)



In these manipulated conditions, the power of robust methods still followed the ranking from previous sections: Dunnett-HC showed higher power than Tukey-HC (comparable to Dunnett-PI), which was higher than Tukey-PI, while Games-Howell showed the lowest power. Within the two HC families, the more complex the HC adjustment (the higher the HC value), the lower the power. In general, under designs where two groups had equal sample sizes and one group was more extreme, considering the power results as well as the false positive rates, the methods with the best overall performance were still Tukey-HC2, Tukey-HC3, and Dunnett-PI.

Chapter V: Discussion

This chapter systematically evaluates the performance of various multiple comparison test (MCT) procedures across differing group structures and distributional assumptions. Beyond reporting numerical trends, we aim to summarize the implications of these findings for applied research settings involving unequal variances and unbalanced samples.

5.1 The Effect of Sample Size, VR and Their Pairing

An integrated analysis of the simulation results reveals that sample size alone substantially influences both the Type I error control and statistical power. As sample size increases from small (n = 10) to moderate (n = 30) and large (n = 50), robust procedures such as HC3 and PI exhibit progressively improved performance. While PI maintains conservative FWER control under small-sample conditions, it tends to sacrifice statistical power compared to HC3 in larger samples, supporting prior findings that PI is more suitable for extremely small designs, whereas HC3 becomes increasingly reliable with greater sample sizes. When n $\geqslant$ 30, HC3 consistently delivers both accurate FWER control and higher power, making it preferable in most applied contexts.

Variance ratio (VR) is a second critical factor that modulates method performance. Increasing VR values (from 1.5 to 16) lead to notable inflation in FWER among classical methods, regardless of sample size. In contrast, robust methods like HC3 and PI retain Type I error control even under high VR conditions. However, their relative power varies, with HC3 outperforming PI at moderate and large sample sizes. Games-Howell, while maintaining low FWER, exhibits considerable variability in CI width and lower power under these same

conditions. These patterns suggest that while HC3 is more adaptable to variance heteroscedasticity, it still benefits from increased sample size, particularly under extreme VR.

The pairing between sample size and variance distribution also shapes the performance of MCT procedures. In positively paired conditions—where larger groups are associated with higher variances—most methods exhibit acceptable FWER and reasonable power. However, in negatively paired conditions such as A5, A6, and B3, where the smallest group has the largest variance, classical methods and lower-order HC procedures become notably liberal. In these scenarios, only higher-order corrections such as HC3, as well as Games-Howell and Dunnett-PI, maintain robust performance. Notably, HC3's performance under negative pairing improves markedly when $n \geqslant 30$, reinforcing the idea that its robustness is conditional on sufficient sample size. These findings confirm the need for careful consideration of design features—including variance structure and group balance—when selecting appropriate multiple comparison methods.

5.2 Comprehensive Performance of all MCT Methods

Based on the systematic evaluation of the MCT methods, we can draw the following conclusions:

Firstly, under balanced group conditions, the classic methods (Bonferroni, Tukey, Dunnett) demonstrate lower test power when handling variance heteroscedasticity, especially when the sample size is small or there is significant variance difference, leading to a significant decrease in their effectiveness. In contrast, robust methods (such as Games-Howell, HC family, and PI methods) show better adaptability under conditions of variance

heteroscedasticity, effectively resisting the impact of heteroscedasticity while maintaining stable power. Specifically, the Games-Howell method, when facing unbalanced samples and variance heteroscedasticity, shows greater flexibility but relatively lower power. As the sample size increases, most robust methods show a noticeable improvement in power, especially when the sample size exceeds 30, reaching high power levels (>0.8).

Secondly, under conditions of unequal sample sizes, as variance heteroscedasticity increases, there is a noticeable difference in the power of the methods. Classic methods show poor power under small sample sizes and large variance conditions, particularly in extreme unbalanced group designs, where power significantly decreases. In contrast, robust methods are less affected by variance heteroscedasticity, allowing them to better handle sample imbalance and variance changes, showing more stable power.

In the analysis of confidence interval width, we observe differences in the performance of confidence intervals under variance heteroscedasticity for different methods. The confidence intervals for classic methods remain relatively stable and slightly expand as VR increases. In contrast, robust methods show more flexibility in their confidence interval widths, adjusting based on variance heteroscedasticity, especially under high VR conditions, where they tend to maintain more compact confidence intervals, demonstrating better accuracy. The Games-Howell method, under certain conditions, shows wider confidence intervals, especially with small samples and large variance. This may be due to the method's tendency to over-adjust confidence intervals to avoid false significance conclusions. However, despite the wider intervals, Games-Howell still effectively controls the false positive rate, showing strong adaptability. In comparison, the HC family and PI methods

adjust their confidence intervals more precisely, especially in complex designs, improving estimation accuracy by narrowing the interval widths.

Additionally, in extreme group imbalance designs (e.g., conditions B1 to C3), the differences between robust and classic methods become more pronounced. Specifically, under large sample and small variance conditions, robust methods perform better in controlling the false positive rate, while classic methods struggle with false positive control. In cases with larger variance, robust methods show better power and can maintain lower false positive rates, demonstrating their advantage in handling heteroscedasticity.

When considering which classic method to combine with robust procedures, comparison reveals that the Tukey strategy, as the underlying structure, shows better adaptability and synergy when combined with PI or HC methods. This combination is suitable for complex comparison structures and high-interference conditions, effectively controlling the probability of false significance. On the other hand, the Dunnett strategy still has value in experimental designs with clearly defined control groups, as it has strong power and relatively precise confidence intervals, but care should be taken regarding potential error accumulation when combined with methods that adjust for sensitivity.

Finally, combining power, false positive control, and confidence interval width, Tukey-HC2, Tukey-HC3, and Dunnett-PI are the optimal robust method choices, balancing high test power and good false positive control, while maintaining compact confidence intervals. In more complex design conditions, although Games-Howell has good false positive control, it has weaker power and is better suited for situations requiring high flexibility. Overall, under conditions of unbalanced samples and complex variance heteroscedasticity, robust methods

outperform classic methods, particularly when sample sizes are large and variance heteroscedasticity is significant, as robust methods effectively improve test power and control false positives while providing more precise confidence interval widths.

5.3 Comparison with Prior Studies

Integrating findings from our study with insights from prior relevant research, we can better interpret our empirical results and validate our methodological choices. Previous comparative work on robust procedures by Herberich et al. (2010), Hasler & Mario (2014), and Hothorn & Mario (2023) provide a critical foundation for understanding the differential behavior of HC3, PI, and other robust MCT methods across a range of variance heteroscedasticity and sample size imbalance scenarios. Our experimental design focuses more on practical psychological research conditions, providing a meaningful perspective.

In terms of methodology, our approach builds on and extends the simulation designs used in earlier studies. Unlike Hasler & Mario (2014), who focused on extremely small samples (e.g., n = 2), our design excludes such extremes, targeting more representative sample sizes for formal psychological studies. Furthermore, we incorporate both balanced and multiple forms of unbalanced group configurations (A1–A6, B1–B3, C1–C3) and introduce gradations of variance heteroscedasticity (VR = 1.5, 4, 16), thus allowing for a more nuanced and generalizable analysis.

Our findings reinforce and elaborate upon those of Herberich et al. (2010), who identified the sandwich estimator (particularly HC3) as a superior alternative to classical methods like Tukey's HSD in heteroscedastic settings. In our balanced conditions, HC3

consistently controlled FWER below 0.05 across all VR levels, even when sample sizes were small (n = 10). Moreover, unlike Herberich's study that used four group levels and linear variance/sample size scaling, our results showed that HC3 maintained robustness even in less structured pairing designs, suggesting its effectiveness under realistic psychological research conditions.

Additionally, our findings expand upon Hasler & Mario's (2014) conclusion that PI performs better than HC3 in extremely small sample settings, but that HC3 becomes increasingly preferable as sample size increases. In our context—focused on small (n = 10), medium (n = 30), and large (n = 50) samples—we observe that PI remains stable and conservative under most heteroscedastic scenarios but begins to lose relative power in moderate sample size settings compared to HC3. This supports Hothorn & Mario's (2023) recommendation of HC3 for moderate sample sizes and aligns with our conclusion that HC3 outperforms PI when $n \geqslant 30$. Furthermore, while Hothorn & Mario introduced the Bonferroni-Welch test as a competitor, our design excluded it to focus on methods more frequently used or adapted in psychological analysis.

The role of sample-size-variance pairings—emphasized in both Hasler & Mario (2014) and Hothorn & Mario (2023)—is further clarified in our results. For instance, in the A5 and A6 conditions (negative pairings), we confirmed that methods like HC0 and HC1 become liberal, especially when combined with Dunnett, mirroring findings from earlier studies that showed HC3 improves robustness only gradually when $n \geqslant 3$. Our results show that while HC3 handles negative pairing better than lower-order HC methods, Games-Howell and

Dunnett-PI maintain the most stable performance under extreme conditions, corroborating Hothorn & Mario's observations.

Moreover, our power analysis supports the assertion made by Herberich et al. (2010) that HC3 yields higher statistical power than Tukey's HSD under heteroscedasticity. In fact, across most of our design settings, HC3, Dunnett-PI, and Tukey-HC3 achieve power levels above 0.80 when sample sizes are moderate or large—even under extreme VR conditions. This extends Hothorn & Mario's (2023) finding that HC3 brings less power loss than PI and Bonferroni-Welch under unbalanced designs, particularly when the group with increased variance is not the target of comparison.

We also contribute new insights into the role of confidence interval width as a secondary robustness indicator. In our study, robust methods such as HC3 and PI produced narrower and more consistent CIs across imbalance designs, unlike Games-Howell, which exhibited greater width variability and higher CI exclusion misalignment rates. This further validates the conclusions from Hothorn & Mario (2023) that although HC3 may be liberal at small sample sizes, it provides more reliable confidence intervals at moderate sizes.

In summary, the convergent findings across studies reinforce the practical value of HC3 and PI, especially when combined with Tukey's structure. Future research should continue to explore robust estimation under even more realistic conditions, such as non-normal distributions and mixed-design ANOVA. We suggest that the direct use of classical MCT methods is no longer justified, and a move toward adaptive, robust alternatives should be advocated in statistical guidelines and educational materials.

5.4 Summary and Implications

Taken together, the findings underscore the inadequacy of classical multiple comparison procedures when faced with realistic data conditions, particularly those involving variance heteroscedasticity or unequal sample sizes. While classical methods such as Tukey's HSD and Dunnett's test are straightforward to implement and have long-standing use in statistical practice, their performance becomes unreliable in scenarios that violate homoscedasticity or balance assumptions—conditions that are common in psychological, biomedical, and social science research. Our simulations reveal that under such violations, classical methods frequently inflate Type I error rates and suffer from substantial power loss, leading to potentially misleading or non-replicable conclusions.

In contrast, robust procedures—particularly Games-Howell, HC3-based, and PI-based methods—demonstrate greater adaptability across a wide range of conditions. These methods offer superior control over familywise error rates and, in many cases, preserve statistical power even under small sample sizes or substantial variance heterogeneity. Their ability to adjust confidence intervals and test statistics in response to data irregularities reflects a methodological flexibility that classical methods lack. This robustness becomes especially critical in experimental designs where group sizes are unbalanced, or variance assumptions are difficult to justify or test.

Our study highlights the need for researchers and practitioners to move beyond routine reliance on classical MCTs and to adopt robust alternatives that better match the complexity of empirical data. The choice of method should be driven not by tradition or software defaults, but by the statistical characteristics of the data at hand. In light of these findings,

future methodological guidelines, educational resources, and statistical software defaults should prioritize robust procedures as the standard approach, especially in fields where small sample sizes or unequal variances are the norm. By encouraging the use of statistically sound and empirically validated methods, the research community can improve the reproducibility, reliability, and interpretability of findings across disciplines.

Chapter VI: Conclusion

This study systematically investigated the robustness of multiple comparisons test (MCT) strategies under conditions of variance heteroscedasticity, using Monte Carlo simulations. The research aimed to compare classical MCT methods with more robust procedures, specifically focusing on their ability to control false positive rates (FWER) and their statistical power under various real-world conditions. These conditions, including group size, variance ratio, and sample size deviation, are commonly encountered in psychological research but often violate the assumptions of normality and homogeneity that many classical MCT methods are based upon.

Through the simulations, we found that classical methods, such as Bonferroni, Tukey, and Dunnett, demonstrated good performance under balanced group conditions but suffered significant power loss and inflated FWER when faced with variance heteroscedasticity or unequal sample sizes. In particular, these methods exhibited substantial power reductions, especially in extreme group imbalance scenarios or when large variances were present. Conversely, robust methods such as Games-Howell, HC family, and PI procedures showed superior adaptability to these violations, maintaining stable power and controlling false positive rates more effectively.

The analysis of confidence interval (CI) widths revealed further insights into the methods' behavior. Classic methods maintained relatively inflexible CI widths, which slightly expanded as variance heteroscedasticity increased. In contrast, robust methods exhibited more flexible CI adjustments, with some methods, such as Games-Howell, showing wider CIs in extreme conditions, though still effectively controlling false positives. The HC

family and PI methods, on the other hand, provided more precise CI adjustments, especially in complex designs, ensuring better estimation accuracy.

The Monte Carlo simulations confirmed that robust methods, particularly Tukey-HC2, Tukey-HC3, and Dunnett-PI, provided the best combination of power and false positive control, especially under conditions of unbalanced samples and variance heteroscedasticity. The better combination for HC procedure is Tukey's test for lower FWER, while the better combination for PI procedure is Dunnett's test for higher power. These methods were found to maintain compact CIs, offering good control over false positives while achieving high power, particularly in larger sample sizes. Games-Howell, although excellent in controlling false positives, had lower power and was best suited for situations where flexibility was prioritized over power.

In conclusion, this study provides clear recommendations for the most appropriate MCT procedures under various conditions of heteroscedasticity. Tukey-HC2, Tukey-HC3, and Dunnett-PI emerge as the best robust methods for general use, offering a balance between test power and the ability to control false positive rates. These findings highlight the importance of considering both statistical power and robustness in psychological research, particularly when dealing with complex, real-world experimental designs. The results suggest that robust methods are not only more adaptable to assumption violations but also capable of improving the reliability and accuracy of statistical inferences in psychological studies.

Reference

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5), 502-508.

Neyman J & Pearson ES. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Bio-metrika*. 20A: 175–240.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *Bmj*, 316(7139), 1236-1238.

Qian Shi; Emily S. Pavey; Rickey E. Carter. (2012). Bonferroni-Based correction factor for multiple, correlated endpoints. *Pharmaceutical Statistics*, 11(4). doi:10.1002/pst.1514

Keselman, H. J. & Rogan, J. C. (1977). The Tukey multiple comparison test: 1953–1976. *Psychological Bulletin,* 84(5), 1050.

Nanda, A., Mohapatra, B. B., Mahapatra, A. P. K., Mahapatra, A. P. K. & Mahapatra, A. P. K. (2021). Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. *International Journal of Statistics and Applied Mathematics*, 6(1), 59-65.

Benjamini, Y. & Braun, H. (2002). John W. Tukey's contributions to multiple comparisons. *Annals of Statistics,* 1576-1594.

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.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75(372), 796-800.

Lee, S. & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test?. *Korean journal of anesthesiology*, 71(5), 353-360.

Kim, H. Y. (2015). Statistical notes for clinical researchers: post-hoc multiple comparisons. Restorative dentistry & endodontics, 40(2), 172-176.

Micceri, Theodore. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166. doi: 10.1037/ 0033-2909.105.1.156.

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology*, 51(1), 1–39.

Counsell, Alyssa; Harlow, Lisa. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology.. *Canadian Psychology/Psychologie Canadienne*, 58(2), 140–147.

Lisa M. Lix, Joanne C. Keselman and H. J. Keselman (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance "F" Test. *Review of Educational Research*, 66(4), 579–619.

Field, A.P., Wilcox, R.R.. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers, *Behaviour Research and Therapy.* 05.013.

Ruscio, J. & Roche, B. (2012). Variance heteroscedasticity in published psychological research. *Methodology*. 8(1), 1-11.

Grissom, R. J. (2000). Heteroscedasticity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68, 155–165.

Ruxton GD, Beauchamp G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral Ecology,* 19(3):690–693. DOI 10.1093/beheco/arn020.

Midway, S., Robertson, M., Flinn, S. & Kaller, M. (2020). Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. *PeerJ*, 8, e10387.

Herberich E., Sikorski J., Hothorn T. (2010). A Robust Procedure for Comparing Multiple Means under Heteroscedasticity in Unbalanced Designs. *PLoS ONE*, 5(3): e9788. https://doi.org/10.1371/journal.pone.0009788

Hothorn, T., Bretz, F. & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3), 346-363.

Achim Zeileis. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17. URL: http://www.jstatsoft.org/v11/i10/.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16, 1-16.

Freedman, D. A. (2006). On the so-called “Huber sandwich estimator” and “robust standard errors”. The American Statistician, 60(4), 299-302.

Hasler, M. Heteroscedasticity: multiple degrees of freedom vs. sandwich estimation. *Statistical Papers* 57, 55–68. (2016). Doi:10.1007/s00362014-0640-4

Carroll, R. J., Wang, S., Simpson, D. G., Stromberg, A. J. & Ruppert, D. (1998). The sandwich (robust covariance matrix) estimator. Unpublished manuscript.

Huber, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, LeCam, L. M. and Neyman, J. editors. University of California Press, pp. 221–233.

Hothorn, Ludwig & Hasler, Mario. (2023). The Dunnett procedure with possibly heterogeneous variances. arXiv:2303.09222

Petz, D. (2002). Covariance and Fisher information in quantum mechanics. *Journal of Physics A: Mathematical and General*, 35(4), 929.

Abt, M. & Welch, W. J. (1998). Fisher Information and Maximum-Likelihood Estimation of Covariance Parameters in Gaussian Stochastic Processes. *Canadian Journal of Statistics*, 26(1), 127-137.

Li, G., Tang, M., Charon, N. & Priebe, C. (2020). Central limit theorems for classical multidimensional scaling.

Durieu, O. & Tusche, M. (2014). An empirical process central limit theorem for multidimensional dependent data. *Journal of Theoretical Probability*, 27(1), 249-277.

Roberts, G. O. & Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1), 95-110.

Mario Hasler; Ludwig A. Hothorn. (2008). Multiple Contrast Tests in the Presence of Heteroscedasticity. *Biometrical Journal*, 50(5), 793–800. doi:10.1002/bimj.200710466.

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

Pallmann, P., & Hothorn, L. A. (2016). Analysis of means: a generalized approach using R. *Journal of Applied Statistics*, 43(8), 1541-1560.

Mondal, A., Sattler, P., & Kumar, S. (2023). Testing for trend in two-way crossed effects model under heteroscedasticity. *Test*, 32(4), 1434-1458.

Tamhane, A. C., & Xi, D. (2023). Multiplicity adjustments for the Dunnett procedure under heterokcedasticity. *Biometrical Journal*, 65(8), 2200300.

Victor J. Yohai. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, 15:642–65.

Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Martin Maechler, and others. (2007). Robust base: Basic Robust Statistics. URL: http://CRAN.R-project.org. R package version 0.2-8.

Konietschke, F., Hothorn, L. A., & Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals.

Konietschke, F., Bösiger, S., Brunner, E., & Hothorn, L. A. (2013). Are multiple contrast tests superior to the ANOVA?. The International Journal of Biostatistics, 9(1), 63-73.

Konietschke, F., Hothorn, L.A. & Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, 6, 738-759.

Konietschke, F., Bösiger, S., Brunner, E. & Hothorn, L. (2013). Are Multiple Contrast Tests Superior to the ANOVA?. *The International Journal of Biostatistics*, 9(1), 63-73. https://doi.org/10.1515/ijb-2012-0020

Hasler, M. (2014). Heteroscedasticity: multiple degrees of freedom vs. sandwich estimation. *Statistical Papers,* 57, 55 - 68.

Pallmann, P. & Hothorn, L.A. (2016). Analysis of means: a generalized approach using R. *Journal of Applied Statistics*, 43, 1541 - 1560.

Hothorn, L.A. & Kluxen, F.M. (2019). Robust multiple comparisons against a control group with application in toxicology. *arXiv: Applications*.

Hothorn, L.A. & Hasler, M. (2023). The Dunnett procedure with possibly heterogeneous variances. *arXiv: Applications.*

Hothorn, L.A. (2023). Consistent ANOVA-type tests for various effect sizes. *arXiv: Applications.*

Herberich, E. (2012). On the behavior of multiple comparison procedures in complex parametric designs.

Hothorn, Torsten & Kneib, Thomas & Bühlmann, Peter. (2012). Conditional Transformation Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 76. 10.1111/rssb.12017.

Dolgun, A. & Demirhan, H. (2017). Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution. *Communications in Statistics - Simulation and Computation*, 46, 5166 - 5183.

Umlauft, M., Placzek, M., Konietschke, F. & Pauly, M. (2017). Wild Bootstrapping Rank-Based Procedures: Multiple Testing in Nonparametric Split-Plot Designs. *arXiv: Statistics Theory*.

Umlauft, M., Placzek, M., Konietschke, F. & Pauly, M. (2019). Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *Journal of Multivariate Annual*., 171, 176-192.

Mondal, A., Pauly, M. & Kumar, S. (2022). Testing for ordered alternatives in heteroscedastic ANOVA under normality. *Statistical Papers*, 64, 19131941.

Tamhane, A.C. & Xi, D. (2023). Multiplicity adjustments for the Dunnett procedure under heteroscedasticity. *Biometrical Journal*, 65.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R. & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit?. *Behavior research methods*, 50(3), 937–962.

Ruscio, J. & Roche, B. (2012). Variance heteroscedasticity in published psychological research: A review and a new index. *Methodology*, 8, 1–11.

Moder, K. (2010). Alternatives to F-test in one way ANOVA in case of heteroscedasticity of variances (a simulation study). *Psychological Test and Assessment Modeling*, 52, 343–353

Lee, S. & Ahn, C. H. (2003). Modified ANOVA for unequal variances. *Communications in Statistics–Simulation and Computation*, 32, 987– 1004.

Grissom, R. J. (2000). Heteroscedasticity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68, 155–165.

Glass, G. V., Peckham, P. D. & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237–288.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11), 2074-2102.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics*, 17(4), 297-313.

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A. L., Heck, D. W., & Pawel, S. (2024). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting. *Psychological methods*.

Sauder, D. C., & DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science*, 2(1), 26-44.

Ryan, Thomas A. (1959). "Multiple comparison in psychological research". *Psychological Bulletin.* 56 (1). https://doi.org/10.1037/h0042478.

Dunn, Olive Jean. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association.* 56 (293): 52–64.

Tukey, John. (1949). "Comparing individual means in the Analysis of Variance". *Biometrics.* 5 (2): 99–114. doi:10.2307/300191.

Dunnett C. W. (1964.) "New tables for multiple comparisons with a control", *Biometrics,* 20:482–491.

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior research methods*, 53(6), 2576-2590.