Brazilian Journal of Probability and Statistics
2022, Vol. 36, No. 4, 704–724
https://doi.org/10.1214/22-BJPS550

Single-stage sampling procedure for heteroscedasticity analysis of means

Wei-Ming Wang[1,a], Chun-Che Wen[2,b], Tsai-Hsiang Hung[3,c], Junjiang Zhong[4,d] and Miin-Jye Wen[1,5,e]

[1]*Department of Statistics and Institute of Data Science, College of Management, National Cheng Kung University, Tainan, Taiwan,* [a]*weiming3524@gmail.com*
[2]*Department of Public Health Sciences, College of Graduate Studies, Medical University of South Carolina, Charleston, South Carolina, USA,* [b]*chechewen18@gmail.com*
[3]*Digital, Data & Technology, Cathay Financial Holdings Co., Ltd., Taipei, Taiwan,* [c]*hungth18@gmail.com*
[4]*School of Mathematics and Statistics, Xiamen University of Technology, Xiamen, China,* [d]*jjzhong@xmut.edu.cn*
[5]*Institute of International Management, College of Management, National Cheng Kung University, Tainan, Taiwan,* [e]*mjwen@ncku.edu.tw*

Abstract. The analysis of means (ANOM) is a method that can compare the mean of each treatment to the overall mean. According to the graphical result of a statistical data analysis, we can specify which one is different from another. One of the assumptions of the classical ANOM model is that the variances are equal. However, it is not always true for the practice. To solve unknown and unequal population variances, Nelson and Dudewicz (2002) proposed a two-stage sampling procedure. However, additional samples need to be added in the second stage of the two-stage sampling procedure, so it is not practical all the time due to limited time and insufficient budget. Thus, under heteroscedasticity, we applied Chen and Lam's (1989) single-stage sampling procedure to solve the drawback of the two-stage sampling procedure. In addition, we also provided an illustrative example and critical values for practical uses. In order to make the procedure user-friendly, we built an interface by using *R* Shiny.

1 Introduction

Comparisons of k populations or treatment groups are required in many situations in the pharmaceutical research and other fields. Analysis of variance (ANOVA) is a commonly used statistical method to test whether the means of all treatments are equal or not. When the null hypothesis is rejected in a one way ANOVA, it indicates that at least one of the treatment mean is statistically significant. In order to find the difference of among treatment groups, we do the post hoc multiple comparisons. There are quite a few different methods for performing multiple comparisons, such as the Tukey method, Scheffé's test, Bonferroni method, Newman-Keuls method, Dunnett method, and so on. However, it is hard for practitioners to select these procedures in practical. In particular, various post-hoc results are often different and sometimes contradict to ANOVA. Analysis of means (ANOM) can be considered as a useful alternative to the ANOVA to solve the above problems. 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. ANOM and ANOVA are different when there are more than two treatments although they are equivalent when there are only two treatments. 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

Key words and phrases. Heteroscedasticity, single-stage sampling procedure, analysis of means.
Received September 2021; accepted August 2022.

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 that displays the upper and lower decision limits, each treatment means and the overall mean. If a treatment mean falls outside the decision limits, it indicates that this particular mean is statistically different from the overall mean. As such, ANOM has the advantages of providing a graphical display that allows one to easily evaluate both the statistical and practical significance of the differences Nelson and Dudewicz (2002). 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 comprehensive review may refer to Rao (2005).

Nelson (1982) proposed the exact critical values for ANOM in balanced designs. The classical assumption of an ANOM model is the same as that of an ANOVA model, that is, the observations of treatment i $(i = 1, \ldots, k)$ are independent and normally distributed with means and equal variances. When the variances are actually unequal, the ANOVA F-test can result in poor power and its p-value can be biased due to their dependence on the unknown variances (Krutchkoff (1988); Weerahandi (1987)). For unknown and unequal variances, Nelson and Dudewicz (2002) developed a two-stage sampling procedure which was originally proposed by Stein (1945) to solve ANOM under heteroscedasticity (HANOM). Furthermore, Dudewicz and Nelson (2003) showed that while comparing HANOM with heteroscedastic ANOVA, the differences in powers are small, and HANOM has the graphical advantages as ANOM. Although the two-stage sampling procedure can solve problems caused by unequal variances, it has some drawbacks. It may not be practical in the real world applications, because extra time and work are necessary for taking additional samples for the second stage. Sometimes, only one sample procedure is available in an experiment due to limited time, budget, etc. Therefore, a single-stage sampling procedure originally developed by Chen and Lam (1989) to construct the confidence interval of the largest normal mean under heteroscedasticity was designed to work out these problems. Wen and Chen (1994) extended to apply the single-stage sampling procedure for multiple comparison with the largest normal mean and a control. The single-stage sampling procedure was extended to one-way and two-way layout ANOVA under heteroscedasticity by Chen and Chen (1998). For testing the hypotheses of the equality of means in the ANOVA setting, general one-stage and two-stage sampling procedures were presented by Chen (2001). Chen, Chen and Chang (2004) used an one-stage procedure for testing the order treatment means in the one-way layout. Wen, Huang and Zhong (2017) proposed a single-stage sampling procedure of the t best populations under heteroscedasticity. Wen, Wen and Wang (2022) also used the single-stage sampling procedure for multiple comparison with a control with unequal sample sizes. In this paper, we will discuss ANOM in one-way layout model under heteroscedastic situation by a single-stage sampling procedure.

2 Methods

The classical two-stage sampling procedure for HANOM that proposed by Nelson and Dudewicz (2002) that uses a weighted sample mean to estimate the population mean. The first step in performing HANOM is to take an initial sample n_0 (≥ 2) and an total sample size n_i that is defined as $\max\{n_0 + 1, [\![(\frac{\omega}{\delta})^2 S^2_{i,n_0}]\!] + 1\}$ where the value of ω is an arbitrary constant to be chosen by specify the significance α, power γ and the effect size δ. Afterwards, calculate the weighted sample mean $\widetilde{\overline{X}}_i$ by the initial sample n_0 and the remaining $n_i - n_0$ for each i $(i = 1, \ldots, k)$ population. As a result, the decision lines of a HANOM are as:

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

where $\widetilde{\overline{X}}$ is the overall mean of all weighted sample means, and $H(\alpha; k; n_0 - 1)$ is the critical value that can be found in Nelson and Dudewicz (2002). The detailed steps is given in Appendix A. If the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ is true, then all $\widetilde{\overline{X}}_i$ lie between the decision lines; otherwise, the null hypothesis must be rejected if any of them fall outside.

Although the two-stage sampling procedure can solve ANOM under heteroscedasticity, it is design-oriented, which means that a final sample size must be determined with a significance α, power γ, and effect size δ between any two treatment means. The second stage needs extra samples, which might be large due to heterogeneous variances. It is also difficult to determine the final sample size and may not be practical in the real world applications. As a result, we used Chen and Lam's (1989) single-stage sampling procedure to eliminates the disadvantages of a two-stage sampling procedure and deals with the HANOM.

2.1 Single-stage sampling procedure

Let X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n)$ be random samples from a normal distribution with unknown mean, μ_i, and unknown and unequal variance, σ_i^2. Bishop (1976) first introduced the idea of splitting a sample into two portions in the ANOVA problem. Among each population i $(i = 1, \ldots, k)$, we randomly selected $n - 1$ observations to define the sample mean and sample variance, respectively, by

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

The sample size, n (≥ 3), of each population i $(i = 1, \ldots, k)$ is split into two parts: the first set of $n - 1$ observations and the remaining observation to calculate the weighted sample mean

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

where

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

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

$$(n-1)U_i + V_i = 1 \tag{2}$$

$$(n-1)U_i^2 + V_i^2 = \frac{S_{[k]}^2}{nS_i^2} \tag{3}$$

where $S_{[k]}^2$ is the maximum of $\{S_1^2, \ldots, S_k^2\}$. The advantage of Equation (2) implies $\sum_{j=1}^{n-1} U_i + V_i = \sum_{j=1}^{n} \omega_{ij} = 1$ that makes the weighted sample mean, $\widetilde{X}_i$ in (1), an unbiased estimate of the population mean, μ_i, and Equation (3) implies $\sum_{j=1}^{n} \omega_{ij}^2 S_i^2 = S_{[k]}^2/n$ to eliminate the influence of the unknown and unequal variances by Wen and Chen (1994). Such a choice pushes the weights U_i and V_i closer to $1/n$ which is a traditional way while calculating the sample mean. Another reason is that it is optimal in the sense that the Student's t distribution has the smallest critical value for a fixed level of significance. Based on the Equations (2) and (3), the solutions of the weights of U_i and V_i are derived as

$$U_i = \frac{1}{n} + \frac{1}{n}\sqrt{\frac{1}{n-1}\left[S_{[k]}^2/S_i^2 - 1\right]} \tag{4}$$

and

$$V_i = \frac{1}{n} - \frac{1}{n}\sqrt{(n-1)\left[S_{[k]}^2/S_i^2 - 1\right]}. \tag{5}$$

Hence, the sampling distribution of weighted sample mean, $\widetilde{X}_i$ $(i = 1, \ldots, k)$, is denoted by

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

In addition, given the sample variance, S_i^2 $(i = 1, \ldots, k)$, the transformation of weighted sample mean

$$T_i = \frac{\widetilde{X}_i - \mu_i}{\sqrt{\sum_{j=1}^{n} \omega_{ij}^2\ S_i^2}} = \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n}}, \quad i = 1, \ldots, k. \tag{6}$$

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 $n - 2$ by Wen and Chen (1994). Furthermore, $T_1, \ldots, T_k$ are mutually independent (for more detail proofs, see Wen and Chen (1994)).

2.2 The one-way layout model

The model is how we conduct the heteroscedasticity analysis of means (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 $\varepsilon_{ij} \sim N(0,\ \sigma_i^2)$, $i = 1, \ldots, k$; $j = 1, \ldots, n$. That is, we need to test if the k treatments which are balanced have an effect on the means, μ_i, $i = 1, \ldots, k$. Thus, the null hypothesis is

$$H_0 : \mu_1 = \cdots = \mu_i = \cdots = \mu_k = \bar{\mu},$$

in contrast to the alternative hypothesis in which one or more means are different from the overall mean $\bar{\mu} = \sum_{i=1}^{k} \mu_i/k$.

The main characteristic in HANOM is facilitating decisions made by a HANOM decision chart (Nelson and Dudewicz (2002)). When there are differences among the treatment means, HANOM can also indicate which treatment means 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 is an overall mean of all weighted sample means.

(ii) Lower Decision Line (LDL) is

$$\overline{\widetilde{X}} - d^* \cdot \frac{S_{[k]}}{\sqrt{n}} \tag{7}$$

(iii) Upper Decision Line (UDL) is

$$\overline{\widetilde{X}} + d^* \cdot \frac{S_{[k]}}{\sqrt{n}} \tag{8}$$

where $d^* = d^*(\frac{\alpha}{2}, k, df) > 0$ is the critical value depending 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 section.

If the null hypothesis is true, all of the k treatment means fall between LDL and UDL. Therefore, all the weighted sample means, $\widetilde{X}_i$, should be close to the weighted overall mean, $\overline{\widetilde{X}}$. If any of the weighted sample means falls outside LDL or UDL, we will reject the null hypothesis, H_0.

2.3 Determination of critical values

Following the idea of the HANOM Procedure proposed by Nelson and Dudewicz (2002), we define the rejection region of treatment i as

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

Therefore, the level of significance, α, for HANOM is

$$\begin{aligned}
1-\alpha =& P\left(\left|\frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}}\right| \le d^*,\ i = 1, \ldots, k\right) \\
=& P\left(-d^* \le \frac{\widetilde{X}_i - \overline{\widetilde{X}}}{S_{[k]}/\sqrt{n}} \le d^*,\ i = 1, \ldots, k\right) \\
=& P\left(-d^* \le \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n}} - \frac{\overline{\widetilde{X}} - \bar{\mu}}{S_{[k]}/\sqrt{n}} + \frac{\mu_i - \bar{\mu}}{S_{[k]}/\sqrt{n}} \le d^*,\ i = 1, \ldots, k\right). \qquad (9)
\end{aligned}$$

From Equation (6), we have $T_i = \frac{\widetilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n}}$, $i = 1, \ldots, k$ in Equation (9) are Student's t distribution with degree of freedom $n-2$, while $\bar{T} = \sum_{i=1}^{k} T_i/k = \frac{\overline{\widetilde{X}} - \bar{\mu}}{S_{[k]}/\sqrt{n}}$. Under the null hypothesis, $H_0 : \mu_1 = \cdots = \mu_k = \bar{\mu}$, the term $\frac{\mu_i - \bar{\mu}}{S_{[k]}/\sqrt{n}}$ in Equation (9) equals zero and $1-\alpha$ will be rewritten by

$$\begin{aligned}
1-\alpha &= P\left(-d^* \le T_i - \bar{T} \le d^*,\ i = 1, \ldots, k\right) \\
&= P\left(-d^* \le \widetilde{T}_i \le d^*,\ i = 1, \ldots, k\right) \qquad (10)
\end{aligned}$$

where $\widetilde{T}_i = T_i - \bar{T}$, $i = 1, \ldots, k$, and $\widetilde{T}_{[1]}$ is the minimum of $\{\widetilde{T}_1, \ldots, \widetilde{T}_k\}$ and $\widetilde{T}_{[k]}$ is the maximum of $\{\widetilde{T}_1, \ldots, \widetilde{T}_k\}$. Here, we can rewrite Equation (10) as follows.

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

Here, we take

$$P\left(\widetilde{T}_{[1]} < -d^*\right) + P\left(\widetilde{T}_{[k]} > d^*\right) \le \alpha \qquad (11)$$

Since $(\widetilde{T}_1, \ldots, \widetilde{T}_k)$ is symmetric by Wu and Chen (1998), we let

$$P\left(\widetilde{T}_{[1]} < -d^*\right) = \frac{\alpha}{2} = P\left(\widetilde{T}_{[k]} > d^*\right) \qquad (12)$$

It is difficult to find the exact joint distribution of $(\widetilde{T}_1, \ldots, \widetilde{T}_k)$. Therefore, the Monte-Carlo simulation is used 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, we sort these $\widetilde{T}_i$, $i = 1, \ldots, k$ and select

the maximum value $\widetilde{T}_{[k]}$ at each run. After 100,000 independent runs, the probability of inclusion being at least $1-\alpha/2$ of the distribution of $\widetilde{T}_{[k]}$ can be approximated by selecting the cumulated number out of the ranking list of the 100,000 values of $\widetilde{T}_{[k]}$. 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 A1–A3 in Appendix B.

2.4 Computing the p-value

The p-value for HANOM can be identified as a value that the weighted sample mean, $\widetilde{X}_i$, is greatest distance away from the weighted overall mean, $\overline{\widetilde{X}}$ (Nelson, Wludyka and Copeland (2005)). 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

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

Using the R simulation program, we increase p by 0.001 from 0 to 1. By running the HANOM procedure, we try to find the $(100\frac{p}{2})$ quantile of d^* distribution and could find the minimized p as the p-value for HANOM that is

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

We can obtain control chart, critical value, and p-value from the above procedure. In order to use the single-stage sampling procedure for HANOM easily, we provided an interface by R shiny to obtain the critical value and the p-value of the procedure for HANOM include the HANOM chart. Figure A1 displays the user interface and the URL for the interface can be accessed at https://weiming3524.shinyapps.io/HANOM/.

3 Simulations

In the simulation study, ANOVA, Brown-Forsythe test (1974), classical ANOM, and the single-stage sampling procedure for HANOM were compared with respect to type I error rate (α) for validating their quality. The difference of number of treatments, variance scenarios and sample size allocations all have an effect on the quality of a test. We took $k = 3, 4, 5$ treatments under normal distributions and other distributions, respectively, in order to be compared by empirical type I error rates; the nominal level is $\alpha = 0.05$. The empirical type I error rates for ANOVA and Brown-Forsythe test (BF) are estimated by determining the total number of null hypotheses that are incorrectly rejected. 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 falls outside the UDL or LDL. 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. To obtain variability, we repeated this procedure 200 simulations. On the other hand, specific constant numbers with standard deviation form (Δ) were added to the one or more groups to supply at least one of the means that falls outside the decision lines. This process was repeated 5000 times, and the number of trials (r) in which at least one of the treatment means which fell outside the decision lines was determined. Afterwards, the test powers were obtained by dividing this number by the total number of trials 5000. That is, test powers were determined as $r/5000$.

Similar to Hasler's (2016) scenarios, five sample size allocations were chosen, that is, $n = 4, 8, 12, 16, 20$. In Scenario 1, the expected value of all treatments was the same, $\mu_i = 100$ $(i = 1, \ldots, k)$ under normal distribution. Three different variance scenarios were considered as follows:

- 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): $\sigma_i = 10 + 60(i-1)/(k-1)$, $i = 1, \ldots, k$.

In Scenario 2, we set a chi-square distribution with 18 degree of freedom, a beta distribution with $a = 10$ and $b = 10$, and a bimodal distribution of the mixture of $t(10)$ and $N(1, 1)$ distributions to evaluate the robustness under different distributions. In a Monte Carlo simulation, we generate multiple scenarios to get the empirical type I error rate and power of a test. Several experimental conditions were included the homogeneity and two heteroscedasticity cases are given in detail in Table 1. It is noted that BF test is available by the *R* package *onewaytests* (Dag, Dolgun and Konar (2021)) and 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)).

Table 1 *Experimental conditions*

	Number of treatment groups		
	$k = 3$	$k = 4$	$k = 5$
Scenario 1		$N(\mu, \sigma^2)$	
Variance ratio			
case 1	100:100:100	100:100:100:100	100:100:100:100:100
case 2	100:900:2500	100:544:1344:2500	100:400:900:1600:2500
case 3	100:1600:4900	100:900:2500:4900	100:625:1600:3025:4900
Effect size (Δ)			
$\Delta 1$	5:(−5):0	5:(−5):0:0	5:(−5):0:0:0
$\Delta 2$	10:(−10):0	10:(−10):0:0	10:(−10):0:0:0
$\Delta 3$	10:0:(−10)	10:0:0:(−10)	10:0:0:0:(−10)
Scenario 2		$\chi^2(\nu)$, $\beta(a, b)$, $0.5t(r) + 0.5N(\theta, \tau^2)$	
Variance ratio			
case 1	1:1:1	1:1:1:1	1:1:1:1:1
case 2	1:1:4	1:1:1:4	1:1:1:1:4
case 3	1:4:7	1:3:5:7	1:3:5:7:9
Effect size (Δ)			
$\Delta 1$	0.5:−0.5):0	0.5:(−0.5):0:0	0.5:(−0.5):0:0:0
$\Delta 2$	1:(−1):0	1:(−1):0:0	1:(−1):0:0:0
$\Delta 3$	1:0:(−1)	1:0:0:(−1)	1:0:0:0:(−1)

In the first scenario, data are all generated from normal distribution $N(\mu, \sigma^2)$ under three variance cases and effect sizes. However, in the second scenario, we generate the data from chi-square distribution $\chi^2(\nu)$, beta distribution $\beta(a, b)$, and mixing distribution $0.5t(r) + 0.5N(\theta, \tau^2)$ under three variance cases and effect sizes.

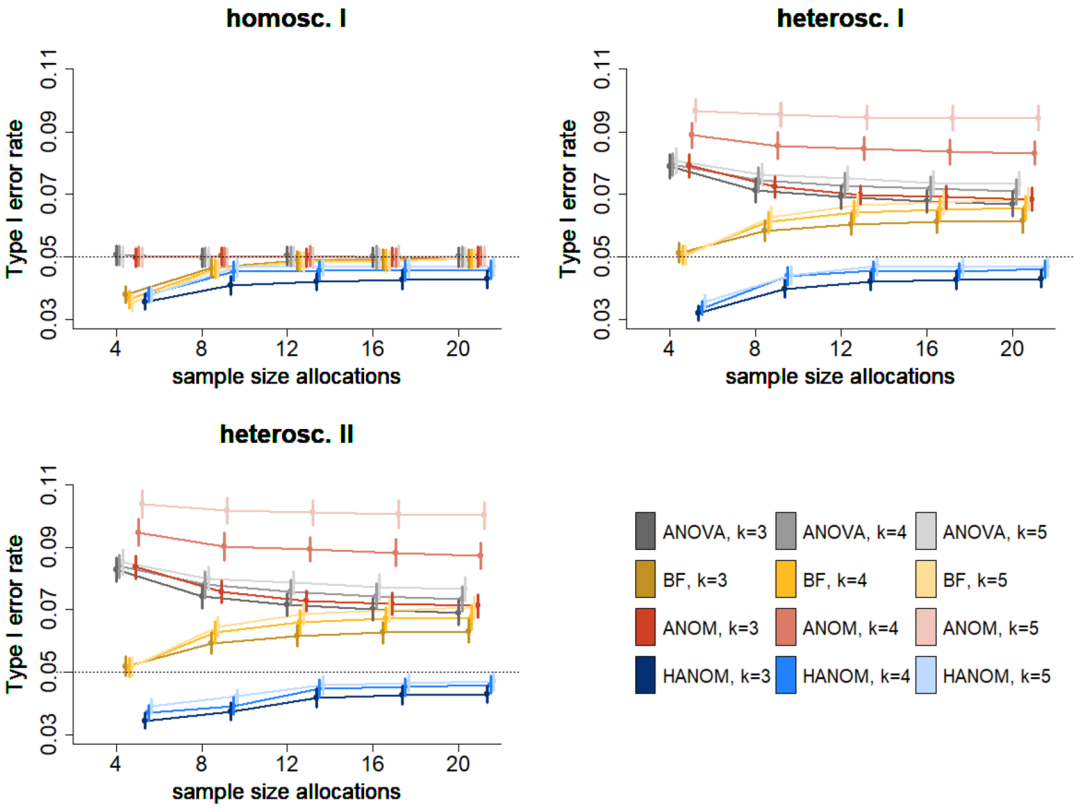

Figure 1 *Simulated type I error rate under normal distribution for several groups, sample sizes allocations and variance scenarios, respectively.*

Figure 1 provides a visual representation of the simulation results with respect to empirical type I error rates with different sample size allocations correspond to the different variance scenarios under normal distributions. The details can be found in Table A4 in Appendix E. Under homogeneity of variances, the results reveal that the ANOVA, BF, ANOM and HANOM tests displayed similar type I error rates. All tests generally display type I error rates close to 5%, but BF and HANOM appear to be a little conservative when the sample size is small. Although the single-stage sampling procedure for HANOM could be conservative in heteroscedastic scenarios due to the small sample size ($n = 4$), the empirical type I error rates ranging between 0.032 and 0.047 were not far 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 is 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.

The simulation results of type I error rates for different distributions can be found in Figures S1–S3 in Supplementary materials (Wang et al. (2022)). Regarding the beta distributions, we can find that HANOM shows conservative results and maintained under 5% regardless of the variation situations. Although other methods can maintained the nominal level under homogeneity, BF is slightly more liberal, while ANOVA and ANOM are obviously influenced from heterogeneity of variances. In contrast, the performance of HANOM in the chi-square

distribution is relatively unstable, and it will be conservative or liberal, and the trend is similar to BF. Under the mixed distribution, HANOM presents a consistent and conservative results. Excluding ANOM, both BF and ANOVA yield conservative findings as well.

The relative performances of the ANOVA, BF, ANOM and HANOM test with respect to empirical test powers are given in Figures S4-S15 in Supplementary materials. Under small effect $\Delta 1$, HANOM displayed the smaller power values than other methods regardless of distribution shapes. When there is a larger effect size ($\Delta 2$ and $\Delta 3$), the power values of HANOM remain at a similar level without improving. It is noted that the power values of HANOM have improved as the larger sample size. Similar findings can be found for distribution shapes, effect sizes, and the number of treatments. An initial sample size n (≥ 3) will work in theory, but Dudewicz and Bishop (1981) suggested that 12 or more giving better results. Under normal distribution, despite HANOM have the low test power, it is the only approach that can control the type I error compared to other methods. We see similar results under beta distribution. All methods have poor power in the bimodal distribution of the mixture of $t(10)$ and $N(1, 1)$ distributions that we created. According to the simulation results, HANOM is more suitable for use under normal distribution than other distributions, although it could be conservative sometimes. Most importantly, the single-stage sampling procedure can provide an exact distribution for its statistics employed in the HANOM.

4 Numerical example

An experiment reported from Juneau (2003) was conducted to test the effects of different 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 available in Juneau (2003) and the summary statistics are shown in the upper part of Table 3.

Shapiro-Wilk test was used to check the normality of the data and Levene's test was use to test the homogeneity of variances. We found that this data do not violate the normality assumption and variances are unequal across four groups (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, and we can also conclude that at 0.05 level, the variances are unequal. Hence, when each is at the level of significance $\alpha = 0.05$, the data do not violate the normality assumption with unequal variances. Therefore, the single-stage sampling procedures can be applied.

The p-value for the traditional one-way ANOVA is 0.038 which is less than the 0.05 significance level. Hence, we rejected the null hypothesis that the means of all populations are the same. 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 displays the results of Scheffé multiple comparison procedure. The results of Scheffé method presented no statistical difference between all pairwise comparisons, and there is a contradictory result between ANOVA and multiple comparisons for difficulty interpretation.

In this example, we can use the Brown-Forsythe test to test the equality of means in the case of small samples under heteroscedasticity. The Brown-Forsythe test produces a p-value of 0.107, which is larger than the 0.05 significance level. Hence, we do not reject the equality of population means. However, if we apply the Games and Howell (1976) post-hoc test to perform multiple comparisons, it does not assume equal variances. The results of Games-Howell method presented a significantly statistical difference between Treatment 1 and Treatment 2, and between Treatment 1 and Treatment 3. Similarly, there is a contradictory result between

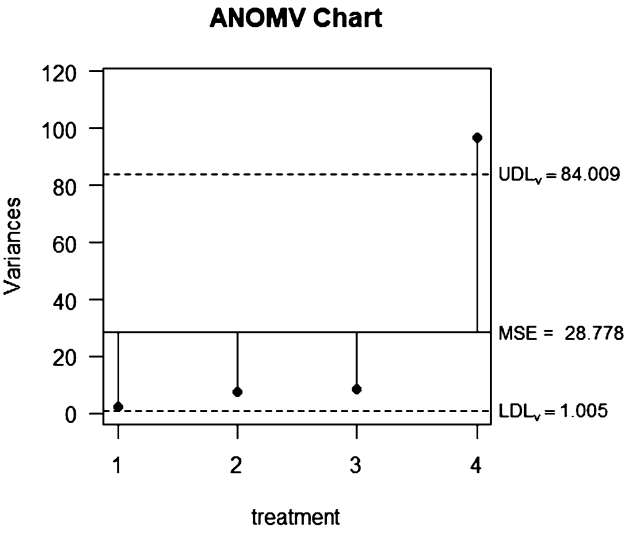

Figure 2 *The ANOMV Chart for Effect of Treatments Data.*

Table 2 *Pairwise comparisons test for Scheffé and Games-Howell*

Contrast	Mean Difference	Std. Error	p-value
	Scheffé Test		
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		
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

the Brown-Forsythe test and the Games-Howell post-hoc test. The second half of Table 2 displays the results of the Games-Howell post-hoc test.

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 applied the single-stage sampling procedure for HANOM to 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 Equation (1) to calculate the weighted sample mean $\widetilde{X}_i$ by weights of U_i and V_i from Equations (2) and (3) for treatment i. The summary statistics are shown in Table 3.

Table 3 *Summary Statistics of Effect of Treatments Data*

	Treatment 1	Treatment 2	Treatment 3	Treatment 4
mean	99.537	110.290	111.032	105.872
standard deviation	1.511	2.742	2.890	9.847
$\overline{X}_i$	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

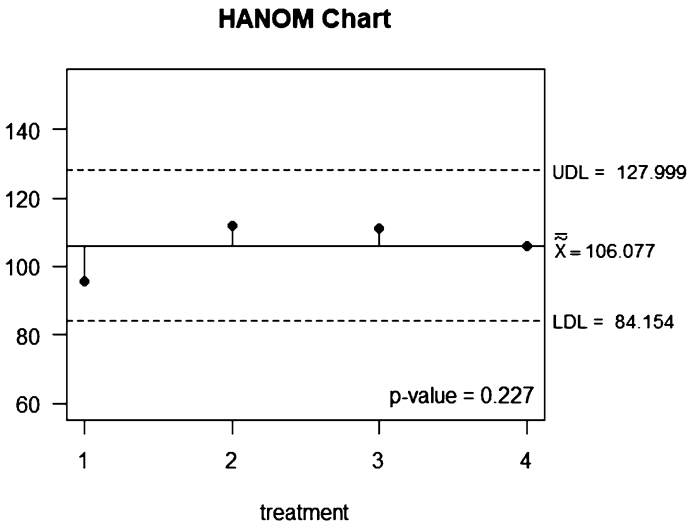

Figure 3 *The HANOM Chart for Effect of Treatments Data.*

For Table 3, we have an overall weighted sample mean. The centerline of HANOM 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 A2 or Appendix D, thus the lower decision line (LDL) and upper decision line (UDL) by Equations (7) and (8) are as follows:

$$\text{LDL} = 106.077 - 7.367 \times \frac{5.952}{\sqrt{4}} = 84.154$$

$$\text{UDL} = 106.077 + 7.367 \times \frac{5.952}{\sqrt{4}} = 127.999$$

where $S_{[k]} = \max\{S_1, S_2, S_3, S_4\} = 5.952$. Figure 3 gives a HANOM decision chart with the centerline, LDL, and UDL. From the results of the HANOM chart, no weighted sample mean falls outside of LDL or UDL which produced a p-value of 0.227, so we conclude that the effect will not be affected by different treatments. This conclusion is same as the result of the Brown and Forsythe test, and no pairwise difference forms the graphic display.

It should be noted that the single-stage sampling procedure for HANOM is based on equal sample size for all samples. In situations where the sample sizes n_i are not all equal, Wu and Chen (2016) suggested that the number n_i be replaced by the harmonic mean of individual sample sizes. We proposed an approximation single-stage sampling procedure for HANOM with unbalanced sample sizes. The procedure employs approximate degrees of freedom to

solve HANOM with unequal sample sizes, and the details are described in Appendix C. Users can also run around this procedure by the shiny app, https://weiming3524.shinyapps.io/HANOM/.

5 Conclusion and discussion

Analysis of Variance (ANOVA) is one of the commonly used statistical methods for testing differences between k population means. Post hoc comparisons can go a step further and examine whether specific population means differ significantly from one another. However, it is difficult to interpret contradictory results between ANOVA and multiple pairwise comparisons that include significant ANOVA with non-significant multiple pairwise comparisons and non-significant ANOVA with significant multiple pairwise comparisons. In contrast to ANOVA, under equal variances assumption, ANOM that can identify which have significant difference from the overall mean through a decision chart is a convenient way to compare k population means. However, equality of variance is not always true for the practice, and it is inconvenient or impossible to transform data to make the variances equal. In this paper, we applied Chen and Lam's (1989) single-stage sampling procedure for the HANOM to eliminate the necessity of such an assumption. From the simulation results, as long as the variances are homogeneous, ANOVA, ANOM and BF tests have generally performed around 5% type I error rate. However, this is not valid in case the variances are heteroscedasticity. In contrast, the single-stage sampling procedure for HANOM can control type I error rates regardless of homoscedasticity or heteroscedasticity, although it could be conservative. The two-stage sampling procedure of HANOM proposed by Nelson and Dudewicz (2002) also eliminates the necessity of such an assumption. But, the two-stage sampling procedure is a design-oriented procedure that requires additional samples, which can be large at the second stage due to heterogeneous variances. It may not be practical in the real world applications and the single-stage sampling procedure is a remedy of the two-stage sampling procedure based on available data on hand. 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, Huang and Zhong (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 $S^2_{[k]}/n$) 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 $S_i^2/n = S_j^2/n$ for all i, j, the single-stage and two-stage sampling procedures have the same decision lines. Under this condition, $(\frac{\omega}{\delta})^2 S^2_{i,n_0}$ except for a rounding error in the sample size determined by a classical two-stage sampling procedure is equal to $S^2_{[k]}/n$ in the single-stage sampling procedure, which gives the same sample sizes in addition to the same decision lines.

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

(iii) When $\min_{1\le j\le k} S_i^2/n > (\frac{\omega}{\delta})^2 S^2_{i,n_0}$, 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 for a 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.

Furthermore, the level of our procedure is completely independent of the unknown variances. The single-stage sampling procedure can also provide an exact distribution for its statistics employed in the HANOM. Therefore, HANOM can be utilized by a single-stage sampling procedure in a data analysis whether the variances are equal or unequal. Finally, we provided a web application using *R* shiny to facilitate the use of the single-stage sampling procedure for HANOM.

Appendix A: Detailed steps for a classical two-stage sampling procedure for HANOM

The classical 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:

1. 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}.
2. 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.
3. Calculate
$$n_i = \max\left\{n_0 + 1, \left[\!\left[\left(\frac{\omega}{\delta}\right)^2 S^2_{i,n_0}\right]\!\right] + 1\right\}$$
for i $(i = 1, \ldots, k)$ population, where $[\![x]\!]$ denoted the greatest integer in x.
4. Take the remaining $n_i - n_0$ $(i = 1, \ldots, k)$ observation to calculate the sample means
$$\overline{X}_{i,n_i-n_0} = \frac{X_{i,n0+1} + \cdots + X_{i,n_i}}{n_i - n_0}$$
5. For each i $(i = 1, \ldots, k)$ population, compute
$$b_i = \frac{n_i - n_0}{n_i}\left[1 + \sqrt{\frac{n_0}{n_i - n_0}\left(\left[\frac{\delta}{\omega}\right]^2 \frac{n_i}{S^2_{i,n_0}} - 1\right)}\right]$$
and
$$\widetilde{\overline{X}}_i = (1 - b_i)\overline{X}_{i,n_0} + b_i \overline{X}_{i,n_i-n_0}$$
and
$$\widetilde{\overline{X}} = \frac{\sum_{i=1}^{k} \widetilde{\overline{X}}_i}{k}$$
6. Calculate decision lines of a HANOM:
$$\widetilde{\overline{X}} \pm H(\alpha; k; n_0 - 1)\frac{\delta}{\omega}$$
where $H(\alpha; k; n_0 - 1)$ is found in Table A.1 in Nelson and Dudewicz (2002).

Appendix B: 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

Table A1 *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 A2 *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 A3 *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

approximate sampling distribution of $\widetilde{T}_{[k]}$ (i.e. solving $P(\widetilde{T}_{[k]} < d^*) = 1 - \alpha/2$). We provide three scenarios when $1 - \alpha = 0.9, 0.95, 0.99$.

Appendix C: Details description for the approximation single-stage sampling procedure for HANOM

In situations where the sample sizes n_i are not all equal, we proposed an approximation single-stage sampling procedure for HANOM that we replaced the different number n_i by the harmonic mean of individual sample sizes, n_0. As a result, the sampling distribution is also an unconditional Student's t distribution with degrees of freedom $n_0 - 1$, and we can modify the single-stage sampling procedure as an approximation single-stage sampling procedure for HANOM. Collecting data for and performing a HANOM consists of the subsequent step:

(i) Calculate the harmonic mean, n_0, of sample sizes n_i from i $(i = 1, \ldots, k)$ population.

$$n_0 = \frac{k}{\frac{1}{n_1} + \cdots + \frac{1}{n_k}}$$

(ii) If $n_i > n_0$, calculate the initial sample mean $\overline{X}_{i,n_0}$, sample variance S^2_{i,n_0} and the weighted sample means $\widetilde{X}_i$; otherwise, calculate the traditional arithmetic means $\overline{X}_i$ and sample variances S_i^2, if $n_i \le n_0$.

$$\begin{cases} \widetilde{X}_i = U_i \sum_{j=1}^{n_0} X_{ij} + V_i \sum_{j=n_0+1}^{n_i} X_{ij}, & \text{for } n_i > n_0 \\ \overline{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i, & \text{for } n_i \le n_0 \end{cases}$$

$\widetilde{X}_i$ is a linear combination of the first set of n_0 observations with weights U_i and the second set of $n_i - n_0$ observations with weights V_i. The weights of U_i and V_i are as

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{n_i - n_0}{n_0^2}\left[n_i S^2_{[k]}/S^2_{i,n_0} - 1\right]}$$

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{\frac{1}{n_i - n_0}\left[n_i S^2_{[k]}/S^2_{i,n_0} - 1\right]}$$

which satisfy the two constraints stated

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

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

where $S^2_{[k]}$ is the maximum of $\{S^2_{1,n_0}, S_i^2\}$ for $i = 1, \ldots, k$.

(iii) The centerline for HANOM decision chart is

$$\overline{\overline{X}} = \frac{\sum_{n_i > n_0} \widetilde{X}_i + \sum_{n_i \le n_0} \overline{X}_i}{k}.$$

Furthermore, the upper and lower decision lines are

$$\overline{\overline{X}} \pm d^* \cdot \frac{S_{[k]}}{\sqrt{n_0}}$$

where the critical value $d^* = d^*(\frac{\alpha}{2}, k, n_0 - 1)$ can also be found in Tables A1–A3 in Appendix B.

Appendix D: *R* shiny usage for the single-stage sampling procedure for HANOM

We developed an interface using *R* shiny, https://weiming3524.shinyapps.io/HANOM/, to simulate the single-stage sampling procedure for HANOM. The interface provides the basic descriptive statistic, test for normality/variance and the HANOM chart by uploading a dataset.

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. Every user can also download a sample file from "Download Toy Dataset" tab on the shiny interface. 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 broad.

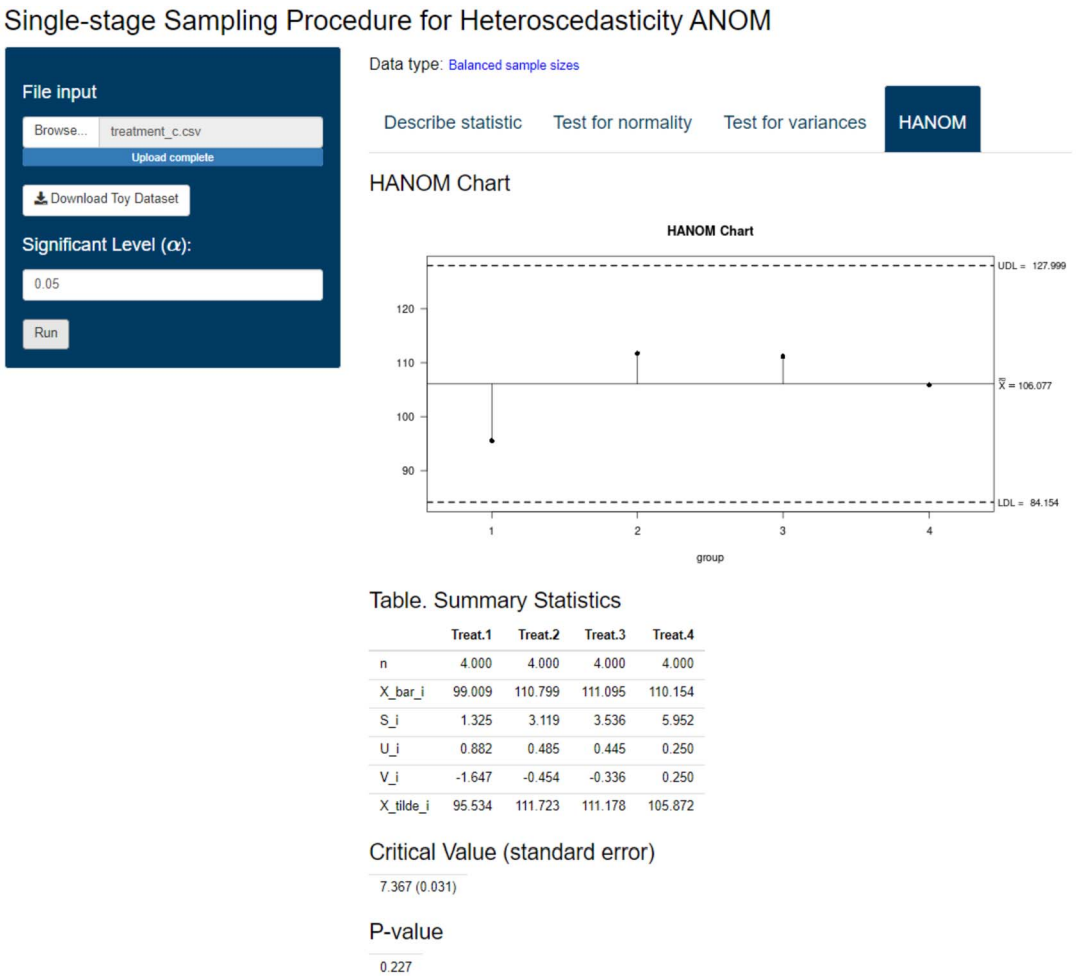

Figure A1 *The R shiny interface for HANOM by using the single-stage sampling procedure.*

Appendix E: Simulation results for type I error rates under normal distribution

Table A4 *Simulated type I error rates under normal distribution*

		$k=3$			$k=4$			$k=5$		
		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.

Funding

Wang's and Wen's research was sponsored by the Ministry of Science and Technology, Taiwan (MOST 109-2118-M-006-004-MY2). The research of Zhong was supported by the Ministry of Educational of China project of Humanities and Social Sciences (21YJC910011), the Education and Scientific Research Foundation for Young Scholars in Fujian Province, China (JAT190665).

Supplementary Material

Supplement to "Single-stage sampling procedure for heteroscedasticity analysis of means" (DOI: 10.1214/22-BJPS550SUPP; .pdf). Supplementary information.

References

Bishop, T. A. (1976). Heteroscedastic anova, manova, and multiple-comparisons. Ph.D. thesis, The Ohio State University. MR2626711

Brown, M. and Forsythe, A. (1974). Small sample behavior of some statistics which test the equality of several means. *Technometrics* **16**, 129–132. MR0334368 https://doi.org/10.2307/1267501

Chen, H. J. and Lam, K. (1989). Single-stage interval estimation of the largest normal mean under heteroscedasnoty. *Communications in Statistics Theory and Methods* **18**, 3703–3718. MR1040672 https://doi.org/10.1080/03610928908830118

Chen, H. J. and Wen, M. J. (2006). Optimal confidence interval for the largest normal mean under heteroscedasticity. *Computational Statistics & Data Analysis* **51**, 982–1001. MR2297501 https://doi.org/10.1016/j.csda.2005.10.004

Chen, S. Y. (2001). One-stage and two-stage statistical inference under heteroscedasticity. *Communications in Statistics Simulation and Computation* **30**, 991–1009. MR1878465 https://doi.org/10.1081/SAC-100107792

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

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**, 49–67. MR2044861 https://doi.org/10.1081/SAC-120028433

Dag, O., Dolgun, A. and Konar, N. M. (2021). Onewaytests: An *R* Package for One-Way Tests in Independent Groups Designs. *R*. package version, 2.6.

Dudewicz, E. J. and Bishop, T. A. (1981). Analysis of variance with unequal variances. *Journal of Quality Technology* **13**, 111–114.

Dudewicz, E. J. and Nelson, P. R. (2003). Heteroscedastic analysis of means (HANOM). *American Journal of Mathematical and Management Sciences* **23**, 143–181. MR2017431 https://doi.org/10.1080/01966324.2003.10737608

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

Hasler, M. (2016). Heteroscedasticity: Multiple degrees of freedom vs. sandwich estimation. *Statistical Papers* **57**, 55–68. MR3461948 https://doi.org/10.1007/s00362-014-0640-4

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

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 Miami, Florida, USA*, *May 4–7, 2003*.

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**, 259–271.

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

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

Nelson, P. R. and Dudewicz, E. J. (2002). Exact analysis of means with unequal variances. *Technometrics* **44**, 152–160. MR1951723 https://doi.org/10.1198/004017002317375109

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*, 25. MR2162661 https://doi.org/10.1137/1.9780898718362

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

Pallmann, P. (2017). ANOM: Analysis of means. http://CRAN.R-project.org/package=ANOM, R package version 0.4.

Rao, C. V. (2005). Analysis of means—a review. *Journal of Quality Technology* **37**, 308–315. MR2065771 https://doi.org/10.1016/S0169-7161(03)23014-5

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics* **16**, 243–258. MR0013885 https://doi.org/10.1214/aoms/1177731088

Wang, W.-M., Wen, C.-C., Hung, T.-H., Zhong, J. and Wen, M.-J. (2022). Supplement to "Single-stage sampling procedure for heteroscedasticity analysis of means." https://doi.org/10.1214/22-BJPS550SUPP

Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica: Journal of the Econometric Society*, 1211–1215. MR0913368 https://doi.org/10.2307/1911268

Wen, M. J. and Chen, H. J. (1994). Single-stage multiple comparison procedures under heteroscedasticity. *American Journal of Mathematical and Management Sciences* **14**, 1–48. MR1287444 https://doi.org/10.1080/01966324.1994.10737368

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**, 9265–9273. MR3684567 https://doi.org/10.1080/03610926.2016.1206935

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

Wludyka, P. S. and Nelson, P. R. (1997). An analysis-of-means-type test for variances from normal populations. *Technometrics* **39**, 274–285. MR1470274 https://doi.org/10.1080/01966324.1997.10737429

Wu, S. F. and Chen, H. J. (1998). Multiple comparisons with the average for normal distributions. *American Journal of Mathematical and Management Sciences* **18**, 193–218. MR1655488 https://doi.org/10.1080/01966324.1998.10737459

Wu, S. F. and Chen, H. J. (2016). An optimal confidence region for the largest and the smallest means from a multivariate normal distribution. *Communications in Statistics Simulation and Computation* **45**, 952–967. MR3473923 https://doi.org/10.1080/03610918.2014.889160