國 立 成 功 大 學

統 計 學 系

博 士 論 文

比較多組平均數的異質變異數問題

Heteroscedasticity Problems in Comparison of Several Means

學生：王偉銘

Student： Wang Wei-Ming

指導教授：温敏杰博士

Advisor： Dr. Miin-Jye Wen

中 華 民 國 111 年 06 月

摘要

存在異質變異數的多組平均數比較，在統計學上是長期存在的問題。過去研究顯示，檢定統計量的分布受未知變異數的影響很大，並且在存在異質變異數時會不穩健。Bishop 和 Dudewicz (1978) 基於 Stein (1945) 的兩階段抽樣程序，提出了在變異數未知且不相等時的精確檢定量分配的方法，並在許多後續研究中用於解決各種多重比較問題。但是，兩階段抽樣程序是研究設計導向，需要在第二階段增加額外的樣本，缺乏實用性。在本論文中，我們分別應用了 Chen 和 Lam (1989)、Chen 和 Chen (1998) 以及 Chen (2001) 的單階段抽樣程序，來處理樣本數相等和不相等的兩種異質變異數下的平均數分析以及與單一控制組的多重比較問題。為了驗證此程序的品質，進行了整體型 I 誤差率的模擬，並提供了多個實例分析來說明程序如何使用。創建 *R* Shiny 的網路介面，供使用者可以更輕易的使用這些程序。

關鍵字: 異質變異數, 一階段抽樣程序, 平均數分析, 聯合信賴區間

Abstract

Comparison of several means in the presence of heteroscedasticity has been a long-standing problem in statistics. According to studies, the distribution of test statistics is highly influenced by unknown variances and is not robust in the presence of heteroscedastic variances. Bishop and Dudewicz (1978) proposed an exact analysis of variance when the variances are unknown and unequal, based on a two-stage sampling process described by Stein (1945). The approach has been used in a number of follow-up research to solve various multiple comparison problems. However, the two-stage sampling procedure is design-oriented which requires additional samples at the second stage, and it is inconvenient for the practice. In this thesis, we applied Chen and Lam's (1989), Chen and Chen's (1998) and Chen's (2001) single-stage sampling procedure, respectively, to analysis of means and multiple comparison with a control under heteroscedasticity with equal and unequal sample sizes. In order to validate the quality of the procedures, simulations of the family-wise error rate were undertaken, and various numerical examples were supplied to illustrate the procedures. R Shiny was utilized in the creation of an interface so that we can make the procedures more user-friendly.

Keyword: heteroscedasticity, single-stage sampling procedure, analysis of means, simultaneous confidence intervals

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Comparisons of k populations or treatment groups are required in many situations in the engineering, medical, pharmaceutical, biological, psychological research and other fields (Kiani et al., 2008; Lopez-Mejia and Roldan-Valadez, 2016; Castillo-Cuenca et al., 2021; Stanciu and Chis, 2021). Analysis of variance (ANOVA) is a regularly used statistical approach for determining whether or not the means of all treatments are equal. In a one-way ANOVA, if the null hypothesis is rejected, it means that at least one of the treatment means is statistically significant. In order to find the difference of among treatment groups or various combinations of groups, we do the multiple comparison tests (MCT). Different MCT methods can be performed depending on the purpose of the study, such as Tukey method for all-pair comparisons (MCA), Scheffé test for all-contrast comparisons (ACC), Dunnett method for multiple comparisons with a control (MCC), and multiple comparisons with the best (MCB), and so on. Previous researches have shown that post-hoc results from various multiple comparisons approaches are inconsistent, and in some cases, contradict ANOVA. An alternate strategy for solving the problem is analysis of means (ANOM). Ott (1967) developed a graphical statistical approach for ANOM that is used to compare k treatment means to see if any of them differ substantially from the overall mean. One of the benefits of ANOM is that it provides a control chart that is a practical and straightforward method of visually

displaying the difference between treatments and overall mean.

Most MCTs including ANOM have the same assumptions as ANOVA model, that is, the observations of treatment i ($i = 1, \ldots, k$) are independent and normally distributed with means and equal variances. Violations of these assumptions will result in poor power and its p-value can be biased due to their dependence on the unknown variances (Weerahandi, 1987; Krutchkoff, 1988). Nevertheless, it is not always reasonable to assume homogeneous variances between treatments, which might result in making conclusions that are not accurate. For instance, research on identifying the optimal dose may suffer from the issue of heteroscedasticity since the variance of the data is dependent on the dose effects. (For example, examine the data in Westfall (1997), which can be found in the R package $multcomp$.) Multiple comparisons in the presence of heteroscedasticity have been a concern in statistics for a very long time. Welch (1938) was the first who proposed a test that approximated the degree of freedom of the resulting t-distribution for comparing of means of two normally distributed populations with heterogeneous variances. Later, Satterthwaite (1946) expanded Welch's method to any linear function of means using the same concept. However, the t-distribution depends on unknown population variances σ_1^2, σ_2^2 and sample size n_1, n_2, it is impossible to derive a exact distribution of test statistic (Dudewicz and Mishra (1988), pp. 500-501). Bishop and Dudewicz (1978) used a two-stage sampling procedure originally proposed by Stein (1945) to construct an exact analysis of variance for the means of k independent normal populations where the variances are unknown and unequal. Bishop (1979) used a two-stage multiple comparisons for main effects and pairwise differences. The two-stage sampling procedure, on the other hand, is design-oriented, with equal sample sizes across treatment groups in the first stage. Extra samples should be obtained during the second step

due to heterogeneous variances. The quantity of samples required might be high, perhaps rendering the procedure non-viable during the second stage. When working on statistical data analysis, we frequently only have one single sample accessible, thus it may not be practical in real-world applications. In analysis of variance problems, Bishop (1976) employed the notion of two-stage sampling procedure for a single sample by separating it into two data sets. Chen and Lam (1989) then used Bishop's single-stage sampling procedure for the interval estimation of the largest normal mean. Lam (1992) also extended the single-stage technique to the subset selection procedure. Chen and Chen (1998) expanded the single-stage sampling procedure to one-way and two-way layout ANOVA in the presence of heteroscedasticity. Chen (2001) compared typical one-stage with two-stage sampling procedures for testing the equality of means hypothesis. Chen et al. (2004) tested the order treatment means in the one-way layout using a one-stage procedure. Under heteroscedasticity, an optimum confidence interval for the largest normal mean developed by Chen and Wen (2006). Chen and Wu (2012) used the findings to develop a simultaneous confidence interval for both the biggest and lowest means of several independent normal populations under heteroscedasticity. Wen et al. (2017) suggested a single-stage sampling procedure for the t best populations.

In this thesis, we used Chen and Lam's (1989), Chen and Chen's (1998) and Chen's (2001) single-stage sampling procedures, respectively, to determine the critical values of the analysis of means (ANOM) in the one-way layout. In addition to ANOM, we apply the single-stage sampling approach to multiple comparison with a control under heteroscedasticity with equal and unequal sample sizes. The rest of this thesis is organized as follows. In Section 2, a single-stage sampling procedure to ANOM under heteroscedasticity is introduced. Simulation studies for the empirical type I error rate are discussed, as well as numerical examples.

In Section 3, the single-stage sampling procedure for one-sided and two-sided intervals is introduced in MCC. The familywise error type I (FWE) was simulated in order to validate the quality of the procedures and a numerical example is demonstrated to show how to apply them. Finally, a conclusion and discussion is given in Section 4.

Chapter 2

Heteroscedastic Analysis of Means

ANOM is a graphical procedure that initially introduced by Ott (1967) for comparing k treatment means to specify if any of them has significant difference from the overall mean. The null hypothesis for ANOM and ANOVA are the same, but the alternative hypothesis of ANOM and ANOVA are different. The alternative hypothesis of ANOVA is that at least one of the means differs from the others, whereas with ANOM, the objective is to see whether or not one or more means are different from the overall mean. In contrast to ANOVA, ANOM is similar to a control chart in that it shows the upper and lower decision limits, the means of each treatment, and the overall mean. If a treatment mean exceeds the decision limits, it means that it is statistically different from the overall mean. As a result, ANOM has the benefit of presenting a graphical representation that allows one to quickly assess both the statistical and practical significance of the differences (Nelson and Dudewicz, 2002). For the analysis of variance problems, we can use a nonparametric method, such as the Kruskal–Wallis test. However, nonparametric methods typically have lower test power (Feir-Walsh and Toothaker, 1974). Nelson (1983) compared ANOM and ANOVA in terms of the sample sizes required for a certain power and found that ANOM is favorable compared to ANOVA. For large numbers of treatments, ANOVA requires substantially large sample sizes than ANOM.

A method developed by Nelson (1982) and Nelson (1991), respectively, for calculating

critical values for ANOM in both balanced designs and unbalanced designs. The variances must be equal in the classical ANOM model, which is the same as in an ANOVA model. However, equality of variance is not always feasible, because converting the data to make the variances equal is either difficult or inconvenient. The two-stage sampling procedure of analysis of mean under heteroscedasticity (HANOM) suggested by Nelson and Dudewicz (2002) eliminates the need for such assumptions for unknown and unequal variances. On the other hand, The two-stage sampling procedure is design-oriented, with equal sample sizes across treatment groups in the first stage and additional samples required in the second step due to heterogeneous variances. The quantity of samples required might be high, perhaps rendering the procedure non-viable during the second stage. When working on statistical data analysis, we frequently only have one single sample accessible, thus it may not be practical in real-world applications.

In this section, we applied several single-stage sampling procedures to eliminates the disadvantages of a two-stage sampling procedure and deals with the HANOM. In Subsection 2.1, we described the steps of a classical two-stage sampling procedure for HANOM. In Subsubsection 2.2.1, we applied Chen and Lam's (1989) single-stage sampling procedure to HANOM in the one-way layout with balanced design. In Subsubsection 2.2.2 and 2.2.3, we used Chen and Chen's (1998) and Chen's (2001) single-stage sampling procedures to determine the critical values and to solve the HANOM with an unbalanced design, respectively. A single-stage sampling procedure with a modified single-stage sampling procedure are introduced, respectively. In Subsection 2.3, we address a simulation study for empirical type I error rates. In Subsection 2.4, numerical examples are given.

2.1 Two-Stage Sampling Procedure for HANOM

In this subsection, we introduce a two-stage sampling procedure for HANOM that proposed by Nelson and Dudewicz (2002) that uses a weighted sample mean to estimate the population mean. Let k be the number of treatments be compared, and X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n_i)$ be random samples from a normal distribution with unknown mean, μ_i, and unknown and unequal variance, σ_i^2. Collecting data for and performing a HANOM consists of the subsequent step:

(i) Take an initial sample n_0 (≥ 2) from each i $(i = 1, \ldots, k)$ population, and calculate the sample means $\overline{X}_{i,n_0}$ and sample variances S^2_{i,n_0}.

(ii) Specify the significance α, power γ, and the difference δ between any two treatment means that will lead to rejection of the all of treatment means are equal hypothesis. From Figure A1-A36 in Dudewicz and Nelson (2003), given k, α, γ, and degree of freedom $df = n_0 - 1$ combination, find the corresponding value of ω. Alternatively, one could choose a value of ω using Equation (2.1) below based on desired sample sizes.

(iii) Calculate

$$n_i = \max\left\{n_0 + 1, [\![\left(\frac{\omega}{\delta}\right)^2 S^2_{i,n_0}]\!] + 1\right\} \tag{2.1}$$

for i $(i = 1, \ldots, k)$ population, where $[\![x]\!]$ denoted the greatest integer in x.

(iv) Take the remaining $n_i - n_0$ $(i = 1, \ldots, k)$ observation to calculate the sample means

$$\overline{X}_{i,n_i-n_0} = \frac{X_{i,n0+1} + \ldots + X_{i,n_i}}{n_i - n_0}.$$

(v) For each i $(i = 1, \ldots, k)$ population, compute the weights

$$b_i = \frac{n_i - n_0}{n_i}\left[1 + \sqrt{\frac{n_0}{n_i - n_0}\left(\left[\frac{\delta}{\omega}\right]^2 \frac{n_i}{S^2_{i,n_0}} - 1\right)}\right]$$

and the weighted sample means

$$\widetilde{\overline{X}}_i = (1 - b_i)\overline{X}_{i,n_0} + b_i\overline{X}_{i,n_i - n_0}$$

and the over all mean

$$\widetilde{\overline{X}} = \frac{\sum_{i=1}^{k} \widetilde{\overline{X}}_i}{k}.$$

(vi) Calculate lower decision line (LDL) and upper decision line (UDL) of a HANOM:

$$\widetilde{\overline{X}} \pm H(\alpha; k; n_0 - 1)\frac{\delta}{\omega}$$

where $H(\alpha; k; n_0 - 1)$ are found in Table A.1 in Nelson and Dudewicz (2002).

If the null hypothesis $H_0 : \mu_1 = \ldots = \mu_k$ is true, then all k treatment means lie between LDL and UDL. We shall reject the null hypothesis, H_0, if any of the weighted sample means goes outside of LDL or UDL.

2.2 Single-Stage Sampling Procedure for HANOM

The two-stage procedure is design-oriented, which implies that a final sample size must be determined with a significance α, power γ, and effect size δ between any two treatment

means. When there is only one sample process available or the requisite two-stage sample quantities are not fulfilled, the single-stage procedure is a viable alternative to the two-stage procedure. We employed Chen and Lam's (1989) single-stage sampling procedure to address the HANOM when the sample size was equal. We used Chen and Chen's (1998) and Chen's (2001) single-stage and modified single-stage sampling procedures, respectively, to determine critical values and solve the HANOM under unbalanced design. The general single-stage procedure is detailed as follows:

Assume there are k treatment groups and n_i observations in the i-th group. Assume X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n_i)$ is a random sample from a normal distribution with an unknown mean, μ_i, and an unknown and unequal variance, σ_i^2. Bishop (1976) was the first to propose splitting a sample into two data sets in an ANOVA problem and many following researchers have adopted this single-stage approach for a number of studies (Lam, 1987; Chen and Lam, 1989; Wen et al., 2007; Chen et al., 2009; Chen et al., 2011; Wen et al., 2017; Wen et al., 2022; Wang et al., 2022). We randomly selected $n_i - 1$ observations from each i-th $(i = 1, \ldots, k)$ population to determine the sample mean and variance, respectively, by

$$\overline{X}_i = \frac{\sum_{j=1}^{n_i-1} X_{ij}}{n_i - 1} \quad \text{and} \quad S_i^2 = \frac{\sum_{j=1}^{n_i-1} (X_{ij} - \overline{X}_i)^2}{n_i - 2} .$$

The sample size, n_i (≥ 3), of each i $(i = 1, \ldots, k)$ population is divided into two parts to calculate the weighted sample mean. These two parts include the first $n_i - 1$ observations and the remaining observation.

$$\widetilde{X}_i = \sum_{j=1}^{n_i} \omega_{ij} X_{ij} = U_i \sum_{j=1}^{n_i-1} X_{ij} + V_i X_{in_i}, \tag{2.2}$$

where

$$\omega_{ij} = \begin{cases} U_i, \ j = 1, 2, \ldots, n_i - 1 \\ V_i, \ j = n_i \ . \end{cases}$$

The weights of U_i and V_i of the n_i observations satisfy the following two conditions

$$(n_i - 1)U_i + V_i = 1 \tag{2.3}$$

$$(n_i - 1)U_i^2 + V_i^2 = \frac{S_{[k]}^2}{n_i S_i^2} \tag{2.4}$$

where $S_{[k]}^2$ is the maximum of $\{S_1^2, \ldots, S_k^2\}$. Equation (2.3) has the benefit of implying $\sum_{j=1}^{n_i-1} U_i + V_i = \sum_{j=1}^{n_i} \omega_{ij} = 1$ that makes the weighted sample mean, $\widetilde{X}_i$ in (2.2), an unbiased estimate of the treatment mean, μ_i, and Equation (2.4) implies $\sum_{j=1}^{n_i} \omega_{ij}^2 S_i^2 = S_{[k]}^2/n_i$ to eliminate the impact of the unknown and unequal variances according to Stein (1945). The solutions of the weights of U_i and V_i are determined using Equations (2.3) and (2.4).

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{1}{n_i - 1}[S_{[k]}^2/S_i^2 - 1]}, \tag{2.5}$$

and

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{(n_i - 1)[S_{[k]}^2/S_i^2 - 1]}. \tag{2.6}$$

As a result, the sampling distribution of the weighted sample mean, $\widetilde{X}_i$ ($i = 1, \ldots, k$),

is represented by

$$\widetilde{X}_i \sim N\left(\mu_i,\ \sum_{j=1}^{n_i} \omega_{ij}^2\ \sigma_i^2\right).$$

Forthermore, the transformation of the weighted sample mean given the sample variance, S_i^2 ($i = 1, \ldots, k$), is as follows:

$$T_i = \frac{\widetilde{X}_i - \mu_i}{\sqrt{\sum_{j=1}^{n_i} \omega_{ij}^2\ S_i^2}} = \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n_i}},\quad i = 1, \ldots, k, \tag{2.7}$$

is a conditional normal distribution with mean zero and variance σ_i^2/S_i^2 ($i = 1, \ldots, k$) and is also an unconditional Student's t distribution with degrees of freedom $n_i - 2$ (Wen and Chen, 1994; Chen and Chen, 1998).

2.2.1 Single-Stage Sampling Procedure with Balanced Design

The One-Way Layout Model

The model is how we conduct the HANOM in the one-way layout. Let X_{ij} be the j-th observation from the i-th population, we assume

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

where μ_i is the mean for the i-th treatment, ε_{ij} is the random error associated with X_{ij}, and the assumptions of normality and unequal variances imply $\varepsilon_{ij} \sim N\left(0,\ \sigma_i^2\right), i = 1, \ldots, k; j = 1, \ldots, n$. That is, we need to test if the k treatments which are balanced and have an effect

on the means, $\mu_i,\ i = 1, \ldots, k$. Thus, the null hypothesis is

$$H_0 : \mu_1 = \ldots = \mu_i = \ldots = \mu_k = \mu_.,$$

and the alternative hypothesis is that one or more treatment means are different from the overall mean.

The primary function of HANOM is to make it easier for people to make decisions by using a HANOM decision chart (Nelson and Dudewicz, 2002). HANOM can also identify which treatment means are significantly different from the overall mean. A HANOM decision chart has three lines:

(i) Centerline is

$$\overline{\widetilde{X}} = \frac{\sum_{i=1}^{k} \widetilde{X}_i}{k}$$

which represents the average of all weighted sample means.

(ii) Lower Decision Line (LDL) is

$$\overline{\widetilde{X}} - d^* \times \frac{S_{[k]}}{\sqrt{n}} \tag{2.8}$$

(iii) Upper Decision Line (UDL) is

$$\overline{\widetilde{X}} + d^* \times \frac{S_{[k]}}{\sqrt{n}} \tag{2.9}$$

where $d^* = d^*(\frac{\alpha}{2}, k, df) > 0$ is the critical value for the HANOM decision chart. The d^*

value is dependent on the level of significance, α, k treatment means being compared, and the degree of freedom, $df = n - 2$. We will discuss the details of critical values in the next paragraph.

If the null hypothesis is correct, then all of the weighted sample means, $\widetilde{X}_i$, should be near to the overall weighted sample mean, $\overline{\widetilde{X}}$, and all of the k treatment means should lie between LDL and UDL. We shall reject the null hypothesis, H_0, if any of the weighted sample means goes outside LDL or UDL.

Determination of Critical Values

We define the rejection region of treatment i as

$$\left| \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}} \right| > d^*, \ i = 1, \ldots, k,$$

which is similar to Nelson (1982) and Nelson and Dudewicz (2002). As a result, the level of significance, α, for HANOM is

$$\begin{aligned} 1 - \alpha =& P\left(\left| \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}} \right| \leq d^*, \ i = 1, \ldots, k \right) \\ =& P\left(-d^* \leq \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}} \leq d^*, \ i = 1, \ldots, k \right) \\ =& P\left(-d^* \leq \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n}} - \frac{\overline{\widetilde{X}} - \mu_.}{S_{[k]}/\sqrt{n}} + \frac{\mu_i - \mu_.}{S_{[k]}/\sqrt{n}} \leq d^*, \ i = 1, \ldots, k \right). \end{aligned} \tag{2.10}$$

From Equation (2.7), we have $T_i = \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n}}, \ i = 1, \ldots, k$ in Equation (2.10) are Student's t distribution with degree of freedom $n - 2$, while $\bar{T} = \sum_{i=1}^{k} T_i/k = \frac{\overline{\widetilde{X}} - \mu_.}{S_{[k]}/\sqrt{n}}$. Under the null hypothesis, $H_0 : \mu_1 = \ldots = \mu_k = \mu_.$, the term $\frac{\mu_i - \mu_.}{S_{[k]}/\sqrt{n}}$ in Equation (2.10) equals zero and

$1-\alpha$ will be rewritten by

$$\begin{aligned} 1-\alpha &= P\left(-d^* \leq T_i - \bar{T} \leq d^*,\ i=1,\ldots,k\right) \\ &= P\left(-d^* \leq \widetilde{T}_i \leq d^*,\ i=1,\ldots,k\right) \end{aligned} \tag{2.11}$$

where $\widetilde{T}_i = T_i - \bar{T},\ i=1,\ldots,k$, and $\widetilde{T}_{[1]}$ is the minimum of $\left\{\widetilde{T}_1,\ldots,\widetilde{T}_k\right\}$ and $\widetilde{T}_{[k]}$ is the maximum of $\left\{\widetilde{T}_1,\ldots,\widetilde{T}_k\right\}$. Here, we can rewrite Equation (2.11) as follows.

$$\begin{aligned} 1-\alpha &= P\left(-d^* \leq \widetilde{T}_{[1]} \leq \widetilde{T}_{[k]} \leq d^*\right) \\ &= 1 - P\left(\left\{\widetilde{T}_{[1]} < -d^*\right\} \cup \left\{\widetilde{T}_{[k]} > d^*\right\}\right) \\ &\geq 1 - P\left(\widetilde{T}_{[1]} < -d^*\right) - P\left(\widetilde{T}_{[k]} > d^*\right). \end{aligned}$$

Here, we take

$$P\left(\widetilde{T}_{[1]} < -d^*\right) + P\left(\widetilde{T}_{[k]} > d^*\right) \leq \alpha.$$

Because $(\widetilde{T}_1,\ldots,\widetilde{T_k})$ is symmetric according to Wu and Chen (1998), we allow

$$P\left(\widetilde{T}_{[1]} < -d^*\right) = \frac{\alpha}{2} = P\left(\widetilde{T}_{[k]} > d^*\right)$$

As a consequence, we use the Monte-Carlo simulation to find out the approximate sampling distribution of $\widetilde{T}_{[k]}$. To do this we need to generate random numbers from k Student's t distribution with specific degrees of freedom and calculate $\widetilde{T}_i,\ i=1,\ldots,k$. Afterwards, for each run, we sort these $\widetilde{T}_i,\ i=1,\ldots,k$ and choose the maximum value $\widetilde{T}_{[k]}$. An approx-

imation of the probability of the distribution of $\widetilde{T}_{[k]}$ can be made by selecting the cumulated number from the ranking list of the 100,000 values of $\widetilde{T}_{[k]}$ by 100,000 independent runs. This gives an approximation of the probability of inclusion being at least $1-\alpha/2$. Using R simulation program for 100,000 simulation runs, we take the $100(1-\alpha/2)$-th percentile of the approximate sampling distribution of $\widetilde{T}_{[k]}$, that is $P(\widetilde{T}_{[k]} < d^*) = 1-\alpha/2$. The simulated values of d^* when $1-\alpha = 0.9, 0.95, 0.99$ are given in Tables A.1-A.3.

Computing the p-value

According to Nelson et al. (2005), the p-value for HANOM can be determined as the value at which the weighted sample mean, $\widetilde{X}_i$, is the furthest distant from the overall weighted sample mean, $\overline{\widetilde{X}}$. More precisely, the p-value of the single-stage sampling procedure for HANOM is determined by

$$
\begin{aligned}
&P\left(\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}}\right| \le d^*\left(\frac{p}{2},k,n-2\right),\ i=1,\ldots,k\right)\\
&= P\left(\max_i\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}}\right| \le d^*\left(\frac{p}{2},k,n-2\right)\right)\\
&= 1-p.
\end{aligned}
$$

So, p is calculated by solving the equation

$$
\max_i\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}}\right| = d^*\left(\frac{p}{2},k,n-2\right).
$$

We raise p from 0 to 1 by 0.001 using the R simulation program. Following the execution of the HANOM procedure, we try to identify the $(100\frac{p}{2})$ percentile of the d^* distribution and

find the minimized p as the p-value for HANOM, which is

$$\min_{p} \left| \left\{ \max_{i} \left| \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}} \right| - d^* \left(\frac{p}{2}, k, n-2 \right) \right\} \right|$$

As a result of the above procedure, we may acquire a control chart, critical value, and p-value. We offered an interface using R Shiny to acquire the critical value and the p-value of the procedure for HANOM including the HANOM chart, so that users could easily apply the single-stage sampling procedure for HANOM. The URL of the interface can be accessed at `https://weiming3524.shinyapps.io/hanom/`.

R Shiny interface for HANOM

The interface provides the HANOM chart by uploading a dataset, as shown in Figure K.1 in Appendix K. To execute the single-stage sampling procedure for HANOM, the first step is to prepare your data files in comma separated values (CSV) format. The CSV file contains 2 columns with column names. The first column is the values of treatments and the second column is the group of treatments. After inputting a data file, the second step is to choose a significant level, 0.05, as the default. Finally, click the Run button and wait a few seconds, the describe statistic, the summary statistics, the p-value for HANOM, and the HANOM chart, respectively, will show on the display board.

2.2.2 Single-Stage Sampling Procedure with Unbalanced Design

The critical value of HANOM under balanced design depends on a k variates joint t distribution that have the same degrees of freedom. However, under an unbalanced design, the degrees of freedom are different among the k variates of t distribution, and this joint t

distribution could be not symmetric which may cause troubles in solving the critical value. Therefore, the following HANOM procedures for unbalanced design are proposed.

The One-Way Layout Model

When the sample size are unequal, we used Chen and Chen's (1998) single-stage sampling procedure that is referred to as $\mathcal{P}_1$ to solve the HANOM in the one-way layout model. We consider

$$X_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, \ldots, k; j = 1, \ldots, n_i,$$

where μ_i is the mean for the i-th treatment, ε_{ij} is the random error associated with X_{ij}, and the assumptions of normality and unknown and unequal variances imply $\varepsilon_{ij} \sim N\left(0,\ \sigma_i^2\right)$. The objective is the same as Subsubsection 2.2.1 to test the null hypothesis that the population means, $\mu_i,\ \ i = 1, \ldots, k$, are all equal. The three lines of HANOM decision chart are as follows:

(i) Centerline

$$\overline{\widetilde{X}} = \frac{\sum_{i=1}^{k} \widetilde{X}_i}{k}$$

which represents the average of all weighted sample means.

(ii) Lower Decision Line (LDL)

$$\overline{\widetilde{X}} - h^* \times \frac{S_{[k]}}{\sqrt{n_i}} \tag{2.12}$$

(iii) Upper Decision Line (UDL)

$$\overline{\widetilde{X}} + h^* \times \frac{S_{[k]}}{\sqrt{n_i}} \tag{2.13}$$

where $h^* = h^*(\frac{\alpha}{2}, k, n_1 - 2, \ldots, n_k - 2)(> 0)$ is the critical value for the HANOM decision chart. The h^* value is dependent on the level of significance, α, k treatment means being compared, and the degree of freedom, $n_i - 2$ where $i = 1, \ldots, k$.

If the null hypothesis is correct, then all the weighted sample means, $\widetilde{X}_i$, should be near to the overall weighted sample mean, $\overline{\widetilde{X}}$, and all of the k treatment means should fall between LDL and UDL. We shall reject the null hypothesis, H_0, if any of them goes outside LDL or UDL.

Determination of Critical Values

The rejection region of treatment i similar to Nelson (1982) as follows:

$$\left| \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}} \right| > h^*, \ i = 1, \ldots, k.$$

As a consequence, the decision limits for HANOM are

$$\begin{aligned}
1 - \alpha =& P\left(\left| \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}} \right| \leq h^*, \ i = 1, \ldots, k \right) \\
\overset{H_0}{=} & P\left(-h^* \leq \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{1}{k} \sum_{z=1}^{k} \frac{\widetilde{X}_z - \mu_z}{S_{[k]}/\sqrt{n_i}} \leq h^*, \ i = 1, \ldots, k \right) \\
=& P\left(-h^* \leq \left(\frac{k-1}{k} \right) T_i - \sqrt{n_i} \left(\frac{1}{k} \sum_{z \neq i}^{k} \frac{T_z}{\sqrt{n_z}} \right) \leq h^*, \ i = 1, \ldots, k \right) \qquad (2.14) \\
=& P\left(-h^* \leq \widetilde{T}_i \leq h^*, \ i = 1, \ldots, k \right) \qquad (2.15)
\end{aligned}$$

where $\widetilde{T}_i = (\frac{k-1}{k})T_i - \sqrt{n_i}(\frac{1}{k}\sum_{z\neq i}^{k}\frac{T_z}{\sqrt{n_z}}),\ i = 1,\ldots,k$. The proof is provided in full in Appendix B. T_i and T_z in Equation (2.14) are Student's t distributions with degrees of freedom $n_i - 2$ $(i = 1,\ldots,k)$ and $n_z - 2$ $(z = 1,\ldots,k)$, respectively, according to Equation (2.7). In this case, $\widetilde{T}_{[1]}$ represents the minimum of $\left\{\widetilde{T}_1,\ldots,\widetilde{T}_k\right\}$ and $\widetilde{T}_{[k]}$ represents the maximum of $\left\{\widetilde{T}_1,\ldots,\widetilde{T}_k\right\}$. The critical values of $\widetilde{T}_{[1]}$ and $\widetilde{T}_{[k]}$ differ because the degrees of freedom of $\widetilde{T}_{[1]}$ and $\widetilde{T}_{[k]}$ are different. The Equation (2.15) has been rewritten as follows.

$$\begin{aligned}
1-\alpha &= P\left(-h^* < \widetilde{T}_{[1]} \leq \widetilde{T}_{[k]} < h^*\right) \\
&\geq 1 - P\left(\widetilde{T}_{[1]} < -h^*\right) - P\left(\widetilde{T}_{[k]} > h^*\right) \\
&\geq 1 - 2\max\left\{P\left(\widetilde{T}_{[1]} < -h^*\right), P\left(\widetilde{T}_{[k]} > h^*\right)\right\}.
\end{aligned}$$

As a result, we choose h^* as the critical value to meet the following condition:

$$\max\left\{P\left(\widetilde{T}_{[1]} < -h^*\right), P\left(\widetilde{T}_{[k]} > h^*\right)\right\} \leq \frac{\alpha}{2}. \tag{2.16}$$

It's difficult to get the exact joint distribution of $(\widetilde{T}_1,\ldots,\widetilde{T_k})$. As a consequence, we use the Monte-Carlo simulation to find out the approximate sampling distribution of $\widetilde{T}_{[1]}$ and $\widetilde{T}_{[k]}$. To do this we need to generate random numbers from k Student's t distribution with specific degrees of freedom and calculate $\widetilde{T}_i,\ i = 1,\ldots,k$. Afterwards, for each run, we sort these $\widetilde{T}_i,\ i = 1,\ldots,k$ and choose the minimum value $\widetilde{T}_{[1]}$ and the maximum value $\widetilde{T}_{[k]}$ at every single run. An approximation of the probability of inclusion in the distribution of $\widetilde{T}_{[1]}$ and $\widetilde{T}_{[k]}$ can be made by selecting the cumulated number from the ranking list of the values of $\widetilde{T}_{[1]}$ and $\widetilde{T}_{[k]}$, respectively, by 100,000 independent runs. This gives an approximation of the

probability of inclusion being at least $1-\alpha/2$. We take the $100(1-\alpha/2)$-th percentile of the approximate sampling distribution of $\widetilde{T}_{[k]}$ and the $100(\alpha/2)$-th percentile of the approximate sampling distribution of $\widetilde{T}_{[1]}$, that is $P(\widetilde{T}_{[k]} < h^*) = 1-\alpha/2$ and $P(\widetilde{T}_{[1]} < h^*) = \alpha/2$, respectively, and select the larger critical value as h^* according to Equation (2.16).

Computing the p-value

Similar to balanced design, the p-value for HANOM can be determined as the value at which the weighted sample mean, $\widetilde{X}_i$, is the furthest distant from the overall weighted sample mean, $\overline{\widetilde{X}}$. Therefore, the p-value of the procedure $\mathcal{P}_1$ for HANOM is determined by

$$P\left(\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}}\right| \le h^*\left(\frac{p}{2}, k, n_1-2, \ldots, n_k-2\right),\ i=1,\ldots,k\right)$$
$$= P\left(\max_i\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}}\right| \le h^*\left(\frac{p}{2}, k, n_1-2, \ldots, n_k-2\right)\right) = 1-p.$$

The p-value can be obtained by solving the equation above. We raise p from 0 to 1 by 0.001 using the R simulation program. Following the execution of the HANOM procedure, we try to identify the $(100\frac{p}{2})$ percentile of the h^* distribution and find the minimized p as the p-value of the $\mathcal{P}_1$ for HANOM, which is

$$\min_p\left|\left\{\max_i\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}}\right| - h^*\left(\frac{p}{2}, k, n_1-2, \ldots, n_k-2\right)\right\}\right|.$$

As a result of the above procedure, we may acquire a control chart, critical value, and p-value.

Determination of Power and Sample Size

The power function for the HANOM is defined as follows.

$$\begin{aligned}\text{power}(\boldsymbol{\mu}) &= 1-P\bigg(-h^* \leq \frac{\widetilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{\overline{\widetilde{X}}-\mu_.}{S_{[k]}/\sqrt{n_i}} + \frac{\mu_i-\mu_.}{S_{[k]}/\sqrt{n_i}} \leq h^*,\ i=1,\ldots,k\bigg) \\ &= 1-P\left(-h^* \leq \widetilde{T}_i + \frac{\mu_i-\mu_.}{S_{[k]}/\sqrt{n_i}} \leq h^*,\ i=1,\ldots,k\right), \end{aligned} \tag{2.17}$$

where $\widetilde{T}_i, i = 1,\ldots,k$ are the same in Equation (2.15) and the power determined by the particular configuration of the $\mu_i's$. The subspace is assigned the least favorable configuration (LFC) of the means

$$M_\delta = \left\{\boldsymbol{\mu} = (\mu_1,\ldots,\mu_k)' : \max_{i,j}|\mu_i-\mu_j| \geq \delta\right\}.$$

For HANOM, Nelson and Dudewicz (2002) showed that the LFC on M_δ is of the form

$$\boldsymbol{\mu} = (\delta/2, -\delta/2, 0, \ldots, 0)'.$$

Consequently, under the LFC configuration, the power function from Equation (2.17) can be expressed as

$$\begin{aligned}\text{power}(\boldsymbol{\mu}) &= 1-P\left(-h^* \leq \widetilde{T}_i^* \leq h^*,\ i=1,\ldots,k\right) \\ &\geq 1-P\left(\widetilde{T}_{[1]}^* < -h^*\right) - P\left(\widetilde{T}_{[k]}^* > h^*\right) \end{aligned} \tag{2.18}$$

where $\widetilde{T}_i^* = \widetilde{T}_i + \frac{\mu_i-\mu_.}{S_{[k]}/\sqrt{n_i}}, i = 1,\ldots,k$, while $\widetilde{T}_{[1]}^*$ and $\widetilde{T}_{[k]}^*$ are the minimum and the maximum of $\left\{\widetilde{T}_1^*,\ldots,\widetilde{T}_k^*\right\}$, respectively. The power function given by the number of treatments, k, level of significance, α, degrees of freedom, $n_i - 2$ $(i = 1,\ldots,k)$, the largest standard

deviation, $S_{[k]}$, and the effect size δ could be calculated in a the same way as the critical values. For example, we have four treatments to test whether the means are equal. The sample sizes n_i are 8, 12, 10, and 9, respectively, and the largest standard deviation $S_{[k]}$ is 0.5. Under the level of significance $\alpha = 0.05$, and the $\delta = 1$ as the difference between any two treatment means that will lead to rejection of the equality of means hypothesis, the critical value h^* is 2.759 according to Equation (2.16) and the power of HANOM is 0.853 by Equation (2.18). These values can obtained using the *R* Shiny interface. The URL of the interface can accessed at `https://weiming3524.shinyapps.io/hanom_ub/`.

The sample size n_i can be calculated with a given level of significance, α, and a desired power of HANOM, $1-\beta$. The test statistic, critical value, and power of the single-stage sampling procedure for HANOM, on the other hand, are dependent on the sample sizes of each treatment. The required sample size for each treatment cannot be obtained directly, therefore we employ a grid-searching approach as a workaround. This procedure is as follows:

(i) Take an initial sample size $n_{0i} = 3, i = 1, \ldots, k$. We may obtain the critical value for HANOM from Equation (2.16) by a given level of significance, α.

(ii) We may calculate the power value for HANOM from Equation (2.18) under the initial sample size. If the power value meets the intended level, the required sample size is n_{0i}. If not, we raise n_{0i} by 1 until it reaches the desired power.

(iii) Iteratively repeat step (i) and (ii) until the power reaches the desired $1-\beta$. The sample size n_i is currently the smallest size necessary to ensure the given power.

For example, we have four treatments to test whether the means are equal or not. Under $\alpha = 0.05$, $\delta = 1$, $S_{[k]} = 0.5$, and the desired power is 0.9, the smaller sample size n_i is 11

and the corresponding power is 0.913. We also can obtain the values through the interface by R Shiny. In addition, we also provide complete tables (Table F.1-F.3) in Appendix F for users to obtain the required sample size.

2.2.3 Modified Single-Stage Sampling Procedure with Unbalanced Design

Another single-stage sampling procedure developed by Chen (2001) that employee the n_0 observations instead of using the $n_i - 1$ observations from each population to test the equality of means for ANOVA with unknown and unequal variances. We employed this procedure, referred to as $\mathcal{P}_2$, to solve the HANOM with an unbalanced design:

$\mathcal{P}_2$: For X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n_i)$ from each i $(i = 1, \ldots, k)$ population, we suggest taking any randomly chosen n_0 observations to be the minimum of n_i−1 $(i = 1, \ldots, k)$ to calculate the initial sample mean and the sample variance, respectively, by

$$\overline{X}_{i,n_0} = \frac{\sum_{j=1}^{n_0} X_{ij}}{n_0} \quad \text{and} \quad S^2_{i,n_0} = \frac{\sum_{j=1}^{n_0} (X_{ij} - \overline{X}_{i,n_0})^2}{n_0 - 1} \, .$$

The sample size, n_i (≥ 3), of each i $(i = 1, \ldots, k)$ population is split into the n_0 observations and the remaining $n_i - n_0$ observations were used for the calculation of the weighted sample mean.

$$\widehat{X}_i = \sum_{j=1}^{n_i} \omega_{ij} X_{ij} = U_i \sum_{j=1}^{n_0} X_{ij} + V_i \sum_{j=n_0+1}^{n_i} X_{ij}, \tag{2.19}$$

which is a linear combination of the first set of observations with weights U_i and the second

set of observations with weights V_i. The weights of U_i and V_i are as

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{n_i - n_0}{n_0}[n_i z^*/S^2_{i,n_0} - 1]} \tag{2.20}$$

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{\frac{n_0}{n_i - n_0}[n_i z^*/S^2_{i,n_0} - 1]} \tag{2.21}$$

which satisfy the two constraints stated

$$n_0 U_i + (n_i - n_0)V_i = 1$$

$$n_0 U_i^2 + (n_i - n_0)V_i^2 = \frac{z^*}{S^2_{i,n_0}}$$

where z^* is the maximum of $\left\{\frac{S_1^2}{n_1}, \ldots, \frac{S_k^2}{n_k}\right\}$.

The transformed random variables, $T_i^* = (\widehat{X}_i - \mu_i)/\sqrt{z^*}, \quad i = 1, \ldots, k$, from the weighted sample mean, $\widehat{X}_i$ $(i = 1, \ldots, k)$, are distributed independently following the t distribution with degrees of freedom $n_0 - 1$. Wen and Chen (1994) have provided rigorous proof (1994).

HANOM Decision Chart for the One-Way Layout Model

Similar to $\mathcal{P}_1$ procedure, the centerline of the $\mathcal{P}_2$ procedure for HANOM decision chart is $\overline{\widehat{X}} = \sum_{i=1}^{k} \widehat{X}_i/k$, which represents the overall mean of all weighted sample means. Furthermore, $\overline{\widehat{X}} \pm d^* \cdot \sqrt{z^*}$ are the upper and lower decision lines, respectively. $d^* = d^*(\frac{\alpha}{2}, k, n_0 - 1)(> 0)$ is the critical value for HANOM decision chart that is dependent on the level of significance, α, the degrees of freedom, $n_0 - 1$, and the number of treatment, k.

Determination of Critical Values

The rejection region of treatment i given the level of significance, α, as follows:

$$\begin{aligned}\alpha =& P\left(\left|\frac{\widehat{X}_i - \overline{\widehat{X}}}{\sqrt{z^*}}\right| > d^*,\ i = 1, \ldots, k\right) \\ \overset{H_0}{=} & P\left(\left|\widehat{T}_i\right| > d^*,\ i = 1, \ldots, k\right)\end{aligned}$$

where $\widehat{T}_i = T_i^* - \bar{T}^*$ and $T_i^*, \quad i = 1, \ldots, k$ are Student's t distributions with degrees of freedom $n_0 - 1$ and $\bar{T}^*$ is the mean of the T_i^*. We provide a comprehensive explanation in Appendix C. The level of significance, α, satisfy $P\left(\widehat{T}_{[1]} < -d^*\right) + P\left(\widehat{T}_{[k]} > d^*\right) \leq \alpha$, and we choose

$$P\left(\widehat{T}_{[1]} < -d^*\right) = \frac{\alpha}{2} = P\left(\widehat{T}_{[k]} > d^*\right)$$

since $(\widehat{T_1}, \ldots, \widehat{T_k})$ is symmetric by Wu and Chen (1998). As a result, the degrees of freedom of $\widehat{T}_{[1]}$ and $\widehat{T}_{[k]}$ are the same.

Similarly, we need to find out the joint distribution $(\widehat{T}_1, \ldots, \widehat{T}_k)$ as $(\widetilde{T}_1, \ldots, \widetilde{T}_k)$. We also employ Monte-Carlo simulation to generate the approximate sampling distribution of the maximum order statistic of $(\widehat{T}_1, \ldots, \widehat{T}_k)$, namely $\widehat{T}_{[k]}$. We take the $100(1 - \alpha/2)$-th percentile of the approximate sampling distribution of $\widehat{T}_{[k]}$ by 100,000 runs as $\mathcal{P}_1$ procedure, that is, $P(\widehat{T}_{[k]} < d^*) = 1 - \alpha/2$. This procedure allows us to determine the critical value, d^*.

Computing the p-value

Similar to the $\mathcal{P}_1$ procedure, the p-value of the $\mathcal{P}_2$ procedure for HANOM can be determined as the value at which the weighted sample mean is the furthest distant from the overall

weighted sample mean that is

$$1-p=P\left(\max_i\left|\frac{\widehat{X}_i-\overline{\widehat{X}}}{\sqrt{z^*}}\right|\leq d^*\left(\frac{p}{2},k,n_0-1\right)\right)$$

We can be obtain the p-value by solving the equation above and find the minimized p as the p-value of the $\mathcal{P}_2$ for HANOM; that is

$$\min_p\left|\left\{\max_i\left|\frac{\widehat{X}_i-\overline{\widehat{X}}}{\sqrt{z^*}}\right|-d^*\left(\frac{p}{2},k,n_0-1\right)\right\}\right|.$$

As a result of the above procedure, we may also acquire a control chart, critical value, and p-value from the $\mathcal{P}_2$ procedure. To make it easier to utilize the procedures, we offered an interface using R Shiny to acquire both $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM. The URL of the interface can be accessed at `https://weiming3524.shinyapps.io/hanom_ub/`.

R Shiny interface of $\mathcal{P}_2$ and $\mathcal{P}_2$ procedures for HANOM

This interface can execute the single-stage sampling procedure $\mathcal{P}_1$ and $\mathcal{P}_2$ for HANOM with unbalanced design, as shown in Figure K.2 in Appendix K. There are three panels on the interface, including the HANOM chart and the summary statistics by uploading a dataset, the power of the HANOM test after entering the relevant parameters, and the determination of the required sample size.

The first panel is to execute the single-stage sampling procedure $\mathcal{P}_1$ and $\mathcal{P}_2$ for HANOM. The first step is to prepare your data files in comma separated values (CSV) format. The CSV file contains 2 columns with column names. The first column is the values of treatments and the second column is the group of treatments. We also provide a toy example for practical

use. After inputting a data file, the second step is to choose a significant level, 0.05, as the default. Finally, click the Run button and wait a few seconds, the describe statistic, both $\mathcal{P}_1$ and $\mathcal{P}_2$ procedure HANOM chart and the summary statistics will be shown on the display board.

The second panel can obtain the critical value and power after entering a significant level, the number of treatments, sample sizes of each treatment group, the largest standard deviation between treatments, and the difference amount between any two treatment means that will lead to rejection of the null hypothesis. To click the Run button and wait a few seconds, the results will show on the display board. The third panel provides the required sample size for a given level of significance and a desired power of HANOM. After entering the relevant parameters, the minimum sample size required in each treatment and its corresponding power will show on the display board.

2.3 Empirical Type I Error Rate Simulations

In this subsection, we compared the quality of classical tests with the single-stage sampling procedures for balanced and unbalanced design HANOM in terms of type I error rate (α) under different sample size allocations and variance scenarios.

2.3.1 Simulations in Balanced Design

In this simulation study, we compare the type I error rate of ANOVA, Brown-Forsythe test (1974), classical ANOM, and the single-stage sampling procedure for HANOM; the nominal level is $\alpha = 0.05$. The difference of number of treatments, variance situations, and sample size allocations all have an effect on the quality of a test. The empirical type I error

rates for ANOVA and the Brown-Forsythe test (BF) were calculated using the total number of incorrectly rejected null hypotheses. For ANOM and HANOM, the empirical type I error rates were estimated by the number of trials in which at least one of the treatment means that fell outside the upper decision line (UDL) or lower decision line (LDL). The empirical type I error rates are calculated by dividing the number of times the null hypothesis was rejected by the total trials number (5000). To obtain variability, we repeat this procedure 200 simulations. Similar to Hasler's (2016) scenarios, the expected value of all treatments was the same, $\mu_i = 100$ $(i = 1, \ldots, k)$. Five sample size allocations were considered, i.e. $n = 4, 8, 12, 16, 20$. Note that an initial sample of size n (≥ 3) will work in theory, but Bishop and Dudewicz (1978) suggested that 10 or more giving better results. The following three alternative variance scenarios were considered:

- the homoscedastic case (case 1): $\sigma_i = 10, i = 1, \ldots, k$,
- the heteroscedastic case I (case 2): $\sigma_i = 10 + 40(i-1)/(k-1), i = 1, \ldots, k$,
- the heteroscedastic case II (case 3): $\sigma_i = 10 + 60(i-1)/(k-1), i = 1, \ldots, k$.

Table D in Appendices shows the simulation results with respect to empirical type I error rates with different sample size allocations correspond to the different variance scenarios. Under homogeneity of variances, the results reveal that the ANOVA, BF, ANOM and HANOM tests display similar type I error rates. All tests generally show type I error rates close to 5%, but BF and HANOM appear to be a little conservative when the sample size is small. Under heteroscedastic cases, although, the single-stage sampling procedure for HANOM could be conservative under small sample size ($n = 4$), the empirical type I error rates changing between 0.032 and 0.047 were not too far away from 5%. It was observed that

ANOVA, BF and ANOM tests were influenced from heterogeneity of variances; however, this negative effect appeared to be more obvious in ANOM test. In addition, the increase in heteroscedastic case (from case 2 to case 3) further seriously affected the results. It is noted that the type I error rates of ANOVA have slightly improved as the larger sample size. In contrast, the single-stage sampling procedure for HANOM are more robust than the ANOVA, BF and ANOM test in heteroscedastic conditions. Regardless of the change in the number of treatments and sample size allocations, the type I error rates of the procedure for HANOM are still maintained under 5% with standard deviations 0.0024 to 0.0040.

Figure 2.1 provides a visual representation of the simulation results, with sub-figures representing the various variance situations. It is noted that the classical ANOM procedure is available by the package $ANOM$ (Pallmann (2017)) of the statistical software R, but the p-values of the ANOM decision chart need to be calculated from the package $SimComp$ (Hasler (2019)).

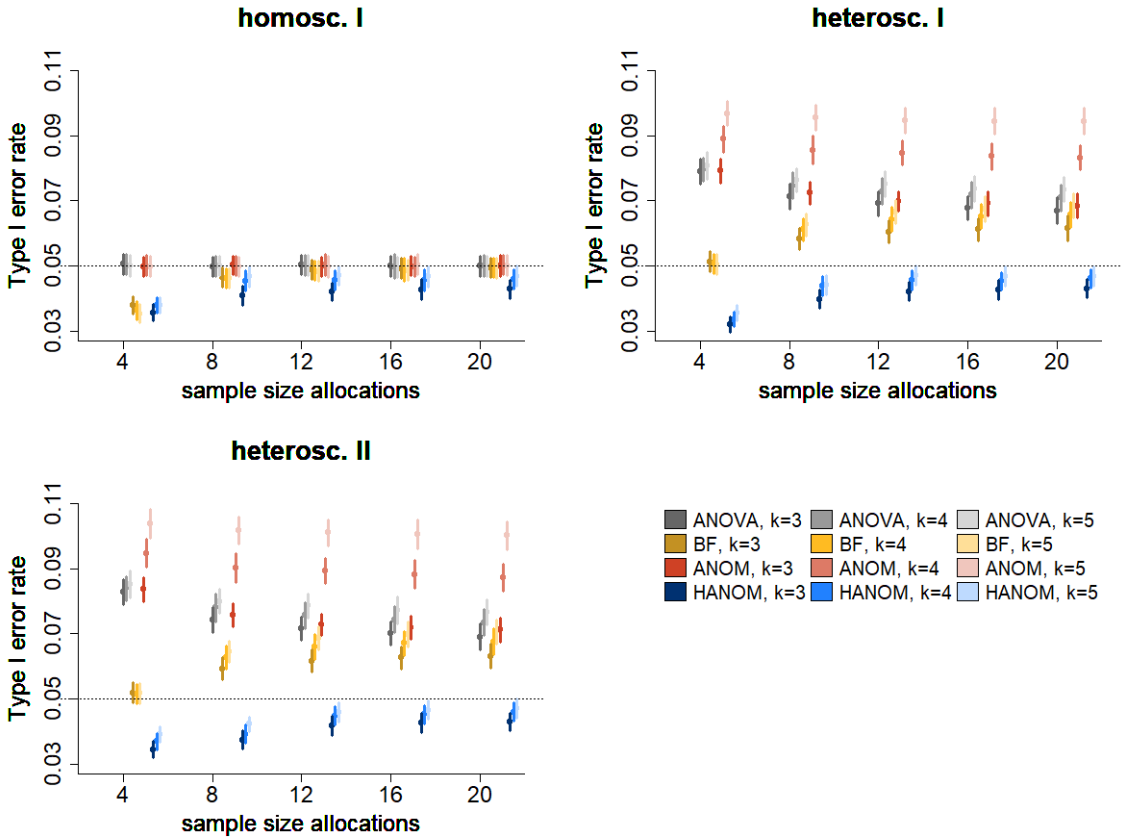


Figure 2.1: Simulated Type I error rates in balanced design under several sample sizes and variance scenarios

2.3.2 Simulations in Unbalanced Design

In the unbalanced design simulation studies, we validated and compared the quality of classical ANOM test, BF test, $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM in terms of type I error rate (α). Based on Mendeş and Yiğit's (2013) scenario setting, we selected $k = 3, 4,$ and 5 treatments under the normal distributions to assess their associated empirical type I error rates. Likewise, the empirical type I error rates are calculated by dividing the number of times we rejected the null hypothesis by the total number of 5000 trials. We repeated this procedure 200 simulations to obtain variability. A number of experimental conditions are detailed in Table 2.1. The equal sample size allocations are n1 to n3, whereas the unequal sample size

allocations are n4 to n9. The direct pairing of sample size and variance represents a larger sample size associated with a larger variance across the unequal sample size conditions n4 to n6. The inverse pairing of sample size and variance represents a larger sample size associated with a smaller variance across the unequal sample size conditions n7 to n9. Furthermore, the scenarios comprised homogeneity (V0) and heteroscedasticity circumstances (V1 to V3).

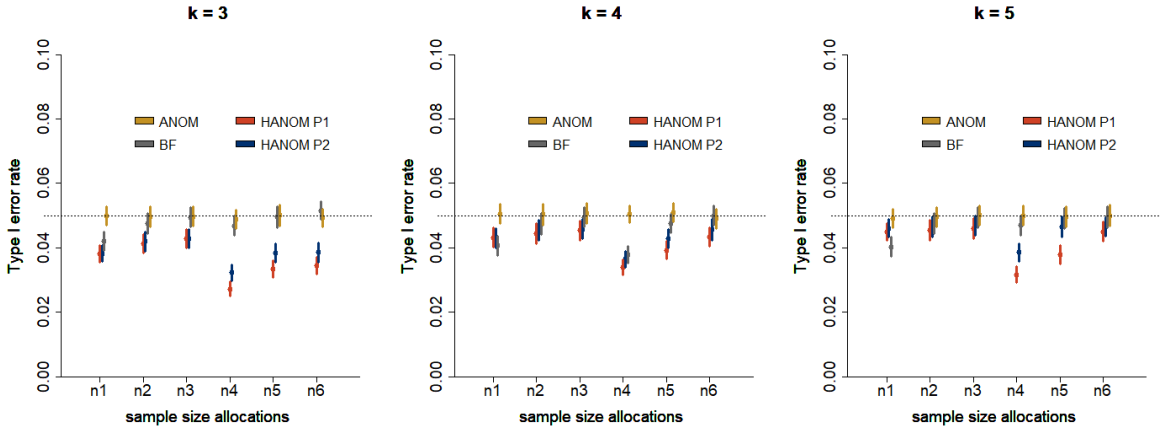
Table 2.1: Experimental conditions in unbalanced design

	Treatment groups number		
	$k = 3$	$k = 4$	$k = 5$
Sample sizes			
n1	5:5:5	5:5:5:5	5:5:5:5:5
n2	10:10:10	10:10:10:10	10:10:10:10:10
n3	20:20:20	20:20:20:20	20:20:20:20:20
n4	3:5:8	3:3:5:5	3:5:7:10:15
n5	5:10:15	5:8:10:15	5:10:15:20:25
n6	5:15:25	10:20:30:40	10:20:30:40:50
n7	8:5:3	5:5:3:3	15:10:7:5:3
n8	15:10:5	15:10:8:5	25:20:15:10:5
n9	25:15:5	40:30:20:10	50:40:30:20:10
Variance ratio			
V0	1:1:1	1:1:1:1	1:1:1:1:1
V1	1:1:4	1:1:1:4	1:1:1:1:4
V2	1:1:10	1:1:1:10	1:1:1:1:10
V3	1:1:20	1:1:1:20	1:1:1:1:20

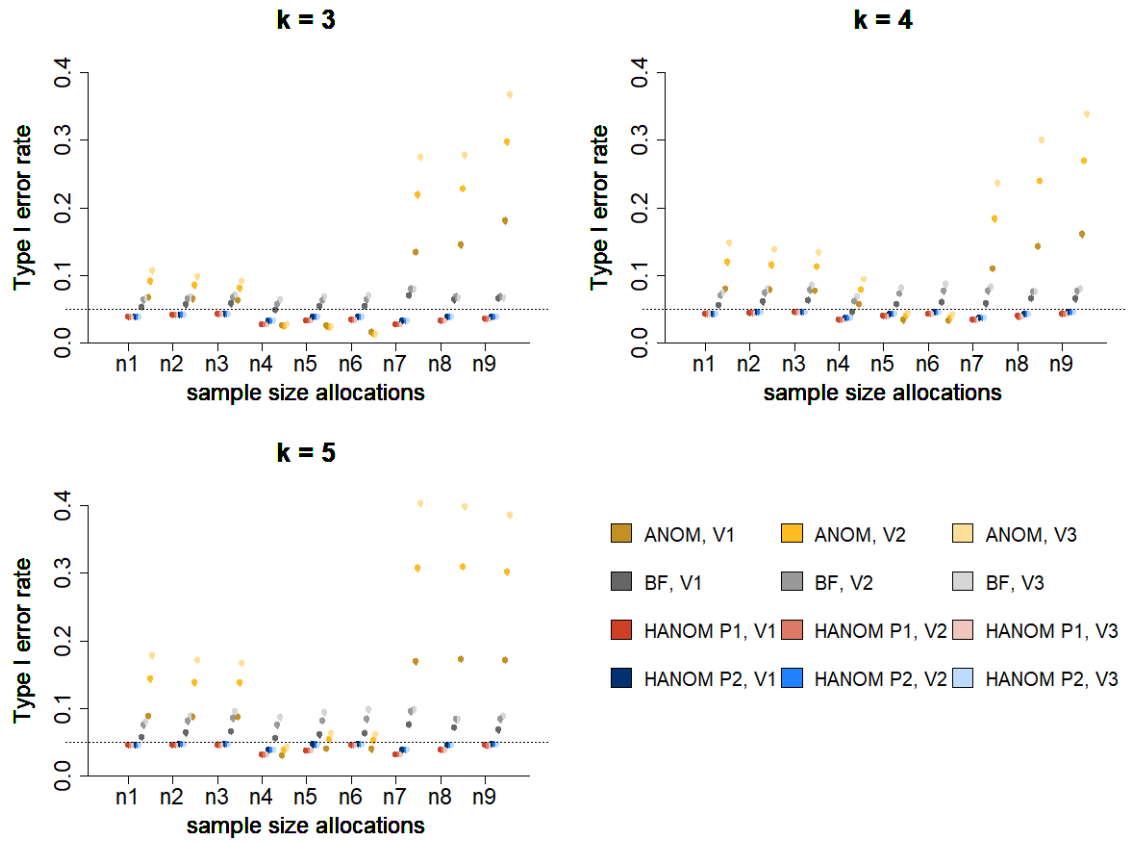
k denotes the number of treatment groups; The equal sample size allocations are n1 to n3, while the unequal sample size allocations are n4 to n7.; V0 represents homogeneity, whereas V1 to V3 represent heterogeneity conditions.

The relative performances of the classical ANOM test, BF test, and $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM in terms of empirical type I error rates corresponding to the different situations are shown in Figure 2.2. The results of homogeneous variances conditions are shown in Figure 2.2(a). The results indicate that ANOM tests performed similarly to the nominal level of 5% and were not significantly affected by unequal sample sizes between groups. Brown-Forsythe test is slightly conservative, $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM are likewise conservative, particularly the small samples size (n4) and less number of group

($k = 3$). Under equal samples sizes (n1 to n3), the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures are equivalent. The $\mathcal{P}_1$ procedure is a little more conservative than the $\mathcal{P}_2$ procedure under unequal sample sizes (n4 to n6).



(a) Type I error rate when variance are homogeneous



(b) Type I error rate when variance are heterogeneous

Figure 2.2: Simulated Type I error rates in unbalanced design with various sample size allocations and variance situations

The results of heterogeneous variances conditions are shown in Figure 2.2(b). The ANOM tests were shown to be more impacted by variance heterogeneity than the HANOM $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures, but BF tests are slightly liberal. Under an equal sample size (n1 to n3), the type I error rates of the ANOM tests surpassed the nominal standard of 5% (liberal). When sample sizes and variances were directly paired (n4 to n6), the type I error rates of the ANOM tests were less than 5%, but substantially above 5% (liberal) when sample sizes and variances were inversely paired (n7 to n9). The BF tests, on the other hand, provided slightly liberal results regardless of sample size allocations and increases in variance ratios (V2 and V3) influenced the results further. Under these conditions, $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM produced more robust results than the ANOM and BF tests, which were identical to the homogeneity settings. There was no significant difference between the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures in all situations. As seen, the type I error rates of the ANOM, BF, and $\mathcal{P}_1$ and $\mathcal{P}_2$ tests exhibited similar patterns as the number of groups increased. The difference was that when the number of treatments increased, the BF test rises over 5%, and the ANOM tests becomes more severe. In certain cases ($k = 5$), ANOM's empirical α can exceed 40%. The $\mathcal{P}_1$ procedure revealed the type I error rates between $0.027 \leq \alpha \leq 0.047$ with standard deviations 0.0023 to 0.0031 and the $\mathcal{P}_2$ procedures were between $0.032 \leq \alpha \leq 0.047$ with standard deviations 0.0025 to 0.0032 in all conditions.

Overall, the ANOM test produced type I error rates of roughly 5% under homogeneity of variances. Regardless of the homogeneous or heterogeneous variances, the BF tests showed slightly liberal. We demonstrated that the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM are robust under heteroscedasticity.

2.4 Numerical Examples

2.4.1 Effects of Different Treatments Data

An experiment reported from Juneau (2003) was carried out to examine the impact of various treatments. A patient received one of four different treatments, namely treatment 1, 2, 3, and 4, and measured the effect. The aim of the experiment is to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

where μ_i denotes the mean effect of treatment i ($i = 1, \ldots, 4$). The data are given in Table G.1 in Appendices.

The normality of the data was checked using the Shapiro-Wilk test, and the homogeneity of variances was tested using Levene's test. The data does not violate the normality assumption, yet the variances are unequal among four treatments. (p-value ≥ 0.379 and 0.031, respectively). Another test for equality of variances is analysis of means for variances (ANOMV), which was described in Wludyka and Nelson (1997). The ANOMV chart is given in Figure 2.3, and we can also conclude that at 0.05 level, the variances are unequal.

When we ignored the effect of unequal variances and used the traditional one-way ANOVA to test the equality of the means of these four treatments. The p-value computed by the one-way ANOVA is 0.038 that is lower than the 0.05 significance level. Hence, We rejected the null hypothesis that the means of all treatments are identical. After rejecting the treatment means are equal, we will try to find a significant difference between pairs of means by a multiple comparison procedure. In the first half of Table 2.2 displays the results

of Scheffé multiple comparison procedure, which revealed no significant difference across all pairwise comparisons, and there is a contradictory result between ANOVA and multiple comparisons for difficulty interpretation.

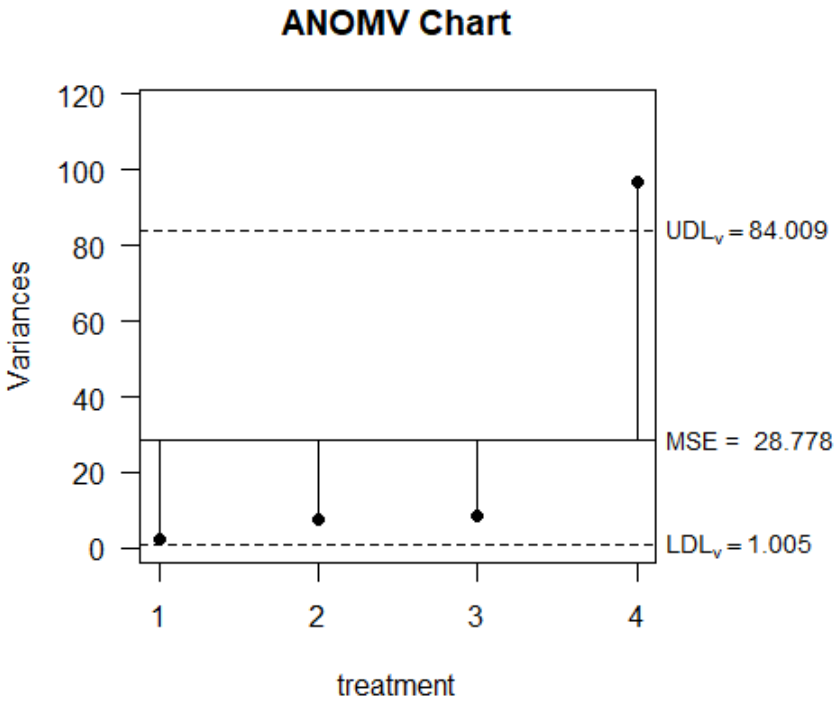


Figure 2.3: The ANOMV Chart for Effect of Treatments Data

In this example, we used Brown and Forsythe's method (1974) to test the equality of means for small samples with heteroscedasticity. The p-value of the Brown-Forsythe test is 0.107 which is larger than the 0.05 significance level. Hence, we do not reject the equality of population means. Therefore, we do not need to do multiple comparison. However, if we do multiple comparisons using the Games-Howell post-hoc test (1976) which does not assume equal variances and sample sizes, the results of Games-Howell method presented a significantly statistical difference between Treatment 1 and Treatment 2, and between Treatment 1 and Treatment 3. The second half of Table 2.2 displays the results of the Games-Howell post-hoc test. Similarly, there is a contradictory result between the Brown-Forsythe test and

the Games-Howell post-hoc test.

Table 2.2: Pairwise comparisons test for Scheffé and Games-Howell

Scheffé Test			
Contrast	Mean Difference	Std. Error	p-value
1 vs 2	-10.753	3.793	0.098
1 vs 3	-11.496	3.793	0.069
1 vs 4	-6.335	3.793	0.456
2 vs 3	-0.742	3.793	0.998
2 vs 4	4.419	3.793	0.772
3 vs 4	5.161	3.793	0.617
Games-Howell Test			
Contrast	Mean Difference	Std. Error	p-value
1 vs 2	-10.753	1.566	**0.005***
1 vs 3	-11.496	1.631	**0.005***
1 vs 4	-6.335	4.981	0.631
2 vs 3	-0.742	1.992	0.981
2 vs 4	4.419	5.111	0.823
3 vs 4	5.161	5.131	0.757

From the analysis above, we find that the results are contradicted between the test equality of means and the pairwise comparison differences. It's not difficult to know their contradictions because the purposes of the tests are different. One is for testing if the means are equal whereas another is for comparing if all pairwise means are different. To overcome this drawback, we adapted the analysis of means, which is a graphic method for comparing means.

We utilized the single-stage sampling procedure for HANOM for this example. We calculated the initial sample mean $\overline{X}_i$ and sample standard deviation S_i with first 3 observations to obtain $\overline{X}_i$ and S_i, respectively. Then, we used weights of U_i and V_i from Equations (2.5) and (2.6) to calculate the weighted sample mean $\widetilde{X}_i$ (Equation (2.2)) for treatment i. Table 2.3 displays the summary statistics.

For Table 2.3, we have an overall weighted sample mean. The centerline of HANOM

Table 2.3: Summary Statistics of Effect of Treatments Data

	Treatment 1	Treatment 2	Treatment 3	Treatment 4
$\overline{X}_i$	99.099	110.799	111.095	110.154
S_i	1.325	3.119	3.536	5.952
U_i	0.882	0.485	0.445	0.250
V_i	-1.647	-0.454	-0.336	0.250
$\widetilde{X}_i$	95.534	111.723	111.178	105.872

decision chart is as follows.

$$\overline{\widetilde{X}} = \frac{95.534 + 111.723 + 111.178 + 105.872}{4} = 106.077$$

The critical value $d^* = d^*(\frac{\alpha}{2} = 0.025, k = 4, df = 2) = 7.367$ obtained from Table A.2 or Appendix K, thus the lower decision line (LDL) and upper decision line (UDL) by Equations (2.8) and (2.9) are as follows:

$$LDL = 106.077 - 7.367 \times \frac{5.952}{\sqrt{4}} = 84.154$$
$$UDL = 106.077 + 7.367 \times \frac{5.952}{\sqrt{4}} = 127.999$$

where $S_{[k]} = \max\{S_1, S_2, S_3, S_4\} = 5.952$. Figure 2.4 gives a HANOM decision chart with the centerline, LDL, and UDL. The HANOM chart revealed that no weighted sample mean fell outside of LDL or UDL which produced a p-value of 0.227. Same as the results of the Brown and Forsythe test that the effect will not be impacted by different treatments.

2.4.2 Reinforcing Bars Strength Data

Weerahandi (2004) conducted an engineer experiment to compare four different types of reinforcing bars. The objective of the experiment is to compare the mean strength of rein-

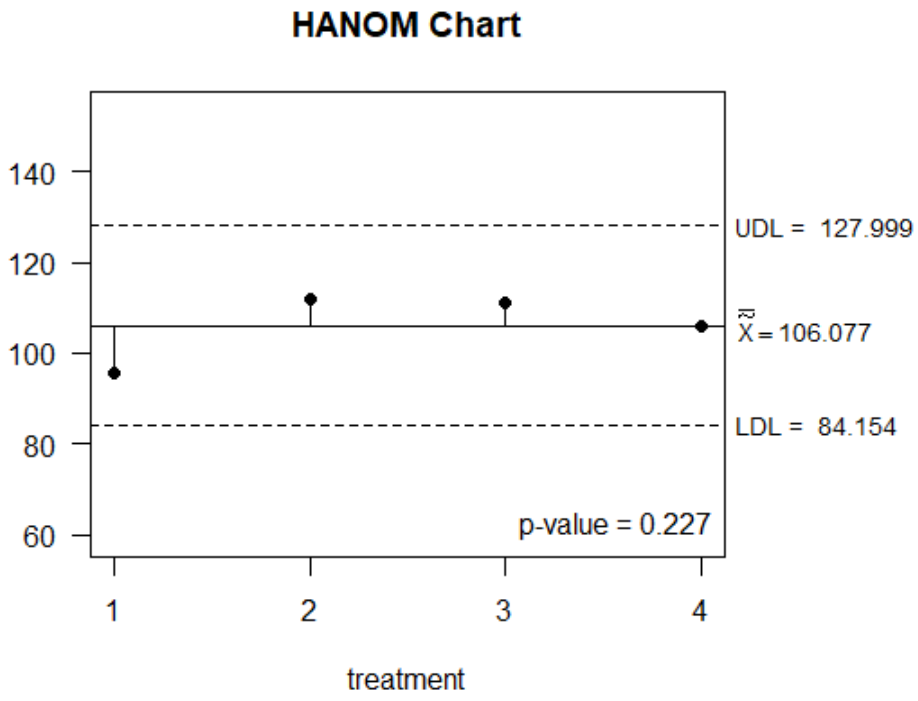


Figure 2.4: The HANOM Chart for Effect of Treatments Data

forcing bars, and the dataset is provided in Table (G.2) in Appendices. The normality of the data was checked using the Shapiro-Wilk test, and the homogeneity of variances was tested using Levene's test. The data does not violate the normality assumption, yet the variances are unequal among four brands (p-value $\geq$ 0.121 and 0.001, respectively).

The traditional Brown-Forsythe test is used to test equality of means in the case for small samples with heteroscedasticity. The Brown-Forsythe test for the data gives a p-value of 0.232, indicating that we do not reject the equality of the population means. After global test, we do multiple comparisons using the Games-Howell post-hoc test, Table 2.4 displays the results that there is a statistically significant difference between group 3 and group 4 (p-value=0.016). There is a contradictory result between the Brown-Forsythe test and the Games-Howell post-hoc test, making it difficult to interpret. Additionally, in this example, we employed the generalized fiducial test and the parametric bootstrap test, both of which are

generalized p-value based tests developed by Weerahandi and Krishnamoorthy (2019). The Fiducial test and Parametric Bootstrap have p-values of 0.035 and 0.026, respectively, which are both less than the 0.05 significance level, therefore we reject the null hypothesis. Afterwards, we do the post-hoc pairwise comparison that are available by the statistical software R package $onewaytests$ (Dag et al., 2021). There was no significant difference between the pairwise comparisons in the Table 2.4. The contradicting result of the global mean test and multiple comparisons is once again difficult to explain.

Table 2.4: The results of the multiple comparisons test to the strength of reinforcing bars data

	p-value		
Contrast	Games-Howell test	Fiducial test	Parametric Bootstrap test
1 vs 2	0.663	0.210	0.154
1 vs 3	0.999	0.211	0.155
1 vs 4	0.225	0.212	0.155
2 vs 3	0.618	0.211	0.154
2 vs 4	0.999	0.212	0.154
3 vs 4	**0.016***	0.209	0.154

Alternatively, we applied the HANOM to this example. The HANOM gives global as well as local inference, which identifies which individual treatment means depart from the grand mean. The HANOM not only gives information on a global information but also makes it possible to draw local inferences about which particular treatment means are different from the overall mean (Pallmann and Hothorn (2016)). For the single-stage sampling procedure $\mathcal{P}_1$ for HANOM, we calculated $\overline{X}_i$ and S_i with the first $n_i - 1$ observations for $i = 1, \ldots, 4$. Next, we used weights of U_i and V_i from Equations (2.5) and (2.6) to calculate the weighted sample mean $\widetilde{X}_i$ (Equation (2.2)) for brand i. Table 2.5 displays the summary statistics.

The upper and the lower decision limits are $\overline{\widetilde{X}} \pm h^* \cdot \frac{S_{[k]}}{\sqrt{n_i}}$, respectively. An overall weighted sample mean, that is the centerline of HANOM decision chart, is as follows, ac-

cording to the results in Table 2.5:

$$\overline{\widetilde{X}} = \frac{17.785 + 20.688 + 18.511 + 20.407}{4} = 19.348$$

We can obtain the larger critical value $h^* = 3.001$ (the standard error is 0.017), where $h^* = h^*(\frac{\alpha}{2} = 0.025, k = 4, df_1 = 5, df_2 = 6, df_3 = 5, df_4 = 7)$ at the 5% level from Appendix K, and $S_{[k]} = \max\{S_1, S_2, S_3, S_4\} = 5.745$. Thus, the upper decision lines (UDLs) and lower decision lines (LDLs) by Equations (2.12) and (2.13) as follows, and the results are shown in Table 2.5 and Figure 2.5(a).

$$UDL_1 = 19.348 + 3.001 \times \frac{5.745}{\sqrt{7}} = 25.864$$
$$LDL_1 = 19.348 - 3.001 \times \frac{5.745}{\sqrt{7}} = 12.832$$
$$UDL_2 = 19.348 + 3.001 \times \frac{5.745}{\sqrt{8}} = 25.443$$
$$LDL_2 = 19.348 - 3.001 \times \frac{5.745}{\sqrt{8}} = 13.253$$
$$UDL_3 = 19.348 + 3.001 \times \frac{5.745}{\sqrt{7}} = 25.864$$
$$LDL_3 = 19.348 - 3.001 \times \frac{5.745}{\sqrt{7}} = 12.832$$
$$UDL_4 = 19.348 + 3.001 \times \frac{5.745}{\sqrt{9}} = 25.094$$
$$LDL_4 = 19.348 - 3.001 \times \frac{5.745}{\sqrt{9}} = 13.601$$

The HANOM chart (Figure 2.5(a)) revealed that no weighted sample mean fell outside of LDLs or UDLs that indicates the reinforcing bars strength would not be impacted by different brands.

Table 2.5: Summary statistics of the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures to the strength of reinforcing bars data

$\mathcal{P}_1$	Brand 1	Brand 2	Brand 3	Brand 4
n_i	7	8	7	9
$\overline{X}_i$	17.900	20.014	18.150	20.788
S_i	3.625	5.745	1.369	1.672
U_i	0.215	0.125	0.339	0.240
V_i	-0.287	0.125	-1.033	-0.922
$\widetilde{X}_i$	17.785	20.688	18.511	20.407
UDL	25.864	25.443	25.864	25.094
LDL	12.832	13.253	12.832	13.601
$\mathcal{P}_2$	Brand 1	Brand 2	Brand 3	Brand 4
n_0	6	6	6	6
$\overline{X}_{i,n_0}$	17.900	20.650	18.150	20.300
S_{i,n_0}	3.625	6.018	1.639	1.430
U_i	0.212	0.125	0.335	0.453
V_i	-0.273	0.125	-1.007	-0.573
$\widehat{X}_i$	17.791	20.688	18.502	17.552
UDL	25.034	25.034	25.034	25.034
LDL	11.962	11.962	11.962	11.962

This example was also subjected to the modified single-stage sampling procedure $\mathcal{P}_2$. For $\mathcal{P}_2$ procedure, $n_0 = 6$ is the minimum of $n_i - 1$ in which the sample size of the example is $n_1 = 7, n_2 = 8, n_3 = 7$, and $n_4 = 9$, respectively. We calculated the initial $\overline{X}_{i,n_0}$ and S_{i,n_0} with the first 6 observations for brand $i = 1, \ldots, 4$, respectively. Next, we used weights of U_i and V_i from Equations (2.20) and (2.21) to calculate the weighted sample mean $\widehat{X}_i$ (Equation (2.19)) for brand i. Table 2.5 displays the summary statistics.

The upper and the lower decision lines are $\overline{\widehat{X}} \pm d^* \cdot \sqrt{z^*}$, respectively, An overall weighted sample mean, that is the $\mathcal{P}_2$ procedure centerline of HANOM decision chart, is as follows, according to the results in Table 2.5:

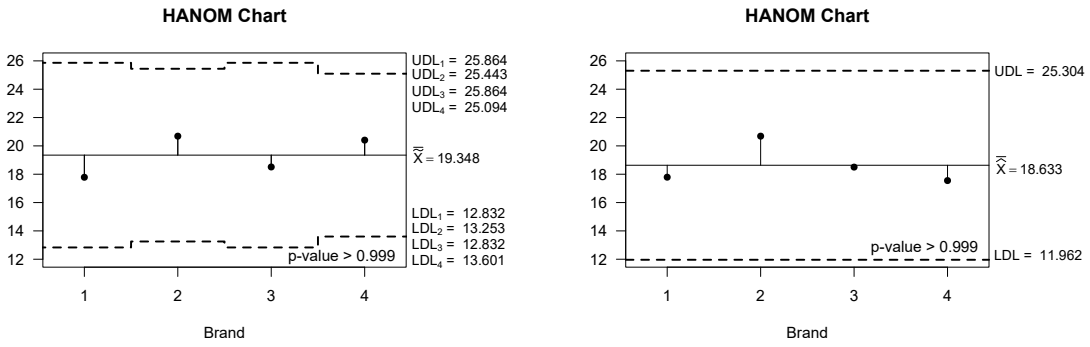
$$\overline{\widehat{X}} = \frac{17.791 + 20.688 + 18.502 + 17.552}{4} = 18.633$$

We can obtain the critical value $d^* = d^*(\frac{\alpha}{2} = 0.025, k = 4, df = 5) = 3.135$ (the standard error is 0.015) at the 5% level from Appendix K, and $z^* = \max\left\{\frac{3.625^2}{7}, \frac{6.018^2}{8}, \frac{1.639^2}{7}, \frac{1.430^2}{9}\right\} = 4.527$. The upper decision limits (UDL) and lower decision line (LDL) are as follows:

$$UDL = 18.633 + 3.135 \times \sqrt{4.527} = 25.304$$

$$LDL = 18.633 - 3.135 \times \sqrt{4.527} = 11.962$$

The HANOM chart (Figure 2.5(b)) revealed that no weighted sample mean fell outside of LDL or UDL that indicates the reinforcing bars strength would not be impacted by different brands. This is the same conclusion as the single-stage sampling procedure $\mathcal{P}_1$. $\mathcal{P}_1$ and $\mathcal{P}_2$ both procedures produce a p-value larger than 0.999. The relatively small sample sizes are the cause for this conservative results. A large and non-significant p-value also produced by the ANOM test to this numerical example. Base on the theory by Bishop and Dudewicz (1978), an initial sample size of greater than or equal to 3 is sufficient, but 10 or more would yield superior results.



(a) HANOM chart by procedure $\mathcal{P}_1$

(b) HANOM chart by procedure $\mathcal{P}_2$

Figure 2.5: HANOM chart by the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for the reinforcing bars strength data

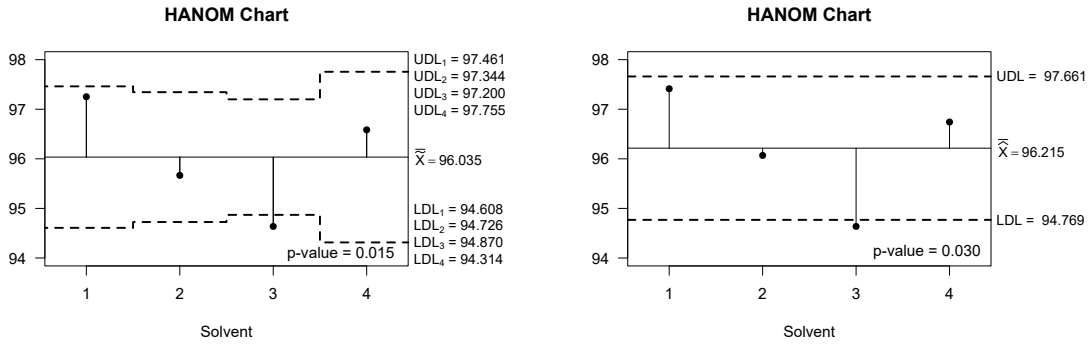
2.4.3 Bacterial Killing Ability Data

Bishop and Dudewicz (1978) described an experiment to test the effect of solvent on the ability of the fungicide methyl-2-benzimidazole-carbamate to destroy the fungus Penicillium expansum (see Table G.3) in Appendices). The aim of this experiment was to test the hypothesis that the mean percentages of the destroyed fungus are all equal. The data does not violate the normality assumption, yet the variances are unequal among four solvents (p-value $\geq$ 0.077 and 0.005, respectively).

Table 2.6: Summary statistics of the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures to the bacterial killing ability data

$\mathcal{P}_1$	Solvent 1	Solvent 2	Solvent 3	Solvent 4
n_i	16	19	24	11
$\overline{X}_i$	97.103	95.614	94.566	97.160
S_i	1.482	2.109	2.301	0.878
U_i	0.082	0.058	0.042	0.161
V_i	-0.225	-0.045	0.042	-0.606
$\widetilde{X}_i$	97.252	95.666	94.637	96.584
UDL	97.461	97.344	97.200	97.755
LDL	94.608	94.726	94.870	94.314
$\mathcal{P}_2$	Solvent 1	Solvent 2	Solvent 3	Solvent 4
n_0	10	10	10	10
$\overline{X}_{i,n_0}$	97.281	96.196	94.590	97.160
S_{i,n_0}	1.396	1.929	2.723	0.878
U_i	0.123	0.091	0.042	0.144
V_i	-0.038	0.010	0.042	-0.440
$\widehat{X}_i$	97.413	96.068	94.637	96.742
UDL	97.661	97.661	97.661	97.661
LDL	94.769	94.769	94.769	94.769

Here, we used the single-stage sampling procedure $\mathcal{P}_1$ for the HANOM to this example. An overall weighted sample mean, $\overline{\widetilde{X}}$, is equal to 96.035 according to the results in Table 2.6. We can obtain the critical value $h^* = h^*(\frac{\alpha}{2} = 0.025, k = 4, df_1 = 15, df_2 = 18, df_3 = 23, df_4 = 10) = 2.480$ (the standard error is 0.009) at the 5% level from Appendix K and $S_{[k]} = \max\{S_1, S_2, S_3, S_4\} = 2.301$, the results of LDL and UDL are shown in Table 2.6.



(a) HANOM chart by procedure $\mathcal{P}_1$ (b) HANOM chart by procedure $\mathcal{P}_2$

Figure 2.6: HANOM chart by the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for the bacterial killing ability data

We conducted a HANOM decision chart as shown in Figure 2.6(a). There is one weighted sample mean falling outside of the LDL_3 based on the HANOM chart (p-value=0.015), which indicates the Solvent 3 destroyed significantly more fungus than others.

For this example, we also used the modified single-stage sampling procedure $\mathcal{P}_2$ for HANOM. The overall weighted sample mean $\overline{\widehat{X}}$ is equal to 96.215, n_0 is equal to 10, and the critical value $d^* = d^*(\frac{\alpha}{2} = 0.025, k = 4, df = 9) = 2.601$ (the standard error is 0.007). $z^* = \max\left\{\frac{1.396^2}{16}, \frac{1.929^2}{19}, \frac{2.723^2}{24}, \frac{0.878^2}{11}\right\} = 0.309$; therefore, the LDL is equal to 94.769 and the UDL is 97.661. We found that the solvent 3 destroyed significantly more fungus than others based on the HANOM chart (p-value=0.030, Figure 2.6(b)). This conclusion is the same as the $\mathcal{P}_1$ procedure.

Chapter 3

Multiple Comparisons with a Control under Heteroscedasticity

Let observations X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n_i)$ from population $\pi_i, i = 1, \ldots, k$ are normally distributed with unknown mean μ_i and unknown variance $\sigma_i^2, i = 1, \ldots, k$. Assume, without loss of generality, that π_1 is the control population with mean of μ_1 and variance of σ_1^2. Our goal is to establish a simultaneous confidence interval on the $k-1$ paired differences $\mu_i - \mu_1, i = 2, \ldots, k$. Such a question, Paulson (1952) and Dunnett (1955) first considered the situation with unknown but equal variances. Miller (1981) provided an overview of these multiple-comparison procedures.

Nevertheless, it is not always reasonable to assume homogeneous variances between treatments, which may lead to wrong conclusions. For example, because the variation dependent on the dosage effects, dose finding studies may suffer from heteroscedasticity. (for example, see the data in Westfall (1997) that are available in the *multcomp* package of *R*.) Tamhane (1977) offered approximation methods for multiple comparisons with all-pairwise and a control comparison under unequal variances. Games and Howell (1976) proposed a approximate simultaneous confidence interval for all-pairwise comparisons under heteroscedasticity. Hasler and Hothorn (2008) proposed a plug-in procedure that provided approximated simultaneous confidence intervals for all-contrast tests. Li and Ning (2012)

developed three approximate simultaneous confidence interval approaches for multiple comparisons with a control based on the Bonferroni approximation, Slepian's inequality, and multivariate t distribution.

Dudewicz and Dalal (1983) proposed a two-stage procedure based on Stein (1945) for comparing all μ_i's with μ_1 in order to find exact distributions of the test statistics under unknown and unequal variances σ_i^2 $(i = 1, \ldots, k)$. However, a two-stage procedure is design-oriented and may not always be applicable to real-world circumstances. We applied a single-stage sampling procedure and a modified single-stage sampling procedure for multiple comparison with a control (MCC) in this section. As mentioned above, the single-stage sampling method has been applied in many studies. The single-stage sampling procedure for MCC was applied to equal sample sizes by Wen and Chen (1994). Here, we extended it to the unequal sample sizes. In Subsection 3.1, a classical two-stage sampling procedure to MCC is introduced. In Subsection 3.2, consideration is given to a single-stage sampling approach and a modified single-stage sampling procedure with one-sided and two-sided intervals, respectively, for MCC. In Subsection 3.3, we simulated the familywise error rates (FWE) to validate the quality of the procedures, and the last subsection is a numerical example for demonstrating to show how to apply these procedures.

3.1 Two-Stage Sampling Procedure for MCC

In this Subsection, we will introduce a classical two-stage sampling procedure for MCC that proposed by Dudewicz and Dalal (1983). The two-stage sampling procedure for MCC is a design-oriented method whose purpose is to obtain simultaneous confidence intervals with a overall confidence coefficient $1 - \alpha$.

Let X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n_i)$ represent an independent random sample drawn from the normal population π_i with an unknown mean μ_i and unknown and unequal variance σ_i^2 $(i = 1, \ldots, k)$. Assume that the purpose is to obtain simultaneous confidence intervals on

$$\mu_i - \mu_1, \quad 2 \leq i \leq k$$

with a overall confidence coefficient $1 - \alpha$. The classical two-stage sampling procedure is detailed as follows:

(i) Take an initial sample n_0 (≥ 2) from each i $(i = 1, \ldots, k)$ population, and calculate the sample means $\overline{X}_{i,n_{0i}}$ and sample variances $S^2_{i,n_{0i}}$.

(ii) Calculate

$$n_i = \max\left\{n_{0i} + 1, [\![\frac{S^2_{i,n_{0i}}}{d^2}]\!] + 1\right\} \tag{3.1}$$

for i $(i = 1, \ldots, k)$ population, where $[\![x]\!]$ denoted the greatest integer in x. Let $d = d(n_{0i}, k, l, 1 - \alpha)$ be the solution of the equation

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{n_{0i}}\left(t_1 + \frac{l}{d}\right)\right] f_{n_{0i}}(t_1)\, dt_1 = 1 - \alpha \tag{3.2}$$

where $F_v(\cdot)$ and $f_v(\cdot)$ are the distribution and density functions respectively of a student's t random variables with $v - 1$ $(= n_{0i} - 1)$ degrees of freedom. Details of the use of (3.2) to obtain independent t-statistics follow as in Theorem (4.1) of Dudewicz and Dalal (1975) and Theorem (1) of Dudewicz and Dalal (1983).

(iii) Take the remaining $n_i - n_{0i}$ $(i = 1, \ldots, k)$ observation to calculate the sample means

$$\overline{X}_{i,n_i-n_{0i}} = \frac{X_{i,n0i+1} + \ldots + X_{i,n_i}}{n_i - n_{0i}}$$

(iv) For each i $(i = 1, \ldots, k)$ population, compute the weights

$$c_i = \frac{n_{0i}}{n_i}\left[1 + \sqrt{1 - \frac{n_i}{n_{0i}}\left(1 - \frac{n_i - n_{0i}}{(S_{i,n_{0i}}/d)^2}\right)}\right]$$

and the weighted sample means

$$\widetilde{\overline{X}}_i = c_i\overline{X}_{i,n_{0i}} + (1 - c_i)\overline{X}_{i,n_i-n_{0i}}$$

(v) Calculate the one-sided confidence intervals:

$$\left(\widetilde{\overline{X}}_i - \widetilde{\overline{X}}_1\right) - l \leq \mu_i - \mu_1$$

or

$$\mu_i - \mu_1 \leq \left(\widetilde{\overline{X}}_i - \widetilde{\overline{X}}_1\right) + l$$

The choice of l value in an actual application should be determined by real-world issues. For the special case $n_{0i} = n_0$, Dudewicz et al. (1975), and Dudewicz and Dalal (1983) provided the critical values tables of l/d for Equations (3.2), and Wen and Chen (1994) supplied a FORTRAN program for the critical values. Under unequal sample sizes, we now can obtain

the critical value by the user-friendly interface by R Shiny. The URL of the interface can be accessed at `https://weiming3524.shinyapps.io/mc-control/`.

R Shiny interface for multiple comparison with a control

The interface simulates the critical values of the single-stage sampling procedure for multiple comparison with a control under heteroscedasticity. As shown in Figure K.3 in Appendix K, the interface provides the one-sided and two-sided critical values after inputting parameters.

To execute the single-stage sampling procedure for multiple comparison with a control under heteroscedasticity, the first step is to choose a significant level, 0.05, as default. The second step is to enter the size of control and treatment groups. After entering the number of treatment groups, you can enter the size in each treatment group respectively. The size in each treatment group can be same or different. Finally, click the Run button and wait for a few seconds, the one-sided and two-sided critical values will be shown on the display board.

3.2 Single-Stage Sampling Procedures for MCC

The single-stage sampling procedure for MCC is the same as that considered in Section 2.2. In the following subsection, two single-stage sampling procedures for simultaneous one-sided and two-sided intervals I for $k-1$ paired differences of means with the control mean $\mu_i - \mu_1$ are considered to be

$$\inf_{\Omega} P(\mu_i - \mu_1 \in I, i = 2, \ldots, k) \geq P^* \tag{3.3}$$

where $P^* \in (\frac{1}{k-1}, 1)$ and $\Omega = \left\{ \underset{\sim}{\mu} = (\mu_1, \ldots, \mu_k)', \underset{\sim}{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)' \right\}$.

3.2.1 Single-Stage Sampling Procedure for MCC

Among each population π_i $(i = 1, \ldots, k)$, we compute the weighted sample mean, $\widehat{X}_i$, using Equation (2.2) by selecting the first $n_i - 1$ observations and the remaining observation to calculate the weighted sample mean. The weights of U_i and V_i of the n_i observations satisfy $(n_i - 1)U_i + V_i = 1$ and $(n_i - 1)U_i^2 + V_i^2 = \frac{z^*}{S_i^2}$, where z^* is the maximum of $\left\{\frac{S_1^2}{n_1}, \ldots, \frac{S_k^2}{n_k}\right\}$. Thus, we can derive the weights of U_i and V_i as

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{1}{n_i - 1}[n_i z^*/S_i^2 - 1]}, \tag{3.4}$$

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{(n_i - 1)[n_i z^*/S_i^2 - 1]}. \tag{3.5}$$

The sampling distribution of the weighted sample mean, $\widehat{X}_i$, given the sample variance, S_i^2 $(i = 1, \ldots, k)$, is a conditional normal distribution with mean zero and variance σ_i^2/S_i^2 $(i = 1, \ldots, k)$, and is also an unconditional Student's t distribution with degrees of freedom $n_i - 2$ (Wen and Chen (1994); Chen and Chen (1998)). That is,

$$T_i = \frac{\widehat{X}_i - \mu_i}{\sqrt{z^*}} \sim t(n_i - 2), \quad i = 1, \ldots, k. \tag{3.6}$$

T_i $(i = 1, \ldots, k)$ is not affected by the unknown population variance σ_i^2 $(i = 1, \ldots, k)$, and Wen and Chen (1994) have demonstrated that $T_1, \ldots, T_k$ are mutually independent.

One-Sided Confidence interval

In this subsection, simultaneous lower and upper confidence intervals for $\mu_i - \mu_1, i =$

$2, \ldots, k$ are considered. Let

$$I_L = (\widehat{X}_i - \widehat{X}_1 - a_i\sqrt{z^*}, \infty) \tag{3.7}$$

be a lower confidence intervals for $\mu_i - \mu_1$, where a_i's are critical values selected to satisfy the requirement of probability (3.3). The constants a_i's are the solution of

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + a_i)\right] f_{v_1}(t_1)\, dt_1 = P^*, \tag{3.8}$$

where $F_{v_i}(\cdot)$ and $f_{v_i}(\cdot)$ are the distribution and density functions respectively of a student's t random variables with $v_i(= n_i - 2), i = 1, \ldots, k$, degrees of freedom. The details proof for a_i's solution is in Appendix H.

Likewise, the upper confidence interval for $\mu_i - \mu_1, i = 2, \ldots, k$ is

$$I_U = (-\infty, \widehat{X}_i - \widehat{X}_1 + b_i\sqrt{z^*}).$$

The critical values b_i's are the solution of

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[1 - F_{v_i}(t_1 - b_i)\right] f_{v_1}(t_1)\, dt_1 = P^*. \tag{3.9}$$

The details proof for b_i's solution is in Appendix H.

Two-Sided Confidence interval

Let

$$I_s = (\widehat{X}_i - \widehat{X}_1 - a_i\sqrt{z^*}, \widetilde{X}_i - \widetilde{X}_1 + b_i\sqrt{z^*})$$

be a two-sided confidence intervals for $\mu_i - \mu_1, i = 2, \ldots, k$, where the critical values a_i's and b_i's are chosen to satisfy the requirement of probability (3.3). The constants a_i's and b_i's are the solution of

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + a_i) - F_{v_i}(t_1 - b_i)\right] f_{v_1}(t_1)\, dt_1 = P^*. \tag{3.10}$$

The details proof for a_i's and b_i's solution is in Appendix H.

Optimal Confidence Intervals

Critical values a_i and b_i $(i = 1, \ldots, k)$ would be chosen differently in practice depending on the purposes. In a normal model with known or unknown variance, the smallest length of confidence interval occurs for the critical values $a = -b$ according Guenther (1969) and Casella and Berger (2002). In other words, if the distribution is symmetric, the shortest confidence interval is equal-tails. The optimal solution for the maximized coverage probability under equal sample sizes have been shown by Wen and Chen (1994) that is when all critical values of t for the Equation (3.10) are the same. As a result, the optimal confidence interval is the shortest based on the symmetrical normal and Student t distributions.

Wen and Chen (1994) have demonstrated that $T_1, \ldots, T_k$ are mutually independent under unequal sample sizes. According to Casella and Berger (2002), each confidence interval has the smallest predicted length of t distribution with the same critical values. As a result, the

Equations (3.8), (3.9) and (3.10) may be rewritten as follows.

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + a)\right] f_{v_1}(t_1)\, dt_1 = P^*, \tag{3.11}$$

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[1 - F_{v_i}(t_1 - b)\right] f_{v_1}(t_1)\, dt_1 = P^*, \tag{3.12}$$

and

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + h) - F_{v_i}(t_1 - h)\right] f_{v_1}(t_1)\, dt_1 = P^*. \tag{3.13}$$

We may consider that the critical values a and b for Equations (3.11) and (3.12) are the same. If sample sizes are equal, the degrees of freedom are the same, $v_1 = \ldots = v_k = v$. Equations (3.11) and (3.13) become

$$\int_{-\infty}^{\infty} \left[F_v(t + a)\right]^{k-1} f_v(t)\, dt = P^*, \tag{3.14}$$

and

$$\int_{-\infty}^{\infty} \left[F_v(t + h) - F_v(t - h)\right]^{k-1} f_v(t)\, dt = P^*. \tag{3.15}$$

The critical values tables of a for Equations (3.14) or h for Equations (3.15) can be found in Dudewicz et al. (1975), and Dudewicz and Dalal (1983). Wen and Chen (1994) supplied a FORTRAN program for the critical values. Under unequal sample sizes, it is impossible to list critical values table for all cases. To obtain the critical values easily from Equations (3.11) and (3.13), we created an interface by using R Shiny. After entering parameters settings,

we can get the one-sided and two-sided critical values from the interface. The URL of the interface can be accessed at `https://weiming3524.shinyapps.io/mc-control/`.

3.2.2 Modified Single-Stage Sampling Procedure for MCC

In this subsection, we propose a modified single-stage sampling procedure for MCC. The weights of U_i and V_i of the n_i observations satisfy two conditions (2.3) and (2.4) and the solutions of weights of U_i and V_i are Equations (2.5) and (2.6). Also, given the sample variance, S_i^2 ($i = 1, \ldots, k$), the sampling distribution of the weighted sample mean, $\widetilde{X}_i$, is an unconditional Student's t distribution with degrees of freedom $n_i - 2$. That is,

$$T_i = \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n_i}} \sim t(n_i - 2), \quad i = 1, \ldots, k.$$

The modified single-stage sampling procedure to the one-sided and two-sided simultaneous confidence intervals for $\mu_i - \mu_1, i = 2, \ldots, k$ are considered as

$$\begin{aligned} I_L &= (\widetilde{X}_i - \widetilde{X}_1 - c^* \frac{S_{[k]}}{\sqrt{n_i}}, \infty) \\ I_U &= (-\infty, \widetilde{X}_i - \widetilde{X}_1 + d^* \frac{S_{[k]}}{\sqrt{n_i}}) \end{aligned} \tag{3.16}$$

and

$$I_s = (\widetilde{X}_i - \widetilde{X}_1 - h^* \frac{S_{[k]}}{\sqrt{n_i}}, \widetilde{X}_i - \widetilde{X}_1 + h^* \frac{S_{[k]}}{\sqrt{n_i}})$$

The critical value c^* in Equation (3.16) needs to satisfy

$$\int_{-\infty}^{\infty} \left[\prod_{i=2}^{k} F_{v_i}(t_1 \sqrt{\frac{n_i}{n_1}} + c^*) \right] f_{v_1}(t_1)\, dt_1 = 1 - \alpha. \tag{3.17}$$

The details proof of Equation (3.17) is given in Appendix I.

Determination of Sample Size

The determination of the total sample size for MCC problem depends on the definition of power, such as any-pair power, all-pairs power and per-pair power. The any-pair power is defined as the probability of rejecting at least one false hypothesis, while the per-pair power is the probability of rejecting a particular false hypothesis, and the all-pairs power is the probability of rejecting all false hypotheses. Among the three aforementioned choices, any-pair power yields the smallest sample size, while all-pairs power yields the largest one. Horn and Dunnett (2004) discussed and compared the differences in the total sample size requirements of these three power definitions in a one-sided test by assuming equal allocation.

Similar to the all-pairs power of Hsu (1996), we use an any-pair power which can be considered as the minimum required sample size to meet a pre-specified significance and power level for the modified single-stage sampling procedure for MCC. The power function for the modified single-stage sampling procedure for MCC is as follows.

$$\begin{aligned}
&P\left\{\mu_i - \mu_1 \notin I_L \text{ and } \exists i, i = 2, \ldots, k, \mu_i - \mu_1 > \delta\right\} \\
=&1 - P\left(T_i < T_1 \sqrt{\frac{n_i}{n_1}} + \frac{\delta}{S_{[k]}/\sqrt{n_i}}, i = 2, \ldots, k \right) \\
=&1 - \int_{-\infty}^{\infty} \left(\prod_{i=2}^{k} \left[F_{v_i}(t_1 \sqrt{\frac{n_i}{n_1}} + \frac{\delta}{S_{[k]}/\sqrt{n_i}}) \right] \right) f_{v_1}(t_1)\, dt_1 \geq 1 - \beta
\end{aligned}$$

The sample size n_i can be determined for a given level of significance, α, and a required power of MCC, $1-\beta$. Because we cannot directly obtained the sample size for each treatment, we use a grid-searching approach as a workaround. This procedure is as follows:

(i) Consider an initial sample size $n_{01} = 3$ and $n_{0i} = 3, i = 2, \ldots, k$. A power value for MCC may be determined from the aforementioned power equation by a given largest standard deviation $S_{[k]}$ between treatments, and a difference amount δ between any two comparisons that will lead to rejection of the null hypothesis.

(ii) If the $1-\beta$ and α for MCC under the initial sample size meet the desired power and the level of significance, α, the required sample size is n_{0i}. If not, to calculate the power equation under $n_i = n_{0i} + 1$ and to check if it meets the desired power and level of significance.

(iii) Iteratively this process until the power and α values meets the desired level. The sample size n_i is currently the smallest size necessary to ensure the given power and significance level.

3.3 Simulations for MCC

Under heteroscedasticity, we have shown that the single-stage sampling procedure can produce an exact test statistic for multiple comparison with a control. Here, we use a simulation study for MCC to evaluate its quality. We validated and compared the quality of classical Dunnett test , plug-in procedure from Hasler and Hothorn (2008), single-stage and modified single-stage sampling procedure for MCC in terms of familywise type I error rate (FWE) (α); the nominal level is $\alpha = 0.05$. Similar to Hasler (2016) scenario setting, we compared using the one-sided multiple comparison method with a control, with first treat-

ment was considered as the control. Based on Hasler's (2016) scenarios, there are three kinds of treatments, $k = 3, 4,$ and 5 respectively, under the normal distributions with expected values of treatments are identical $\mu_i = 100$ $(i = 1, \ldots, k)$. Five sample size allocations ($n_1 = 4, 8, 12, 16, 20$) and unbalanced sample sizes per allocation ($n_i = n_{i-1} + 2$ $(i = 2, \ldots, k)$) were selected. Note that an initial sample of size n ($n_1 \geq 3$) will work in theory, but Bishop and Dudewicz (1978) suggested that 10 or more giving better results. The following three alternative variance scenarios were considered:

• the homoscedastic case (case 1): $\sigma_i = 10, i = 1, \ldots, k$,

• the heteroscedastic case I (case 2): $\sigma_i = 10 + 40(i-1)/(k-1), i = 1, \ldots, k$,

• the heteroscedastic case II (case 3): $\sigma_i = 50 - 40(i-1)/(k-1), i = 1, \ldots, k$.

Case 2 represents the treatment with the smallest standard deviation has the smallest sample size, while the treatment with the highest standard deviation has the smallest sample size. The familywise error rates are calculated by dividing the number of times we rejected the null hypothesis by the total number of 5000 trials. To obtain variability, we repeat this procedure 200 simulations. Here, Dunn indicates the Dunnett Test, PI indicates the plug-in procedure, and SS and MSS indicate the single-stage and modified single-stage sampling procedure, respectively.

The simulation results with varying sample sizes correspond to the various variance situations, as shown in Table J in Appendices. Overall, regardless of whether the variances are homogeneous or heterogeneous, PI, SS and MSS all show similar FWEs around 0.05. In contrast, Dunnett test can only maintain FWE under homogeneity, and it will be conservative or liberal when the variance is heterogeneous. Under directly paired sample sizes and variances (case 2) that is the control with the smallest standard deviation and the smallest sample

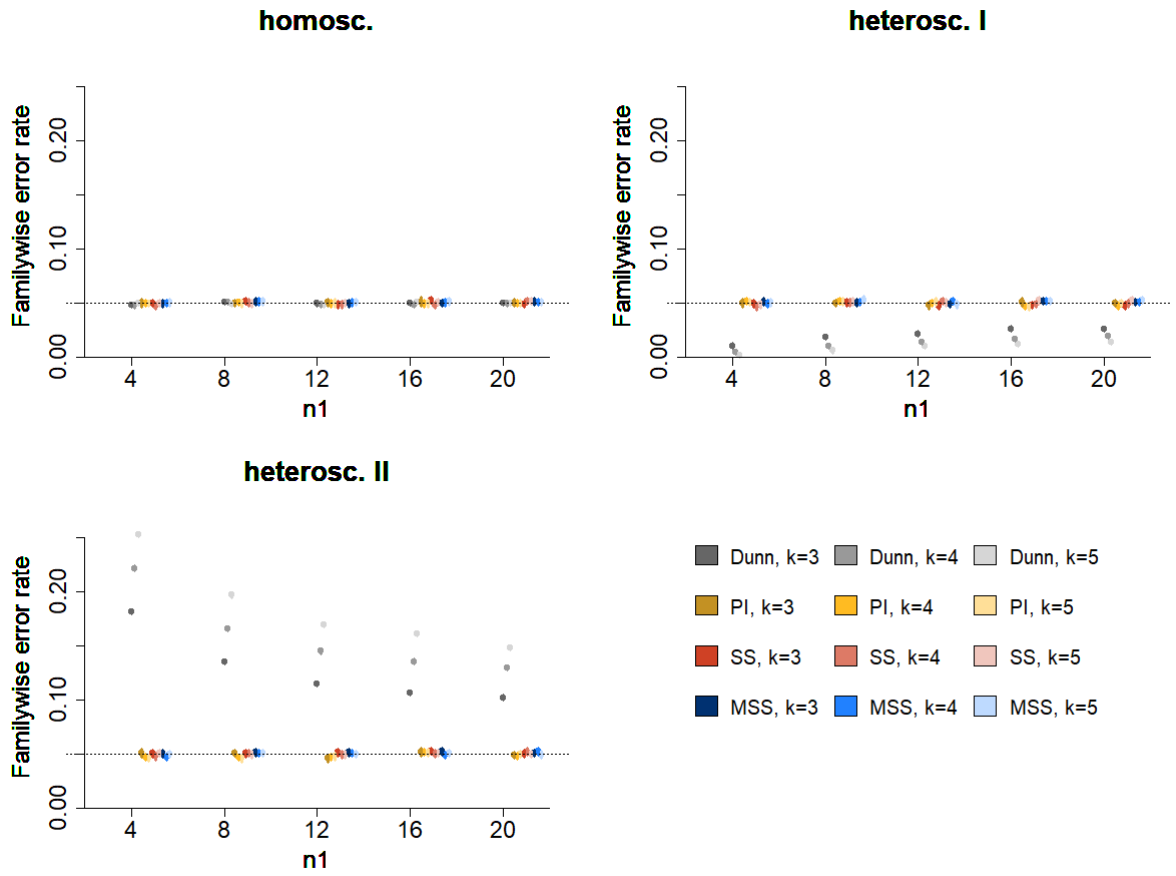


Figure 3.1: Simulated FWEs under different sample sizes allocations and variance scenarios

size, the Dunnett test appears conservative results. The Dunnett test are substantially above 5% (liberal) under inversely paired sample sizes and variances (case 3). The SS and MSS procedures exhibit the same robustness as PI in all scenarios, with FWEs between 0.047 to 0.053 with standard deviations 0.0031 to 0.0048. When the sample sizes are small, the PI procedure is slightly more conservative than the SS and MSS procedures.

The simulation results with sub-figures correspond to the different variance scenarios are shown in Figure 3.1. Although it seems that there is no significant difference between the SS and MSS procedures, the numerical results results in Table J reveal that the FWEs of MSS are a slightly larger than that of SS in some cases. More simulations can be employed in future studies to determine when to adopt either of the two procedures.

3.4 Numerical Example

The mutagenicity data set can be found in the R package $mratios$ (Djira et al., 2020); Adler and Kliesch are the sources for these data (1990). Mutagenicity assay for 4 dosages (30, 50, 70, 100 mg/kg) of a compound (hydroquinone) against a vehicle control. After 24 hours, counts of micronuclei in polychromatic erythrocytes are used to assess the ability to produce chromosomal damage. The goal is to see if the underlying chemical substance may cause chromosomal damage. Table G.4 in Appendices is a statistical summary of the number of micronuclei per male mouse.

The data are used as an example for a multiple comparison to a control problem, despite the fact that the trial's goal was different. The Shapiro-Wilk test was employed to confirm the normality of the data, yielding a p-value of at least 0.05 with the exception of 0.045 in the 100 mg/kg treatment. Despite the data deviates somewhat from normality, it is often robust under a statistical model with a Gaussian error structure. The degree of significance of a test is typically unaffected as long as the distributions are not significantly different from a normal distribution (Knief and Forstmeier, 2021). Dudewicz and Van Der Meulen (1983) have also demonstrated robustness results of the procedures suitable for general non-normal distributions. The homogeneity of variances was tested using Levene's test. The variances of the data are unequal among treatments (p-value $= 0.028$).

We used the plug-in, single-stage, and modified single-stage sampling procedures, respectively to analyze this example. Among them, the plug-in approach may be employed using R package $SimComp$ (Hasler, 2019). For the single-stage sampling procedure (SS), we calculated $\overline{X}_i$ and S_i with the first $n_i - 1$ observations for $i = 1, \ldots, 5$. Next, we used weights of U_i and V_i from Equations (3.4) and (3.5) to calculate the weighted sample mean

$\widehat{X}_i$ (Equation (2.2)) for treatment i. Table 3.1 displays the summary statistics.

Table 3.1: Summary statistics with MCC for the mutagenicity data

		Treatment groups			
	Vehicle control	30 mg/kg	50 mg/kg	75 mg/kg	100 mg/kg
SS procedure					
n_i	7	5	5	5	5
$\overline{X}_i$	2.167	3.500	5.750	13.000	19.250
S_i	0.753	1.000	1.258	3.742	4.272
U_i	0.530	0.615	0.524	0.255	0.200
V_i	-2.181	-1.461	-1.098	-0.020	0.200
$\widehat{X}_i$	-4.012	1.308	3.280	12.898	20.000
Lower conf. intervals		$(-3.915, \infty)$	$(-1.943, \infty)$	$(7.675, \infty)^*$	$(14.776, \infty)^*$
MSS procedure					
U_i	0.469	0.615	0.524	0.255	0.200
V_i	-1.812	1.461	-1.098	-0.020	0.200
$\widetilde{X}_i$	-2.967	1.308	3.280	12.090	20.000
Lower conf. intervals		$(-4.580, \infty)$	$(-2.608, \infty)$	$(7.010, \infty)^*$	$(14.112, \infty)^*$
PI procedure					
Lower conf. intervals		$(-0.557, \infty)$	$(1.391, \infty)^*$	$(5.495, \infty)^*$	$(11.319, \infty)^*$

Multiple comparisons with a control for the mutagenicity data with the single-stage sampling procedure (SS), the modified single-stage sampling procedure (MSS) and the plug-in approach (PI), respectively.

The simultaneous lower confidence intervals for the SS procedure are $I_L = (\widehat{X}_i - \widehat{X}_1 - a\sqrt{z^*}, \infty)$, where $z^* = \max\left\{\frac{0.753^2}{7}, \frac{1^2}{5}, \frac{1.258^2}{5}, \frac{3.742^2}{5}, \frac{4.272^2}{5}\right\} = 3.650$ and the critical value $a = 4.834$ can be obtain from R Shiny interface which is shown in Appendix K. The lower confidence limits are as follows:

$$IL_1 = 1.308 - (-4.012) - 4.834 \times \sqrt{3.65} = -3.915$$

$$IL_2 = 3.280 - (-4.012) - 4.834 \times \sqrt{3.65} = -1.943$$

$$IL_3 = 12.898 - (-4.012) - 4.834 \times \sqrt{3.65} = 7.675$$

$$IL_4 = 20.000 - (-4.012) - 4.834 \times \sqrt{3.65} = 14.776.$$

Based on the results of the SS procedure, the two lower dosages of 30 and 50 mg/kg are non-

significant to the control. The two higher dosages of 75 and 100 mg/kg show significance higher than the control.

Furthermore, we applied the modified single-stage sampling procedure for MCC (MSS) to this example. Using the weights of U_i and V_i from Equations (2.5) and (2.6) to calculate the weighted sample mean, $\widetilde{X}_i$, for treatment $i, \ldots, 5$. The simultaneous lower confidence intervals of the MSS procedure are $I_L = (\widetilde{X}_i - \widetilde{X}_1 - c^* \frac{S_{[k]}}{\sqrt{n_i}}, \infty)$, where the critical value $c^* = 4.635$. The lower confidence limits are as follows:

$$IL_1 = 1.308 - (-2.967) - 4.635 \times \frac{4.272}{\sqrt{5}} = -4.580$$
$$IL_2 = 3.280 - (-2.967) - 4.635 \times \frac{4.272}{\sqrt{5}} = -2.608$$
$$IL_3 = 12.090 - (-2.967) - 4.635 \times \frac{4.272}{\sqrt{5}} = 7.010$$
$$IL_4 = 20.000 - (-2.967) - 4.635 \times \frac{4.272}{\sqrt{5}} = 14.112.$$

The results of The MSS procedure show only two higher dosages of 75 and 100 mg/kg significance difference than the control that is same as the SS procedure. In addition to the aforementioned significance dosages results, the PI approach additionally present significantly higher than the control at dosage of 50 mg/kg. In this example, although the inference of the SS and MSS procedures that take the largest variance to eliminate the influence of unequal variances display more conservative results, the SS and MSS procedure provide an exact distribution for test statistics distribution of the MCC under heteroscedasticity problems.

Chapter 4

Conclusion and Discussion

4.1 Conclusion

The single-stage sampling procedures for analysis of means (ANOM) applied in this paper can be regarded as an alternative to the analysis of variance (ANOVA). The result of overall means test, such as ANOVA, sometimes contradicts the pairwise comparisons causing difficulties interpretation. The ANOVA model does not ensure consistent results for multiple comparisons between treatments, whether or not these comparisons are statistically significant. In contrast, ANOM can generate a clear graphical procedure and thereby ensure consistent results. When the treatment means go outside of the decision limits, they are regarded statistically different from the overall mean. As a result, ANOM is a useful and convenient method for testing multiple population means. The assumption of homogeneity, however, restricts its use, and it is inconvenient or impossible to transform data to make the variances equal. In this thesis, we modified Chen and Lam's (1989), Chen and Chen's (1998) and Chen's (2001) single-stage sampling procedure for the ANOM under heteroscedasticity, respectively, to eliminate the necessity of such an assumption.

The simulation results showed that the single-stage sampling procedure for HANOM can maintain robust regardless of homoscedasticity or heteroscedasticity. It was observed

that ANOVA, BF and ANOM tests were influenced from heterogeneity of variances; however, this negative effect appeared to be more obvious in ANOM test. We believe that the suggested single-stage sampling process $\mathcal{P}_1$ and the modified single-stage sampling technique$\mathcal{P}_2$ for HANOM are suitable to the heteroscedasticity and varying sample sizes. The $\mathcal{P}_1$ and $\mathcal{P}_2$ techniques are more conservative in homogeneity circumstances, with fewer groups and smaller sample sizes. Traditional methods of testing for heteroscedasticity such as Brown-Forsythe test produce similar simulated results. Brown and Forsythe (1974), Li et al. (2012), and Wang (2018) also conducted related comparison studies. When there are more groups or more observations per treatment group, the simulation results for the single-stage sampling approach for HANOM may be improved. We suggested that the sample size n_i for each treatment in the single-stage sampling procedure for HANOM should consist of at least 10 observations as suggested by Bishop and Dudewicz (1978). In fact, we can achieve reasonably decent results in our simulations as long as the number of groups is more than 4 and the size of each treatment is greater than 5 observations.

In addition to the application of HANOM, the single-stage procedure is also applied to MCC under heteroscedasticity and it also can provide an exact distribution for its statistic. Based on the FWE simulation results, the single-stage and mortified single-stage sampling procedure for MCC maintains the FWEs for all sample sizes and variance scenarios. Compared with the procedures for HANOM, the critical values of MCC calculated by the integral method is more able to meet the pre-specified significance level.

We provided several numerical examples to illustrate how to apply the procedures to find the critical values and make a decision for comparing several means including the HANOM and MCC under heteroscedasticity. To make it easier for people to use the procedures, a

web application using R Shiny was developed for the single-stage sampling procedure on HANOM with balanced or unbalanced design or MCC under heteroscedasticity.

In conclusion, the level of the single-stage sampling procedures are completely independent of the unknown variances and can also provide an exact distribution for its statistics employed in the ANOM and MCC under heteroscedasticity. Therefore, ANOM and MCC can be utilized by a single-stage sampling procedure in a data analysis whether the variances are equal or unequal.

4.2 Discussion

Despite the similar results of the two procedures for HANOM in the simulation studies and numerical examples, we offered some recommendations for selecting an appropriate procedure. We recommend using the procedure $\mathcal{P}_1$ when the difference in the number of samples in each group is substantial, the $\mathcal{P}_1$ procedure can be employed more efficiently in terms of significance. However, the $\mathcal{P}_2$ procedure can be employed to control the large variation when the variances betweens treatments are large. That is because the standard error of $\mathcal{P}_2$ for HANOM is the square root of the maximum of S_i^2/n_i. Which procedure is more suitable depends on the number of samples and the configuration of different heterogeneity variances. In future studies, more different scenarios will be set to conduct simulation comparisons and find out the recommendations for using the procedures.

The two-stage sampling procedure of HANOM proposed by Nelson and Dudewicz (2002) also eliminates the heteroscedasticity problems. But, the two-stage sampling procedure is a design-oriented procedure that requires additional samples at the second stage. Due to time, budget and other constraints, there is only one sample or the necessary two-stage

sample size is not met, which may not be feasible in practical applications. The advantages of the single-stage sampling procedure are that it is data-analysis oriented, and it works for any given dataset. Chen (2001), Chen and Wen (2006) and Wen et al. (2017) addressed the relative advantages of the single-stage and two-stage sampling procedures. The decision lines of the HANOM by the single-stage sampling procedure are data dependent (depending on $S^2_{[k]}/n_i$) which could be equal to, smaller than, or larger than the decision lines specified by the two-stage sampling procedure. The comparison is concluded as follows:

(i) When $S_i^2/n_i = S_j^2/n_j$ for all i, j, the single-stage and two-stage sampling procedures have the same decision lines. Under this condition, $\left(\frac{\omega}{\delta}\right)^2 S^2_{i,n_0}$ except for a rounding error in the sample size determined by a classical two-stage sampling procedure is equal to $S^2_{[k]}/n_i$ in the single-stage sampling procedure, which gives the same sample sizes in addition to the same decision lines.

(ii) When $\underset{1\le j\le k}{max}\, S_i^2/n_j < \left(\frac{\omega}{\delta}\right)^2 S^2_{i,n_0}$, the single-stage sampling procedure has a smaller decision line than that of two-stage sampling procedure.

(iii) When $\underset{1\le j\le k}{min}\, S_i^2/n_j > \left(\frac{\omega}{\delta}\right)^2 S^2_{i,n_0}$, the single-stage sampling procedure has a larger decision line than that of two-stage sampling procedure.

(iv) In other situations, the decision lines of the single-stage sampling procedure could be larger than, smaller than or equal to those of the two-stage sampling procedure depending on actual sample data.

The theoretical comparison between the single-stage and two-stage sampling procedures is very difficult. This is because the expected decision lines by the single-stage sampling procedure are determined by a given significance level and unknown population variances, but the

parameters for a two-stage sampling procedure are derived simultaneously from significance, power, and effect size.

In conclusion, in the study design phase, one can use a two-stage sampling procedure to perform multiple comparisons in the presence of heteroscedasticity, but if only one available data on hand, a single-stage procedure is a feasible approach. In the case that there is just one accessible data, users also can consider a nonparametric method, especially if the data is not normally distribution. We still recommend using a single-stage sampling procedure rather than a nonparametric approach to solve this issue. That is because that a statistical models with a Gaussian error structure is generally robust to violations of the normality assumption (Knief and Forstmeier, 2021), and a nonparametric method typically has lower test power. In future research, we can compare the applicability betrween nonparametric methods and single-stage sampling procedures using a simulation study.

4.3 Future Research

After completing these studies, some further problems remain to be resolved. The areas of further research include the following:

(i) More efficient for determining sample sizes under different least favorable configurations and power definitions.

(ii) To use and modify the single-stage sampling procedures to the correlated populations.

(iii) HANOM in a two-way layout or higher-way model may be developed.

(iv) Consider the arithmetic mean instead of weighted sample mean to modify a two-stage sampling procedure for HANOM and MCC problems.

References

1. Adler, I., and Kliesch, U. (1990). Comparison of single and multiple treatment regimens in the mouse bone marrow micronucleus assay for hydroquinone (HQ) and cyclophosphamide (CP). *Mutation Research*, 234(3-4), 115-123.

2. Bishop, T. A. (1976). Heteroscedastic ANOVA, MANOVA, and multiple-comparisons. *Unpublished Ph.D. thesis*, The Ohio State University.

3. Bishop, T. A. (1979). Some results on simultaneous inference for analysis of variance with unequal variances. *Technometrics*, 21(3), 337-340.

4. Bishop, T. A., and Dudewicz, E. J. (1978). Exact analysis of variance with unequal variances: test procedures and tables. *Technometrics*, 20(4), 419-430.

5. Brown, M., and Forsythe, A. (1974). Small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16(1), 129-132.

6. Casella, G., and Berger, R. L. (2002). *Statistical inference*. Section 9.3: Methods of Evaluating Interval Estimation, page 443. 2^{nd} Edition, Duxbury, CA, USA.

7. Castillo-Cuenca, J. C., Martínez-Moreno, Á., Diaz-Cao, J. M., Entrena-García, A., Fraga, J., Arias, P. C., Almería, S., and García-Bocanegra, I. (2021). Seroprevalence of Toxoplasma gondii and associated risk factors in domestic pigs raised from Cuba. *Parasitology Research*, 120(8), 2897-2903.

8. Chen, H. J., and Lam, K. (1989). Single-stage interval estimation of the largest nor-

mal mean under heteroscedasnoty. *Communications in Statistics-Theory and Methods*, 18(10), 3703-3718.

9. Chen, H. J., and Wen, M. J. (2006). Optimal confidence interval for the largest normal mean under heteroscedasticity. *Computational statistics and data analysis*, 51(2), 982-1001.

10. Chen, H. J., Wen, M. J., and Chuang, C. J. (2011). On testing the equivalence of treatments using the measure of range. *Computational Statistics and Data Analysis*, 55, 603−614.

11. Chen, H. J., Wen, M. J., and Wang, A. M. (2009). On testing the bioequivalence of several treatments using the measure of distance. *Statistics*, 43(5), 513−530.

12. Chen, H. J., and Wu, S. F. (2012). A confidence region for the largest and the smallest means under heteroscedasticity. *Computational Statistics and Data Analysis*, 56(6), 1692-1702.

13. Chen, S. Y. (2001). One-stage and two-stage statistical inference under heteroscedasticity. *Communications in Statistics-Simulation and Computation*, 30(4), 991-1009.

14. Chen, S. Y., and Chen, H. J. (1998). Single-stage analysis of variance under heteroscedasticity. *Communications in Statistics-Simulation and Computation*, 27(3), 641-666.

15. Chen, S. Y., Chen, H. J., and Chang, H. F. (2004). A one-stage procedure for testing homogeneity of means against an ordered alternative under unequal variances. *Communications in Statistics-Simulation and Computation*, 33(1), 49-67.

16. Dag, O., Dolgun, A., Konar, N. M., Weerahandi, S., Ananda, M., and Dag, M. O. (2021). Package 'onewaytests'. Available at:`https://cran.r-project.org/web/packages/onewaytests/index.html`. *R package version, 2.6*.

17. Djira, G., Hasler, M., Gerhard, D., and Schaarschmidt, F. (2020). mratios: ratios of coefficients in the general linear model. `https://carn.r-project.org/package=mratios`. *R package version, 1.4.2*.

18. Dudewicz, E. J., and Dalal, S. R. (1983). Multiple-comparisons with a control when variances are unknown and unequal. *American Journal of Mathematical and Management Sciences*, 3(4), 275-295.

19. Dudewicz, E. J., and Mishra, S. (1988). *Modern mathematical statistics*, pp. 500-501. John Wiley, NY.

20. Dudewicz, E. J., and Nelson, P. R. (2003). Heteroscedastic analysis of means (HANOM). *American Journal of Mathematical and Management Sciences*, 23(1-2), 143-181.

21. Dudewicz, E. J., Ramberg, J. S., and Chen, H. J. (1975). New tables for multiple comparisons with a control (unknown variances). *Biometrische Zeitschrift*, 17(1), 13-26.

22. Dudewicz, E. J., and Van Der Meulen, E. C. (1983). Entropy-based statistical inference, ii: selection-of-the-best/complete ranking for continuous distributions on (0, 1), with applications to random number generators. *Statistics and Risk Modeling*, 1(2), 131–146.

23. 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.

24. Feir-Walsh, B. J., and Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34(4), 789-799.

25. 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(2), 113-125.

26. Guenther, W. C. (1969). Shortest confidence intervals. *The American Statistician*, 23(1), 22-25.

27. Hasler, M. (2016). Heteroscedasticity: multiple degrees of freedom vs. sandwich estimation. *Statistical Papers*, 57(1), 55-68.

28. Hasler, M. (2019). SimComp: simultaneous comparisons for multiple endpoints. `http://CRAN.R-project.org/package=SimComp`, *R package version 3.3*.

29. Hasler, M., and Hothorn, L. A. (2008). Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 50(5), 793-800.

30. Horn, M., and Dunnett, C. W. (2004). Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. *Lecture Notes-Monograph*, Series 47, 48-64.

31. Hsu, J. (1996). *Multiple comparisons: theory and methods* CRC Press, NY.

32. Juneau, P. (2003). Using SAS to perform a single-stage multiple comparison procedure for all pair-wise comparisons in a one-way layout with unequal variances. In

Proceedings of the PharmaSUG 2003 Annual Conference, May 4 – 7, 2003, Miami, Florida, U.S.A.

33. Kiani, M., Panaretos, J., and Psarakis, S. (2008). A new procedure to monitor the mean of a quality characteristic. *Communications in Statistics—Simulation and Computation*, 37(9), 1870-1880.

34. Knief, U., and Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*. doi:`https://doi.org/10.3758/s13428-021-01587-5`.

35. Krutchkoff, R. G. (1988). One—way fixed effects analysis of variance when the error varianes may be unequal. *Journal of Statistical Computation and Simulation*, 30(4), 259-271.

36. Lam, K. (1987). Subset selection of normal populations under heteroscedasticity. Presented at *the second International Advanced Seminar/Workshop on Inference Procedures Associated with Statistal Ranking and Selection*, August 9-15, 1987, Sydney, Australia.

37. Lam, K. (1992). Subset selection of normal populations under heteroscedasticity. In *proceedings and discussions of the second international conference on inference procedures associated with statistal ranking and selection on The frontiers of modern statistical inference procedures, II*, Sydney, Australia.

38. Li, X., Tian, L., Wang, J., and Muindi, J. R. (2012). Comparison of quantiles for several normal populations. *Computational Statistics and Data Analysis*, 56(6), 2129-2138.

39. Lopez-Mejia, M., and Roldan-Valadez, E. (2016). Comparisons of apparent diffusion

coefficient values in penumbra, infarct, and normal brain regions in acute ischemic stroke: Confirmatory data using bootstrap confidence intervals, analysis of variance, and analysis of means. *Journal of Stroke and Cerebrovascular Diseases*, 25(3), 515-522.

40. Mendeş, M., and Yiğit, S. (2013). Comparison of ANOVA-F and ANOM tests with regard to type I error rate and test power. *Journal of Statistical Computation and Simulation*, 83(11), 2093-2104.

41. Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*, 2^{nd} Edition. Springer, New York.

42. Nelson, P. R. (1982). Exact critical points for the analysis of means. *Communications in Statistics: Theory and Methods*, 11(6), 699-709.

43. Nelson, P. R. (1983). A comparison of sample sizes for the analysis of means and the analysis of variance. *Journal of Quality Technology*, 15(1), 33-39.

44. Nelson, P. R. (1991). Numerical evaluation of multivariate normal integrals with correlations $\rho_{lj} = -\alpha_l \alpha_j$. *The Frontiers of Statistical Scientific Theory and Industrial Applications*, Vol. II, 97-114.

45. Nelson, P. R., and Dudewicz, E. J. (2002). Exact analysis of means with unequal variances. *Technometrics*, 44(2), 152-160.

46. Nelson, P. R., Wludyka, P. S., and Copeland, K. A. (2005). The analysis of means: a graphical method for comparing means, rates, and proportions. *Society for Industrial and Applied Mathematics*, pp.25.

47. Ott, E. R. (1967). Analysis of means - a graphical procedure. *Industrial Quality Control*, 24(2), 101-109.

48. Pallmann P. (2017). Package 'ANOM'. Available at:`http://CRAN.R-project.org/package=ANOM`. *R package version 0.4*.

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

50. Paulson, E. (1952). On the comparison of several experimental categories with a control. *The Annals of Mathematical Statistics*, 23(2), 239-246.

51. Stanciu, D., and Chis, A. (2021). A study of cyberstander reactions under community-related influence: when gender complicates matters. *Computers in human behavior*, 115, 106589. doi:`https://doi.org/10.1016/j.chb.2020.106589`.

52. Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3), 243-258.

53. Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6), 110-114.

54. Tamhane, A. C. (1977). Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics-Theory and Methods*, 6(1), 15-32.

55. Wang, W. M., Wen, C. C., Das, R., and Wen, M. J. (2022). A modified single-stage sampling procedure for heteroscedasticity analysis of means with unbalanced design, to appear in *Communications in Statistics-Simulation and Computation*. doi:`https://doi.org/10.1080/03610918.2022.2074456`.

56. Wang, Y., Pham, T., Nguyen, D., Kim, E. S., Chen, Y. H., Kromrey, J., Yi, Z., and Yin, Y. (2018). Evaluating the Efficacy of Conditional Analysis of Variance under Heterogeneity and Non-Normality. *Journal of Modern Applied Statistical Methods*, 17(2), eP2701. doi: 10.22237/jmasm/1555340224.

57. Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica: Journal of the Econometric Society*, 1211-1215.

58. Weerahandi, S. (2004). Generalized inference in repeated measures: Exact methods in MANOVA and mixed models. *John Wiley & Sons*, pp.43.

59. Weerahandi, S., and Krishnamoorthy, K. (2019). A note reconciling ANOVA tests under unequal error variances. *Communications in Statistics-Theory and Methods*, 48(3), 689-693.

60. Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.

61. Wen, M. J., and Chen, H. J. (1994). Single-stage multiple comparison procedures under heteroscedasticity. *American Journal of Mathematical and Management Sciences*, 14(1-2), 1-48.

62. Wen, M. J., Chen, S. Y., and Chen, H. J. (2007). On testing a subset of regression parameters under heteroskedasticity. *Computational Statistics and Data Analysis*, 51(12), 5958−5976.

63. Wen, M. J., Huang, L. C, and Zhong, J. (2017). Single-stage sampling procedure of the t best populations under heteroscedasticity. *Communications in Statistics-Theory and Methods*, 46(18), 9265-9273.

64. Wen, M. J., Wen, C. C., and Wang, W. M. (2022). Single-stage sampling procedure for heteroscedasticity in multiple comparisons with a control, to appear in *Communications in Statistics-Simulation and Computation*. doi:https://doi.org/10.1080/03610918.2022.2053863.

65. Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, 92(437), 299-306.

66. Wludyka, P. S., and Nelson, P. R. (1997). An analysis-of-means-type test for variances from normal populations. *Technometrics*, 39(3), 274-285.

67. Wu, S. F., and Chen, H. J. (1998). Multiple comparisons with the average for normal distributions. *American Journal of Mathematical and Management Sciences*, 18(1-2), 193-218.

Appendix A. Critical Values of decision lines (d^*)

Tables A1-A3 give the critical values d^* in Equation (3) and (4) for HANOM decision lines. The following critical values d^* are simulated form the upper $(100\alpha/2)^{th}$ percentile of the approximate sampling distribution of $\widetilde{T}_{[k]}$ (i.e. solving $P(\widetilde{T}_{[k]} < d^*) = 1 - \alpha/2$). We provide three scenarios when $1 - \alpha = 0.9, 0.95, 0.99$.

Table A.1: HANOM Critical Values d^* where $\frac{\alpha}{2} = 0.05$

	Number of Treatments, k																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	12.69	19.65	26.13	32.54	38.71	45.05	51.42	57.88	63.81	70.03	76.31	82.48	88.65	94.64	101.38	107.79	113.84	120.71	126.73
2	3.25	4.39	5.27	6.04	6.74	7.38	7.98	8.57	9.11	9.61	10.10	10.58	11.01	11.44	11.84	12.24	12.61	12.99	13.39
3	2.29	2.99	3.48	3.89	4.24	4.55	4.83	5.08	5.32	5.54	5.74	5.94	6.13	6.29	6.47	6.62	6.77	6.92	7.07
4	1.97	2.54	2.92	3.22	3.48	3.70	3.89	4.06	4.23	4.37	4.50	4.63	4.75	4.86	4.96	5.05	5.15	5.24	5.33
5	1.81	2.32	2.65	2.91	3.12	3.30	3.45	3.59	3.72	3.83	3.94	4.03	4.12	4.20	4.28	4.35	4.43	4.49	4.56
6	1.73	2.19	2.50	2.72	2.92	3.07	3.21	3.33	3.43	3.53	3.62	3.69	3.77	3.84	3.91	3.96	4.03	4.08	4.13
7	1.66	2.11	2.39	2.61	2.78	2.92	3.04	3.15	3.25	3.33	3.41	3.48	3.55	3.61	3.66	3.72	3.77	3.81	3.86
8	1.62	2.06	2.33	2.53	2.69	2.82	2.93	3.03	3.12	3.20	3.27	3.33	3.39	3.45	3.50	3.55	3.60	3.64	3.68
9	1.59	2.01	2.28	2.47	2.62	2.75	2.86	2.94	3.03	3.10	3.17	3.23	3.28	3.33	3.39	3.43	3.47	3.51	3.55
10	1.57	1.98	2.23	2.42	2.57	2.69	2.79	2.88	2.96	3.03	3.09	3.15	3.20	3.25	3.30	3.34	3.38	3.41	3.45
11	1.55	1.96	2.21	2.38	2.53	2.64	2.74	2.83	2.90	2.97	3.03	3.08	3.13	3.18	3.22	3.27	3.30	3.34	3.37
12	1.54	1.94	2.18	2.36	2.49	2.61	2.70	2.78	2.86	2.92	2.98	3.03	3.08	3.13	3.17	3.21	3.24	3.28	3.31
13	1.52	1.92	2.16	2.33	2.47	2.58	2.67	2.75	2.82	2.88	2.94	2.99	3.04	3.08	3.12	3.16	3.19	3.22	3.25
14	1.51	1.91	2.14	2.31	2.44	2.55	2.64	2.72	2.79	2.85	2.91	2.95	3.00	3.04	3.08	3.12	3.15	3.18	3.21
15	1.50	1.89	2.13	2.30	2.43	2.53	2.62	2.70	2.77	2.83	2.88	2.92	2.97	3.01	3.05	3.08	3.12	3.15	3.17
16	1.49	1.88	2.11	2.28	2.41	2.51	2.60	2.68	2.74	2.80	2.85	2.90	2.94	2.98	3.02	3.05	3.09	3.11	3.14
17	1.49	1.87	2.10	2.27	2.40	2.50	2.58	2.66	2.72	2.78	2.83	2.88	2.92	2.96	2.99	3.03	3.06	3.09	3.11
18	1.48	1.86	2.09	2.26	2.38	2.48	2.57	2.64	2.71	2.76	2.81	2.86	2.90	2.94	2.97	3.00	3.03	3.06	3.09
19	1.47	1.86	2.08	2.24	2.37	2.47	2.55	2.63	2.69	2.74	2.80	2.84	2.88	2.92	2.95	2.98	3.01	3.04	3.07
20	1.47	1.85	2.08	2.24	2.36	2.46	2.54	2.61	2.68	2.73	2.78	2.82	2.86	2.90	2.94	2.97	3.00	3.02	3.05
40	1.43	1.79	2.01	2.15	2.27	2.36	2.43	2.50	2.56	2.60	2.65	2.68	2.72	2.75	2.78	2.81	2.83	2.86	2.88
60	1.41	1.77	1.99	2.13	2.24	2.33	2.40	2.46	2.52	2.56	2.60	2.64	2.67	2.70	2.73	2.76	2.78	2.81	2.83
120	1.40	1.76	1.96	2.10	2.21	2.30	2.37	2.43	2.48	2.52	2.56	2.60	2.63	2.66	2.69	2.71	2.73	2.76	2.78
∞	1.39	1.74	1.94	2.08	2.18	2.27	2.33	2.39	2.44	2.48	2.52	2.55	2.59	2.61	2.64	2.66	2.69	2.71	2.73

Table A.2: HANOM Critical Values d^* where $\frac{\alpha}{2} = 0.025$

	Number of Treatments, k																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	25.48	38.97	51.83	64.79	76.83	89.34	101.83	115.45	127.64	139.84	152.69	165.16	178.66	191.08	204.01	216.46	229.7	241.20	252.48
2	4.58	6.11	7.37	8.46	9.48	10.46	11.31	12.15	12.95	13.66	14.40	15.06	15.69	16.28	16.77	17.36	17.88	18.39	19.00
3	2.92	3.76	4.38	4.89	5.35	5.74	6.09	6.42	6.74	7.01	7.28	7.52	7.76	7.98	8.22	8.40	8.59	8.78	8.97
4	2.42	3.07	3.52	3.89	4.21	4.46	4.69	4.91	5.10	5.29	5.46	5.60	5.75	5.87	6.00	6.11	6.23	6.33	6.44
5	2.18	2.75	3.14	3.43	3.69	3.89	4.07	4.23	4.37	4.51	4.63	4.74	4.85	4.95	5.04	5.12	5.20	5.29	5.36
6	2.05	2.57	2.90	3.17	3.39	3.56	3.72	3.85	3.97	4.09	4.18	4.28	4.35	4.44	4.50	4.58	4.65	4.71	4.77
7	1.96	2.45	2.76	3.00	3.19	3.35	3.49	3.61	3.72	3.81	3.90	3.98	4.05	4.12	4.19	4.24	4.30	4.36	4.41
8	1.90	2.37	2.67	2.89	3.07	3.22	3.34	3.44	3.54	3.63	3.71	3.79	3.85	3.91	3.97	4.03	4.07	4.12	4.16
9	1.85	2.31	2.60	2.81	2.98	3.11	3.24	3.34	3.43	3.50	3.58	3.64	3.70	3.76	3.81	3.86	3.90	3.95	3.99
10	1.82	2.27	2.55	2.75	2.91	3.04	3.15	3.25	3.33	3.41	3.48	3.54	3.59	3.65	3.70	3.74	3.78	3.82	3.86
11	1.80	2.24	2.51	2.70	2.86	2.98	3.09	3.18	3.26	3.33	3.39	3.45	3.50	3.55	3.60	3.65	3.68	3.72	3.75
12	1.78	2.21	2.47	2.66	2.81	2.93	3.03	3.12	3.20	3.26	3.33	3.38	3.43	3.48	3.52	3.56	3.60	3.63	3.67
13	1.77	2.19	2.45	2.63	2.78	2.89	2.99	3.07	3.15	3.21	3.27	3.32	3.37	3.42	3.46	3.50	3.53	3.57	3.60
14	1.75	2.17	2.42	2.60	2.74	2.86	2.95	3.04	3.11	3.17	3.23	3.28	3.32	3.37	3.41	3.45	3.48	3.51	3.54
15	1.74	2.16	2.40	2.58	2.72	2.83	2.93	3.01	3.07	3.14	3.19	3.24	3.29	3.33	3.37	3.40	3.43	3.46	3.50
16	1.73	2.14	2.39	2.56	2.69	2.80	2.90	2.98	3.04	3.11	3.16	3.21	3.25	3.30	3.33	3.37	3.39	3.42	3.45
17	1.72	2.13	2.37	2.54	2.67	2.79	2.88	2.95	3.02	3.08	3.13	3.17	3.22	3.27	3.30	3.33	3.36	3.39	3.42
18	1.71	2.12	2.36	2.53	2.66	2.77	2.86	2.93	3.00	3.05	3.10	3.15	3.20	3.23	3.27	3.30	3.33	3.36	3.39
19	1.70	2.11	2.35	2.51	2.65	2.75	2.84	2.91	2.98	3.03	3.08	3.13	3.17	3.21	3.24	3.28	3.31	3.34	3.36
20	1.70	2.10	2.34	2.51	2.64	2.74	2.82	2.90	2.96	3.02	3.06	3.11	3.15	3.19	3.23	3.26	3.28	3.31	3.34
40	1.63	2.02	2.25	2.40	2.52	2.61	2.68	2.74	2.80	2.85	2.89	2.93	2.97	3.00	3.03	3.05	3.08	3.11	3.12
60	1.62	2.00	2.22	2.37	2.48	2.57	2.64	2.70	2.75	2.80	2.84	2.88	2.91	2.94	2.97	2.99	3.02	3.04	3.06
120	1.60	1.98	2.19	2.34	2.44	2.53	2.60	2.65	2.71	2.75	2.79	2.82	2.86	2.88	2.91	2.94	2.96	2.98	3.00
∞	1.59	1.96	2.16	2.30	2.41	2.49	2.56	2.61	2.66	2.70	2.74	2.78	2.81	2.83	2.85	2.88	2.90	2.92	2.94

Table A.3: HANOM Critical Values d^* where $\frac{\alpha}{2} = 0.005$

	Number of Treatments, k																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	123.54	191.21	254.03	309.41	370.90	429.62	498.14	567.66	617.79	671.75	733.49	799.37	866.57	928.19	995.82	1054.71	1116.55	1183.59	1232.61
2	10.03	13.24	15.88	18.29	20.61	22.83	24.76	26.62	28.64	30.25	31.81	33.36	34.82	36.18	37.43	38.62	39.87	41.05	42.17
3	5.01	6.31	7.42	8.39	9.18	9.92	10.52	11.11	11.68	12.19	12.65	13.04	13.43	13.78	14.18	14.55	14.84	15.20	15.45
4	3.69	4.59	5.33	5.89	6.37	6.78	7.13	7.43	7.77	8.04	8.31	8.52	8.77	8.92	9.08	9.24	9.40	9.56	9.76
5	3.13	3.86	4.43	4.87	5.23	5.52	5.79	6.01	6.25	6.44	6.64	6.76	6.89	7.06	7.15	7.26	7.37	7.50	7.60
6	2.83	3.48	3.97	4.35	4.64	4.88	5.09	5.28	5.44	5.61	5.73	5.84	5.95	6.06	6.16	6.24	6.31	6.41	6.48
7	2.65	3.27	3.70	4.01	4.27	4.49	4.65	4.81	4.96	5.06	5.18	5.27	5.35	5.45	5.52	5.58	5.66	5.72	5.80
8	2.55	3.11	3.50	3.78	4.02	4.21	4.36	4.50	4.61	4.73	4.83	4.91	4.98	5.05	5.12	5.18	5.22	5.30	5.34
9	2.46	3.01	3.37	3.63	3.85	4.02	4.16	4.30	4.39	4.49	4.58	4.65	4.72	4.79	4.84	4.89	4.94	4.99	5.04
10	2.39	2.94	3.27	3.53	3.73	3.88	4.01	4.13	4.23	4.32	4.38	4.46	4.52	4.58	4.63	4.68	4.73	4.78	4.81
11	2.36	2.87	3.19	3.43	3.63	3.78	3.90	4.02	4.10	4.18	4.26	4.32	4.40	4.44	4.48	4.53	4.57	4.61	4.65
12	2.32	2.82	3.14	3.36	3.55	3.69	3.80	3.91	4.00	4.08	4.14	4.20	4.27	4.31	4.36	4.40	4.44	4.49	4.51
13	2.30	2.78	3.09	3.31	3.48	3.63	3.74	3.83	3.91	3.99	4.06	4.12	4.18	4.22	4.26	4.29	4.33	4.37	4.40
14	2.26	2.75	3.05	3.26	3.43	3.57	3.69	3.77	3.85	3.93	3.98	4.04	4.09	4.14	4.18	4.21	4.25	4.28	4.32
15	2.24	2.72	3.02	3.21	3.39	3.51	3.62	3.71	3.79	3.87	3.92	3.98	4.03	4.07	4.10	4.14	4.17	4.22	4.24
16	2.22	2.70	2.98	3.19	3.35	3.46	3.57	3.66	3.76	3.81	3.87	3.92	3.96	4.01	4.04	4.08	4.11	4.14	4.18
17	2.21	2.68	2.97	3.16	3.31	3.43	3.54	3.63	3.70	3.76	3.81	3.86	3.92	3.96	3.99	4.03	4.06	4.09	4.11
18	2.19	2.65	2.93	3.13	3.30	3.40	3.51	3.59	3.67	3.73	3.79	3.83	3.88	3.92	3.95	3.98	4.02	4.05	4.07
19	2.18	2.64	2.92	3.11	3.26	3.38	3.48	3.56	3.63	3.69	3.74	3.79	3.84	3.88	3.91	3.94	3.98	4.01	4.03
20	2.17	2.63	2.90	3.10	3.24	3.35	3.46	3.53	3.60	3.65	3.72	3.76	3.81	3.84	3.87	3.92	3.94	3.97	3.99
40	2.07	2.50	2.74	2.91	3.04	3.13	3.22	3.28	3.34	3.40	3.43	3.47	3.51	3.55	3.57	3.59	3.62	3.65	3.67
60	2.04	2.47	2.71	2.86	2.98	3.07	3.15	3.21	3.26	3.31	3.35	3.38	3.42	3.45	3.48	3.50	3.52	3.55	3.56
120	2.02	2.42	2.66	2.81	2.92	3.01	3.08	3.14	3.19	3.24	3.28	3.32	3.34	3.37	3.39	3.42	3.44	3.46	3.47
∞	1.99	2.40	2.62	2.76	2.87	2.95	3.03	3.08	3.13	3.16	3.21	3.23	3.26	3.28	3.31	3.33	3.35	3.37	3.39

Appendix B. Details description for the rejection region of $\mathcal{P}_1$ procedure

According to the rejection region of treatment i for $\mathcal{P}_1$ procedure as $\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}}\right| > h^*,\ i = 1,\ldots,k$, the level of significance, α, for HANOM is

$$\begin{aligned}1-\alpha =& P\left(\left|\frac{\widetilde{X}_i-\overline{\widetilde{X}}}{S_{[k]}/\sqrt{n_i}}\right| \leq h^*,\ i=1,\ldots,k\right)\\ =& P\bigg(-h^* \leq \frac{\widetilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{\overline{\widetilde{X}}-\mu_.}{S_{[k]}/\sqrt{n_i}} + \frac{\mu_i-\mu_.}{S_{[k]}/\sqrt{n_i}} \leq h^*,\ i=1,\ldots,k\bigg),\end{aligned}$$

where $\mu_. = \sum_{i=1}^{k}\mu_i/k$. Under the null hypothesis, $H_0 : \mu_1 = \ldots = \mu_k$, the term $\frac{\mu_i-\mu_.}{S_{[k]}/\sqrt{n}}$ equals zero and $1-\alpha$ will be rewritten by

$$\begin{aligned}1-\alpha &= P\left(-h^* \leq \frac{\widetilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{\overline{\widetilde{X}}-\mu_.}{S_{[k]}/\sqrt{n_i}} \leq h^*,\ i=1,\ldots,k\right)\\ &= P\left(-h^* \leq \frac{\widetilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{1}{k}\sum_{z=1}^{k}\frac{\widetilde{X}_z-\mu_z}{S_{[k]}/\sqrt{n_i}} \leq h^*,\ i=1,\ldots,k\right)\\ &= P\left(-h^* \leq \left(\frac{k-1}{k}\right)\frac{\widetilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{1}{k}\sum_{z\neq i}^{k}\sqrt{\frac{n_i}{n_z}}\frac{\widetilde{X}_z-\mu_z}{S_{[k]}/\sqrt{n_z}} \leq h^*,\ i=1,\ldots,k\right)\\ &= P\left(-h^* \leq \left(\frac{k-1}{k}\right)T_i - \sqrt{n_i}\left(\frac{1}{k}\sum_{z\neq i}^{k}\frac{T_z}{\sqrt{n_z}}\right) \leq h^*,\ i=1,\ldots,k\right)\\ &= P\left(-h^* \leq \widetilde{T}_i \leq h^*,\ i=1,\ldots,k\right)\end{aligned}$$

We let $\widetilde{T}_{[1]}$ is the minimum of $\left\{\widetilde{T}_1,\ldots,\widetilde{T}_k\right\}$ and $\widetilde{T}_{[k]}$ is the maximum of $\left\{\widetilde{T}_1,\ldots,\widetilde{T}_k\right\}$, then

we can rewrite the equation as follows.

$$\begin{aligned}
1-\alpha &= P\left(-h^* < \widetilde{T}_{[1]} \leq \widetilde{T}_{[k]} < h^*\right) \\
&\geq 1 - P\left(\widetilde{T}_{[1]} < -h^*\right) - P\left(\widetilde{T}_{[k]} > h^*\right) \\
&\geq 1 - 2\max\left\{P\left(\widetilde{T}_{[1]} < -h^*\right), P\left(\widetilde{T}_{[k]} > h^*\right)\right\}
\end{aligned}$$

Therefore, the critical value satisfy the following equation.

$$\max\left\{P\left(\widetilde{T}_{[1]} < -h^*\right), P\left(\widetilde{T}_{[k]} > h^*\right)\right\} \leq \frac{\alpha}{2}$$

Appendix C. Details description for the rejection region of $\mathcal{P}_2$ procedure

According to the rejection region of treatment i for $\mathcal{P}_2$ procedure as $\left|\frac{\widehat{X}_i-\overline{\widehat{X}}}{\sqrt{z^*}}\right| > d^*,\ i = 1,\ldots,k$, the level of significance, α, for HANOM is

$$\begin{aligned}
1-\alpha =& P\left(\left|\frac{\widehat{X}_i-\overline{\widehat{X}}}{\sqrt{z^*}}\right| \leq d^*,\ i=1,\ldots,k\right)\\
=& P\left(-d^* \leq \frac{\widehat{X}_i-\overline{\widehat{X}}}{\sqrt{z^*}} \leq d^*,\ i=1,\ldots,k\right)\\
=& P\Bigg(-d^* \leq \frac{\widehat{X}_i-\mu_i}{\sqrt{z^*}} - \frac{\overline{\widehat{X}}-\mu_.}{\sqrt{z^*}} + \frac{\mu_i-\mu_.}{\sqrt{z^*}} \leq d^*,\ i=1,\ldots,k\Bigg)
\end{aligned}$$

where $\mu_. = \sum_{i=1}^{k}\mu_i/k$. We have $T_i = \frac{\widehat{X}_i-\mu_i}{\sqrt{z^*}},\ i=1,\ldots,k$ and then $\bar{T} = \sum_{i=1}^{k} T_i/k = \frac{\overline{\widehat{X}}-\mu_.}{\sqrt{z^*}}$. Under the null hypothesis, $H_0 : \mu_1 = \ldots = \mu_k$, the term $\frac{\mu_i-\mu_.}{\sqrt{z^*}}$ equals zero and $1-\alpha$ will be rewritten by

$$\begin{aligned}
1-\alpha &= P\left(-d^* \leq T_i^* - \bar{T}^* \leq d^*,\ i=1,\ldots,k\right)\\
&= P\left(-d^* \leq \widehat{T}_i \leq d^*,\ i=1,\ldots,k\right)
\end{aligned}$$

where $\widehat{T}_i = T_i - \bar{T},\ i=1,\ldots,k$. And, $\widehat{T}_{[1]}$ is the minimum of $\left\{\widehat{T}_1,\ldots,\widehat{T}_k\right\}$ and $\widehat{T}_{[k]}$ is the maximum of $\left\{\widehat{T}_1,\ldots,\widehat{T}_k\right\}$, then we can rewrite the equation as follows.

$$1-\alpha = P\left(-d^* < \widehat{T}_{[1]} \leq \widehat{T}_{[k]} < d^*\right)$$

$$= 1 - P\left(\left\{\widehat{T}_{[1]} < -d^*\right\} \cup \left\{\widehat{T}_{[k]} > d^*\right\}\right)$$
$$\geq 1 - P\left(\widehat{T}_{[1]} < -d^*\right) - P\left(\widehat{T}_{[k]} > d^*\right)$$

Here, we take

$$P\left(\widehat{T}_{[1]} < -d^*\right) + P\left(\widehat{T}_{[k]} > d^*\right) \leq \alpha$$

As, $(\widehat{T_1}, \ldots, \widehat{T_k})$ is symmetric by Wu and Chen (1998) and the degrees of freedom are the same, we let

$$P\left(\widehat{T}_{[1]} < -d^*\right) = \frac{\alpha}{2} = P\left(\widehat{T}_{[k]} > d^*\right)$$

Appendix D. Simulated type I error rates in balanced design

Table D: Simulated type I error rates in balanced design

		k=3			k=4			k=5		
		case 1	case 2	case 3	case 1	case 2	case 3	case 1	case 2	case 3
$n = 4$	ANOVA	0.051	0.079	0.083	0.050	0.080	0.084	0.050	0.081	0.085
	BF	0.038	0.051	0.052	0.036	0.051	0.052	0.035	0.050	0.052
	ANOM	0.050	0.079	0.084	0.050	0.089	0.095	0.050	0.097	0.104
	HANOM	0.036	0.032	0.034	0.038	0.034	0.037	0.038	0.036	0.039
$n = 8$	ANOVA	0.050	0.071	0.074	0.050	0.075	0.078	0.050	0.076	0.080
	BF	0.047	0.058	0.059	0.046	0.061	0.063	0.046	0.063	0.064
	ANOM	0.050	0.072	0.076	0.050	0.086	0.090	0.050	0.096	0.102
	HANOM	0.041	0.040	0.037	0.045	0.044	0.039	0.047	0.044	0.042
$n = 12$	ANOVA	0.050	0.069	0.072	0.050	0.073	0.076	0.050	0.075	0.079
	BF	0.049	0.060	0.062	0.049	0.064	0.066	0.048	0.066	0.069
	ANOM	0.050	0.070	0.073	0.050	0.085	0.089	0.050	0.095	0.101
	HANOM	0.042	0.042	0.042	0.046	0.046	0.045	0.047	0.047	0.046
$n = 16$	ANOVA	0.050	0.068	0.070	0.050	0.072	0.074	0.050	0.074	0.077
	BF	0.049	0.061	0.063	0.049	0.065	0.067	0.049	0.067	0.070
	ANOM	0.050	0.069	0.072	0.050	0.084	0.088	0.050	0.094	0.101
	HANOM	0.043	0.043	0.043	0.046	0.045	0.045	0.047	0.047	0.047
$n = 20$	ANOVA	0.050	0.067	0.069	0.050	0.071	0.074	0.050	0.074	0.077
	BF	0.049	0.062	0.063	0.049	0.066	0.068	0.049	0.068	0.071
	ANOM	0.050	0.068	0.071	0.050	0.083	0.087	0.050	0.094	0.100
	HANOM	0.043	0.043	0.043	0.046	0.046	0.046	0.047	0.047	0.047

k is the number of the treatment groups; five sample size allocations from n=4 to n=20; case 1 is the homogeneity and case 2 to case 3 are the heterogeneity situations.

Appendix E. Simulated type I error rates in unbalanced design

Table E.1-E.2 show the empirical type I rate of the classical ANOM test, BF test, $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM with respect to empirical type I error rates corresponding to the different scenarios.

Table E.1: Simulated type I error rate in unbalanced design under homoscedasticity

	sample sizes	5:5:5	10:10:10	20:20:20	3:5:8	5:10:15	5:15:25
$k = 3$	ANOM	0.050	0.050	0.050	0.049	0.050	0.049
	BF	0.042	0.048	0.049	0.047	0.050	0.052
	$\mathcal{P}_1$	0.038	0.041	0.043	0.027	0.033	0.035
	$\mathcal{P}_2$	0.038	0.042	0.043	0.032	0.038	0.039
	sample sizes	5:5:5:5	10:10:10:10	20:20:20:20	3:3:5:5	5:8:10:15	10:20:30:40
$k = 4$	ANOM	0.050	0.050	0.050	0.049	0.050	0.049
	BF	0.042	0.048	0.049	0.047	0.050	0.052
	$\mathcal{P}_1$	0.038	0.041	0.043	0.027	0.033	0.035
	$\mathcal{P}_2$	0.038	0.042	0.043	0.032	0.038	0.039
	sample sizes	5:5:5:5:5	10:10:10:10:10	20:20:20:20:20	3:5:7:10:15	5:10:15:20:25	10:20:30:40:50
	ANOM	0.050	0.050	0.050	0.049	0.050	0.049
	BF	0.042	0.048	0.049	0.047	0.050	0.052
$k = 5$	$\mathcal{P}_1$	0.038	0.041	0.043	0.027	0.033	0.035
	$\mathcal{P}_2$	0.038	0.042	0.043	0.032	0.038	0.039

Table E.2: Simulated type I error rate in unbalanced design under heteroscedasticity

$k = 3$ sample sizes		5:5:5	10:10:10	20:20:20	3:5:8	5:10:15	5:15:25	8:5:3	15:10:5	25:15:5
	ANOM	0.067	0.065	0.063	0.026	0.025	0.016	0.134	0.145	0.181
	BF	0.053	0.057	0.058	0.048	0.053	0.054	0.070	0.064	0.066
V1	$\mathcal{P}_1$	0.038	0.041	0.043	0.027	0.033	0.034	0.027	0.033	0.035
	$\mathcal{P}_2$	0.038	0.042	0.043	0.032	0.038	0.039	0.032	0.038	0.038
	ANOM	0.091	0.085	0.081	0.025	0.023	0.013	0.220	0.228	0.297
	BF	0.063	0.065	0.067	0.058	0.063	0.064	0.080	0.068	0.069
V2	$\mathcal{P}_1$	0.038	0.041	0.043	0.027	0.033	0.034	0.027	0.033	0.035
	$\mathcal{P}_2$	0.038	0.042	0.043	0.032	0.038	0.039	0.032	0.038	0.038
	ANOM	0.107	0.098	0.092	0.027	0.024	0.012	0.275	0.277	0.367
	BF	0.066	0.068	0.072	0.064	0.068	0.070	0.079	0.067	0.067
V3	$\mathcal{P}_1$	0.038	0.041	0.043	0.027	0.033	0.034	0.027	0.033	0.035
	$\mathcal{P}_2$	0.038	0.042	0.043	0.032	0.038	0.039	0.032	0.038	0.038
$k = 4$ sample sizes		5:5:5:5	10:10:10:10	20:20:20:20	3:3:5:5	5:8:10:15	10:20:30:40	5:5:3:3	15:10:8:5	40:30:20:10
	ANOM	0.080	0.078	0.077	0.056	0.034	0.033	0.109	0.143	0.161
	BF	0.056	0.061	0.063	0.046	0.057	0.060	0.058	0.066	0.066
V1	$\mathcal{P}_1$	0.043	0.045	0.045	0.034	0.039	0.043	0.034	0.039	0.043
	$\mathcal{P}_2$	0.043	0.045	0.046	0.036	0.043	0.045	0.036	0.043	0.046
	ANOM	0.120	0.115	0.113	0.079	0.040	0.036	0.183	0.240	0.270
	BF	0.070	0.074	0.078	0.062	0.073	0.077	0.077	0.076	0.077
V2	$\mathcal{P}_1$	0.043	0.044	0.045	0.034	0.039	0.043	0.034	0.039	0.043
	$\mathcal{P}_2$	0.043	0.045	0.046	0.036	0.043	0.045	0.036	0.043	0.046
	ANOM	0.148	0.139	0.133	0.094	0.044	0.042	0.236	0.300	0.338
	BF	0.074	0.079	0.086	0.069	0.081	0.087	0.082	0.076	0.080
V3	$\mathcal{P}_1$	0.043	0.044	0.045	0.034	0.039	0.043	0.034	0.039	0.043
	$\mathcal{P}_2$	0.043	0.045	0.046	0.036	0.043	0.045	0.036	0.043	0.046
$k = 5$ sample sizes		5:5:5:5:5	10:10:10:10:10	20:20:20:20:20	3:5:7:10:15	5:10:15:20:25	10:20:30:40:50	15:10:7:5:3	25:20:15:10:5	50:40:30:20:10
	ANOM	0.088	0.087	0.088	0.030	0.040	0.040	0.169	0.173	0.171
	BF	0.057	0.064	0.066	0.056	0.061	0.063	0.076	0.071	0.068
V1	$\mathcal{P}_1$	0.045	0.046	0.046	0.032	0.038	0.045	0.032	0.038	0.045
	$\mathcal{P}_2$	0.046	0.047	0.047	0.039	0.046	0.047	0.038	0.046	0.047
	ANOM	0.143	0.138	0.138	0.038	0.054	0.053	0.308	0.309	0.302
	BF	0.075	0.081	0.085	0.075	0.082	0.085	0.096	0.084	0.084
V2	$\mathcal{P}_1$	0.045	0.046	0.046	0.032	0.038	0.045	0.032	0.038	0.045
	$\mathcal{P}_2$	0.046	0.047	0.047	0.039	0.046	0.046	0.038	0.046	0.047
	ANOM	0.178	0.171	0.167	0.043	0.062	0.062	0.402	0.398	0.386
	BF	0.080	0.088	0.095	0.087	0.094	0.098	0.098	0.084	0.089
V3	$\mathcal{P}_1$	0.045	0.046	0.046	0.032	0.038	0.045	0.032	0.038	0.045
	$\mathcal{P}_2$	0.046	0.047	0.047	0.039	0.046	0.046	0.038	0.046	0.047

Appendix F. HANOM sample size

Table F.1: HANOM sample size where $\alpha = 0.1$

	$\delta/S_{[k]} = 2.25$ power					$\delta/S_{[k]} = 2.00$ power					$\delta/S_{[k]} = 1.75$ power				
k	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95
3	6	7	7	8	9	7	8	8	9	10	8	9	9	10	12
4	7	7	8	8	9	7	8	9	10	11	8	10	10	12	13
5	7	8	8	9	10	8	9	9	10	12	9	10	11	12	14
6	8	8	9	9	11	9	9	10	11	12	10	11	12	13	15
7	8	9	9	10	11	9	10	11	11	13	10	12	13	14	16
8	8	9	9	10	11	9	10	11	12	13	11	12	13	14	16
9	9	9	10	11	12	10	11	11	12	14	11	13	13	15	17
10	9	10	10	11	12	10	11	12	13	14	12	13	14	15	17
15	10	11	11	12	13	11	12	13	14	16	13	14	15	17	19
20	10	11	12	13	14	12	13	14	15	16	14	15	16	18	20
	$\delta/S_{[k]} = 1.50$ power					$\delta/S_{[k]} = 1.25$ power					$\delta/S_{[k]} = 1.00$ power				
k	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95
3	9	11	12	13	15	12	14	15	17	21	16	20	22	25	31
4	10	12	13	14	17	13	15	17	19	23	18	22	24	28	34
5	11	13	14	15	18	14	17	18	21	25	20	24	27	30	37
6	12	14	15	17	19	15	18	20	22	26	22	26	29	33	39
7	13	14	16	17	20	16	19	21	23	27	23	28	30	34	41
8	13	15	16	18	21	17	20	22	24	28	24	29	31	35	42
9	14	15	17	19	21	17	21	22	25	29	25	30	32	37	44
10	14	16	17	19	22	18	21	23	26	30	26	31	37	38	45
15	16	18	19	21	24	20	24	26	29	33	29	34	38	42	50
20	17	19	20	23	26	22	25	28	30	35	31	37	40	45	52

Table F.2: HANOM sample size where $\alpha = 0.05$

	$\delta/S_{[k]} = 2.25$ power					$\delta/S_{[k]} = 2.00$ power					$\delta/S_{[k]} = 1.75$ power				
k	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95
3	7	8	8	9	10	8	9	9	10	12	9	10	11	12	14
4	8	9	9	10	11	9	10	10	11	13	10	11	12	13	15
5	8	9	10	10	11	9	10	11	12	14	11	12	13	14	16
6	9	10	10	11	12	10	11	12	13	14	12	13	14	15	17
7	9	10	11	11	13	11	11	12	13	15	12	14	15	16	18
8	10	10	11	12	13	11	12	13	14	15	13	14	15	16	19
9	10	11	11	12	13	11	12	13	14	16	13	15	15	17	19
10	10	11	12	12	14	12	13	13	14	16	14	15	16	17	20
15	11	12	13	14	15	13	14	15	16	17	15	17	18	19	22
20	12	13	14	14	16	14	15	16	17	18	16	18	19	20	23
	$\delta/S_{[k]} = 1.50$ power					$\delta/S_{[k]} = 1.25$ power					$\delta/S_{[k]} = 1.00$ power				
k	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95
3	11	13	14	15	18	14	17	18	20	24	20	24	26	30	35
4	12	14	15	17	19	16	18	20	23	27	23	27	29	33	39
5	13	15	16	18	21	17	20	22	24	28	25	29	32	36	42
6	14	16	17	19	22	18	21	23	26	30	27	31	34	38	45
7	15	17	18	20	23	19	23	24	27	31	28	33	36	40	47
8	16	18	19	21	24	20	23	25	28	32	29	34	38	42	49
9	16	18	20	21	25	21	24	26	29	33	30	35	38	43	50
10	17	29	20	22	25	22	25	27	30	35	31	36	40	44	52
15	19	21	22	24	28	24	28	30	33	38	35	40	44	49	57
20	20	22	24	26	29	26	30	32	35	40	38	43	47	52	59

Table F.3: HANOM sample size where $\alpha = 0.01$

	$\delta/S_{[k]} = 2.25$ power					$\delta/S_{[k]} = 2.00$ power					$\delta/S_{[k]} = 1.75$ power				
k	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95
3	10	10	11	12	13	11	12	12	13	15	13	14	15	16	18
4	10	11	12	13	14	12	13	14	15	16	14	15	16	18	20
5	11	12	13	14	15	13	14	15	16	17	15	16	17	19	21
6	12	13	13	14	15	13	15	15	16	18	16	17	18	20	22
7	12	13	14	15	16	14	15	16	17	19	16	18	19	21	23
8	13	14	14	15	17	14	16	17	18	19	17	19	20	22	24
9	13	14	15	15	17	15	16	17	18	20	18	19	21	22	25
10	13	14	15	16	17	15	17	17	19	21	18	20	21	23	25
15	15	16	17	17	19	17	18	19	20	22	20	22	23	25	27
20	16	17	17	18	20	18	19	20	21	23	21	23	24	26	28
	$\delta/S_{[k]} = 1.50$ power					$\delta/S_{[k]} = 1.25$ power					$\delta/S_{[k]} = 1.00$ power				
k	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95	0.70	0.80	0.85	0.9	0.95
3	15	17	19	21	23	20	23	25	27	32	28	34	36	41	47
4	17	19	21	23	25	23	25	28	30	35	33	37	41	45	52
5	18	21	22	24	27	24	28	30	33	37	35	40	44	49	56
6	19	22	23	25	29	26	29	31	35	40	37	43	46	51	59
7	20	23	24	27	37	27	31	32	36	41	39	45	48	53	61
8	21	24	25	28	31	28	32	34	37	42	40	46	50	55	63
9	22	24	26	28	31	29	32	35	38	43	42	48	51	57	65
10	23	25	27	29	33	30	33	36	39	44	43	49	53	58	66
15	25	28	29	31	35	33	37	39	43	48	48	53	58	63	72
20	26	29	30	33	37	35	39	41	45	50	50	56	60	66	75

Appendix G. Numerical Example Data

Table G.1: Effect of Treatments Data

	Treatment 1	Treatment 2	Treatment 3	Treatment 4
Responses	99.891	113.408	115.138	104.133
	99.651	111.645	108.581	116.034
	97.486	107.345	109.565	110.294
	101.120	108.763	110.846	93.026

Table G.2: Reinforcing bars strength data

Brand 1	21.4	13.5	21.1	13.3	18.9	19.2	18.3		
Brand 2	27.3	22.3	16.9	11.3	26.3	19.8	16.2	25.4	
Brand 3	18.7	19.1	16.4	15.9	18.7	20.1	17.8		
Brand 4	19.9	19.3	18.7	20.3	22.8	20.8	20.9	23.6	21.2

Table G.3: Bacterial killing ability data

Solvent 1	98.29	95.58	97.39	98.43	95.41	96.87	98.66	95.29	98.3	98.59
	95.28	98.89	97.41	97.52	94.63	96.44				
Solvent 2	98.47	94.67	95.68	97.52	97.52	92.43	96.86	97.57	94.03	97.21
	98.06	98.2	92.09	93.63	94.61	95.28	94.14	93.09	94.47	
Solvent 3	92.29	91.57	94.13	98.05	91.43	93.38	97.55	98.93	94.96	93.61
	92.52	94.2	94.21	92.65	92.68	98.79	92.42	97.48	94.34	96.89
	93.86	95.31	93.76	96.27						
Solvent 4	96.06	96.33	95.9	96.71	98.38	97.97	97.13	97.42	97.65	98.05
	98.11									

Table G.4: Mutagenicity data

Treatment group	Mean	Standard deviation	Sample size
Vehicle control	2.571	1.272	7
30 mg/kg	3.800	1.095	5
50 mg/kg	6.200	1.483	5
75 mg/kg	14.000	3.937	5
100 mg/kg	20.000	4.062	5

Appendix H. Details Proof for the critical values of the simultaneous confidence interval of single-stage sampling procedure for MCC

Theorem 1: The constants a_i's are the solution of

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + a_i)\right] f_{v_1}(t_1)\, dt_1 = P^*,$$

where $F_{v_i}(\cdot)$ and $f_{v_i}(\cdot)$ are the distribution and density functions respectively of a student's t random variables with $v_i (= n_i - 2), i = 1, \ldots, k$, degrees of freedom.

Proof:

$$\begin{aligned}
&P(\mu_i - \mu_1 \in I_L, i = 2, \ldots, k) \\
=&P(\mu_i - \mu_1 \geq \tilde{X}_i - \tilde{X}_1 - a_i\sqrt{z^*}, i = 2, \ldots, k) \\
=&P(\tilde{X}_i - \mu_i \leq \tilde{X}_1 - \mu_1 + a_i\sqrt{z^*}, i = 2, \ldots, k) \\
=&P(\frac{\tilde{X}_i - \mu_i}{\sqrt{z^*}} \leq \frac{\tilde{X}_1 - \mu_1}{\sqrt{z^*}} + a_i, i = 2, \ldots, k) \\
=&P(T_i \leq T_1 + a_i, i = 2, \ldots, k) \text{ by Equation}(3.6) \\
=&P(T_i \leq t_1 + a_i, i = 2, \ldots, k | T_1 = t_1) P(t_1) \\
=&P(T_i \leq t_1 + a_i, i = 2, \ldots, k) P(t_1) \ (\because T_1 \perp T_i) \\
=&\prod_{i=2}^{k} P(T_i \leq t_1 + a_i) P(t_1) \ (\because T_2, \ldots, T_k \text{ are independent})
\end{aligned}$$

$$= \int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + a_i) \right] f_{v_1}(t_1) dt_1$$

Theorem 2: The constants b_i's are the solution of

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[1 - F_{v_i}(t_1 - b_i) \right] f_{v_1}(t_1)\, dt_1 = P^*.$$

Proof:This proof is similar to that of Theorem 1. Note that the property of symmetry for t distribution, $1 - F_{v_i}(t_1 - b_i) = F_{v_i}(t_1 + b_i)$, Therefore, Equation (3.9) becomes

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + b_i) \right] f_{v_1}(t_1)\, dt_1 = P^*.$$

Theorem 3: The constants a_i's and b_i's are the solution of

$$\int_{-\infty}^{\infty} \prod_{i=2}^{k} \left[F_{v_i}(t_1 + a_i) - F_{v_i}(t_1 - b_i) \right] f_{v_1}(t_1)\, dt_1 = P^*.$$

Proof: This proof is similar to that of Theorem 1. From Equation (3.10), if we choose all a_i's to go infinite, it reduces the equation down to Equation (3.9). Similarly, if we choose all b_i's to go infinite, it reduces the equation down to Equation (3.8).

Appendix I. Details Proof for the critical values of the simultaneous confidence interval and power function of modified single-stage sampling procedure for MCC

Theorem 4: The critical value c^* needs to satisfy

$$\int_{-\infty}^{\infty}\left[\prod_{i=2}^{k}F_{v_i}(t_1\sqrt{\frac{n_i}{n_1}}+c^*)\right]f_{v_1}(t_1)\,dt_1=1-\alpha.$$

Proof:

$$\begin{aligned}
&P(\mu_i-\mu_1\in I_L, i=2,\ldots,k)\\
=&P(\mu_i-\mu_1\geq \tilde{X}_i-\tilde{X}_1-c^*\frac{S_{[k]}}{\sqrt{n_i}}, i=2,\ldots,k)\\
=&P(\frac{\tilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}}\leq \frac{\tilde{X}_1-\mu_1}{S_{[k]}/\sqrt{n_i}}+c^*, i=2,\ldots,k)\\
=&P(\frac{\tilde{X}_i-\mu_i}{S_{[k]}/\sqrt{n_i}}\leq \frac{\tilde{X}_1-\mu_1}{S_{[k]}/\sqrt{n_1}}\sqrt{\frac{n_i}{n_1}}+c^*, i=2,\ldots,k)\\
=&P(T_i\leq T_1\sqrt{\frac{n_i}{n_1}}+c^*, i=2,\ldots,k)\\
=&P(T_i\leq t_1\sqrt{\frac{n_i}{n_1}}+c^*, i=2,\ldots,k)P(t_1)\\
=&\prod_{i=2}^{k}P(T_i\leq t_1\sqrt{\frac{n_i}{n_1}}+c^*)P(t_1)\\
=&\int_{-\infty}^{\infty}\left[\prod_{i=2}^{k}F_{v_i}(t_1\sqrt{\frac{n_i}{n_1}}+c^*)\right]f_{v_1}(t_1)\,dt_1=1-\alpha
\end{aligned}$$

Theorem 5: The power function for the modified single-stage sampling procedure for MCC is as follows.

$$
\begin{aligned}
& P\left\{\mu_i-\mu_1 \notin I_L \text{ and } \exists i, i=2, \ldots, k, \mu_i-\mu_1>\delta\right\} \\
=& 1-P\left(T_i<T_1 \sqrt{\frac{n_i}{n_1}}+\frac{\delta}{S_{[k]} / \sqrt{n_i}}, i=2, \ldots, k\right) \\
=& 1-\int_{-\infty}^{\infty}\left(\prod_{i=2}^{k}\left[F_{v_i}(t_1 \sqrt{\frac{n_i}{n_1}}+\frac{\delta}{S_{[k]} / \sqrt{n_i}})\right]\right) f_{v_1}(t_1)\, dt_1 \geq 1-\beta
\end{aligned}
$$

Proof:

$$
\begin{aligned}
& P\left\{\mu_i-\mu_1 \notin I_L \text{ and } \exists i, i=2, \ldots, k, \mu_i-\mu_1>\delta\right\} \\
=& 1-P\left\{\mu_i-\mu_1>\widetilde{X}_i-\widetilde{X}_1-c^* \frac{S_{[k]}}{\sqrt{n_i}} \text{ and } c^* \frac{S_{[k]}}{\sqrt{n_i}}>\delta, i=2, \ldots, k\right\} \\
=& 1-P\left(\mu_i-\mu_1>\widetilde{X}_i-\widetilde{X}_1-\delta, i=2, \ldots, k\right) \\
=& 1-P\left(\frac{\tilde{X}_i-\mu_i}{S_{[k]} / \sqrt{n_i}}<\frac{\tilde{X}_1-\mu_1}{S_{[k]} / \sqrt{n_i}}+\frac{\delta}{S_{[k]} / \sqrt{n_i}}, i=2, \ldots, k\right) \\
=& 1-P\left(\frac{\tilde{X}_i-\mu_i}{S_{[k]} / \sqrt{n_i}}<\frac{\tilde{X}_1-\mu_1}{S_{[k]} / \sqrt{n_1}} \sqrt{\frac{n_i}{n_1}}+\frac{\delta}{S_{[k]} / \sqrt{n_i}}, i=2, \ldots, k\right) \\
=& 1-P\left(T_i<T_1 \sqrt{\frac{n_i}{n_1}}+\frac{\delta}{S_{[k]} / \sqrt{n_i}}, i=2, \ldots, k\right) \\
=& 1-\int_{-\infty}^{\infty}\left(\prod_{i=2}^{k}\left[F_{v_i}(t_1 \sqrt{\frac{n_i}{n_1}}+\frac{\delta}{S_{[k]} / \sqrt{n_i}})\right]\right) f_{v_1}(t_1)\, dt_1 \geq 1-\beta
\end{aligned}
$$

Appendix J. Simulated familywise error rates for MCC

Table J: Simulated familywise error rates for MCC

		k=3			k=4			k=5		
		case 1	case 2	case 3	case 1	case 2	case 3	case 1	case 2	case 3
$n_1 = 4$	Dunn	0.048	0.011	0.181	0.048	0.004	0.222	0.050	0.002	0.253
	PI	0.050	0.050	0.051	0.050	0.052	0.047	0.050	0.051	0.046
	SS	0.050	0.050	0.051	0.047	0.047	0.048	0.050	0.049	0.051
	MSS	0.050	0.051	0.051	0.050	0.049	0.047	0.050	0.050	0.050
$n_1 = 8$	Dunn	0.051	0.018	0.135	0.051	0.010	0.166	0.049	0.006	0.197
	PI	0.050	0.050	0.051	0.050	0.052	0.047	0.050	0.051	0.046
	SS	0.052	0.051	0.050	0.050	0.051	0.051	0.050	0.051	0.050
	MSS	0.051	0.050	0.052	0.051	0.051	0.051	0.051	0.053	0.051
$n_1 = 12$	Dunn	0.050	0.022	0.115	0.049	0.014	0.145	0.048	0.010	0.170
	PI	0.051	0.048	0.046	0.050	0.050	0.047	0.050	0.051	0.048
	SS	0.048	0.048	0.051	0.049	0.052	0.050	0.050	0.050	0.049
	MSS	0.050	0.049	0.051	0.050	0.052	0.051	0.051	0.049	0.050
$n_1 = 16$	Dunn	0.050	0.026	0.107	0.048	0.017	0.136	0.051	0.012	0.161
	PI	0.051	0.051	0.052	0.050	0.047	0.052	0.050	0.047	0.051
	SS	0.053	0.048	0.053	0.048	0.049	0.050	0.051	0.052	0.051
	MSS	0.050	0.052	0.052	0.051	0.052	0.049	0.051	0.052	0.051
$n_1 = 20$	Dunn	0.050	0.026	0.102	0.050	0.020	0.130	0.050	0.014	0.148
	PI	0.050	0.050	0.049	0.050	0.047	0.048	0.050	0.050	0.050
	SS	0.049	0.048	0.050	0.052	0.049	0.052	0.052	0.052	0.050
	MSS	0.051	0.051	0.051	0.051	0.051	0.052	0.051	0.053	0.049

Simulated FWEs for Dunnett test (Dunn), plug-in procedure (PI), the single-stage sampling procedure (SS) and the modified single-stage sampling procedure (MSS), respectively; k is the number of the treatment groups; five sample size allocations, and also, unbalanced sample sizes per allocations, i.e. $n_i = n_{i-1} + 2$ $(i = 2, \ldots, k)$; case 1 is the homogeneity and case 2 to case 3 are the heterogeneity situations.

Appendix K. R Shiny usage for the single-stage sampling procedures

We developed three R shiny interfaces for HANOM with balanced design and unbalanced design and MCC under heteroscedasticity, respectively. The URL of the interface for HANOM with balanced design can be accessed at `https://weiming3524.shinyapps.io/HANOM/`, as shown in Figure K.1. The interface can provide the describe statistics, the summary statistics, the p-value for HANOM, and the HANOM chart, respectively. Under unbalanced design, another URL of the interface for HANOM can be accessed at `https://weiming3524.shinyapps.io/HANOM_UB/`, as shown in Figure K.2. This interface can also provide HANOM chart and the summary statistics by uploading a dataset. It can also calculate the power of the HANOM test and the necessary sample size. The third R shiny interface, as shown in Figure K.3, can simulate the critical values of the single-stage sampling procedure for multiple comparison with a control under heteroscedasticity. The one-sided and two-sided critical values can be obtained after inputting parameters on the interface. The URL are `https://weiming3524.shinyapps.io/mc-control/`.

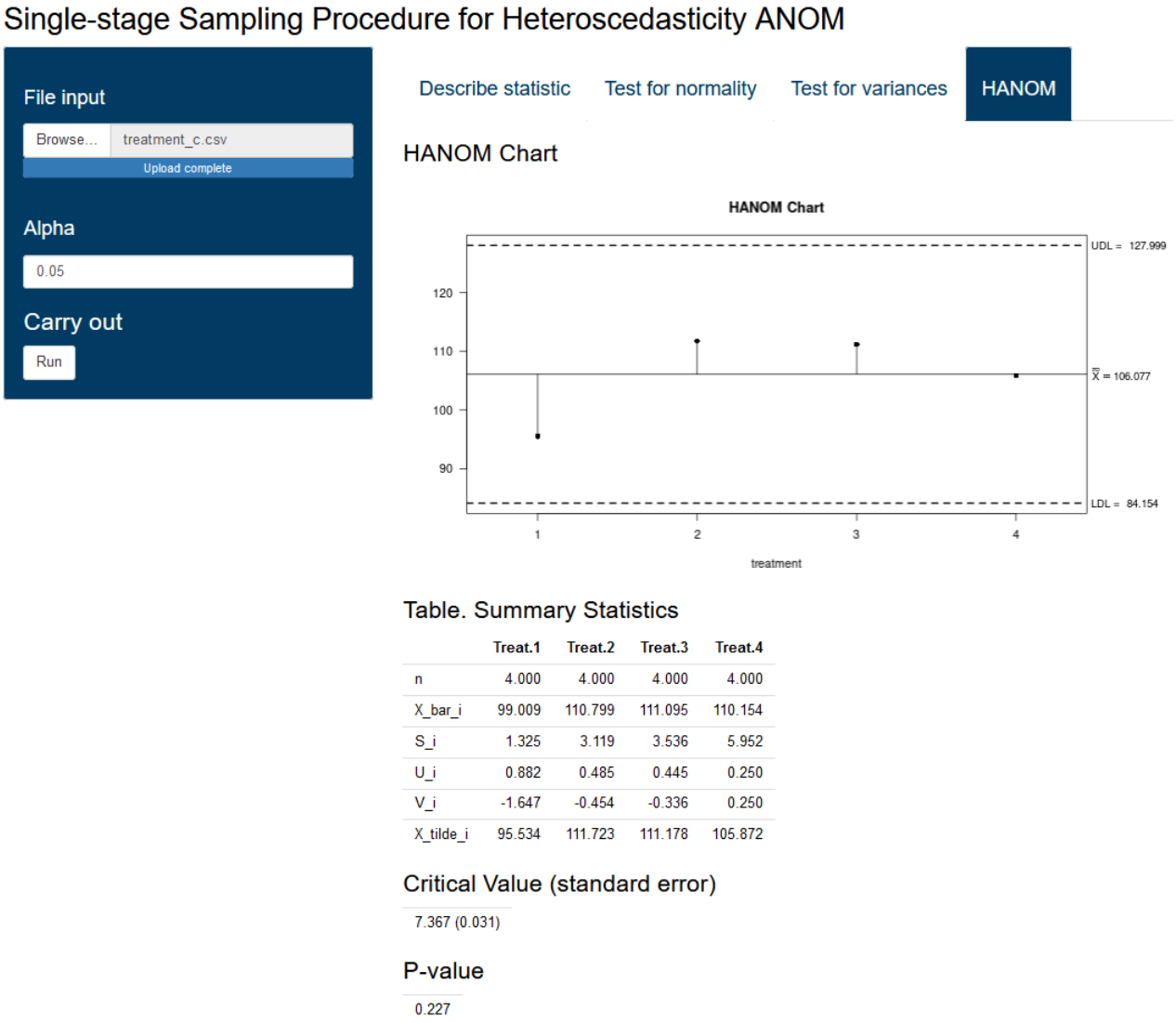


Figure K.1: The *R* Shiny interface for HANOM by using the single-stage sampling procedure

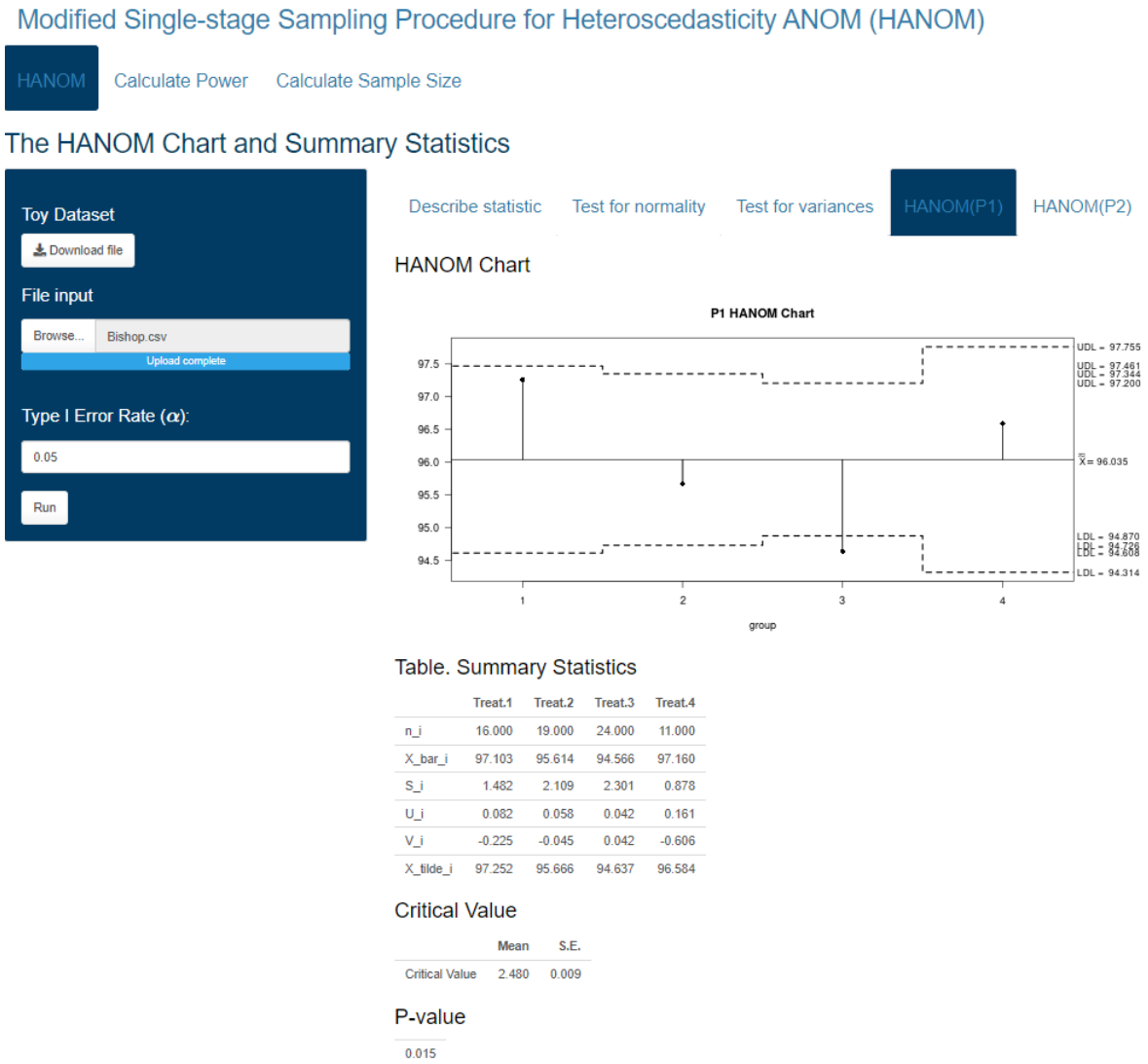


Figure K.2: The R Shiny interface for HANOM by using the single-stage sampling procedure $\mathcal{P}_1$ and $\mathcal{P}_2$

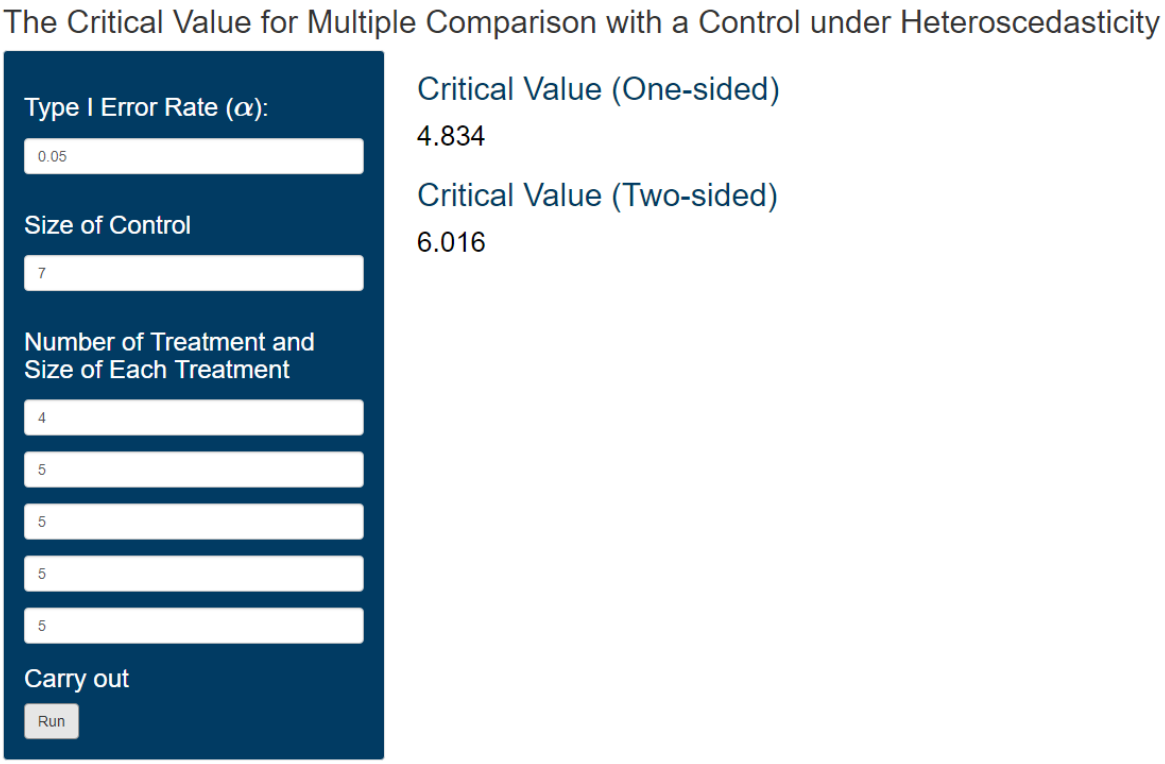


Figure K.3: The R shiny interface for multiple comparison with a control under heteroscedasticity by using a single-stage sampling procedure