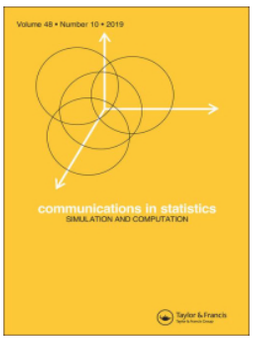


Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

A modified single-stage sampling procedure for heteroscedasticity analysis of means with unbalanced design

Wei-Ming Wang, Chun-Che Wen, Rajdeep Das & Miin-Jye Wen

To cite this article: Wei-Ming Wang, Chun-Che Wen, Rajdeep Das & Miin-Jye Wen (2022): A modified single-stage sampling procedure for heteroscedasticity analysis of means with unbalanced design, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2022.2074456

To link to this article: https://doi.org/10.1080/03610918.2022.2074456

Published online: 12 May 2022.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2022.2074456

A modified single-stage sampling procedure for heteroscedasticity analysis of means with unbalanced design

Wei-Ming Wang[a], Chun-Che Wen[b], Rajdeep Das[c], and Miin-Jye Wen[a,c]

[a]Department of Statistics and Institute of Data Science, College of Management, National Cheng Kung University, Tainan, Taiwan; [b]Department of Public Health Sciences, College of Medicine, Medical University of South Carolina, Charleston, South Carolina, USA; [c]Institute of International Management, College of Management, National Cheng Kung University, Tainan, Taiwan

ABSTRACT

The analysis of means (ANOM) is a method that can compare the mean of each treatment with the overall mean. As one-way ANOVA, the conventional ANOM is not robust under unequal variances. We applied two inference procedures, single-stage and modified single-stage sampling, to solve heteroscedastic analysis of means (HANOM) under unbalanced design. Using proposed weighted average across groups, the influence of the unknown mean and variance are both eliminated. To validate the stability of HANOM under several scenarios of unequal variances and sample sizes, simulation studies of empirical type I error rate are conducted to test the quality of procedures. A real application is provided for illustrating these two procedures clearly. In addition, we build an user-friendly interface to show the results of the HANOM by using Shiny package in *R* software.

ARTICLE HISTORY
Received 21 October 2020
Accepted 26 April 2022

KEYWORDS
Heteroscedasticity; Single-stage sampling procedure; Unbalanced design; Analysis of means

1. Introduction

Analysis of variance (ANOVA) is a commonly used statistical method to test whether the means of all treatments are equal. Once the null hypothesis is rejected, the following goal is to find the statistical difference pair of the means. Previous studies show post-hoc results from different multiple comparisons procedures are not consistent or even contradict to ANOVA. Analysis of means (ANOM) is an alternative method to solve the problem above. Ott (1967) introduced a graphical statistical technique for ANOM, which is used for comparing k treatment means to find out whether any of the mean varies significantly away from the overall mean. One of the advantages of ANOM is to provide a control chart that is useful and simpler way to visually display the difference between groups and overall mean. If a treatment mean falls outside the decision limits, it indicates that this particular mean is statistically different from the overall average. Therefore, ANOM has the advantage of providing a graphical display that allows one to easily evaluate both the statistical and practical significances of the differences (Nelson and Dudewicz 2002). The use of ANOM technology has already exerted its advantages in different fields, such as engineering, clinical medicine, and biomedicine, and so on, (Kiani, Panaretos, and Psarakis 2008; Lopez-Mejia and Roldan-Valadez 2016; Castillo-Cuenca et al. 2021)

Nelson (1982) and Nelson (1991) developed a method for computing exact critical values for ANOM in balanced designs and unbalanced designs, respectively. One assumption of the classical ANOM model is that the variances are equal, which is the same as an ANOVA model. However,

CONTACT Wei-Ming Wang r28051012@gs.ncku.edu.tw Department of Statistics and Institute of Data Science, College of Management, National Cheng Kung University, Tainan, Taiwan

equality of variance is not always feasible, and it takes more steps to convert the data with equal variance. This might lost the interpretability of data even in doing so. For unequal and unknown variances, the two-stage sampling procedure of analysis of mean under heteroscedasticity (HANOM) proposed by Nelson and Dudewicz (2002) eliminates the necessity of such assumptions. The two-stage sampling procedure is design-oriented that requires equal sample sizes across the treatment groups in the first stage. Due to the heterogeneous variances, the additional samples should be achieved in a fair amount. The number of required samples can be large and perhaps make the procedure non-viable during the second stage. When it comes to real-world applications, it may not be practical because we often have only one single sample available while working on statistical data analyses. Therefore, a single-stage sampling procedure originally developed by Chen and Lam (1989) to construct the confidence interval of the largest normal mean was designed to work out these problems. Chen and Chen (1998) extended the single-stage sampling procedure to two-way layout ANOVA. 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 a one-stage procedure for testing the ordered treatment means in the one-way layout. Wen, Huang, and Zhong (2017, Wen, Wen, and Wang 2022) proposed a single-stage sampling procedure of the t best populations and the multiple comparison with a control under unequal sample size, respectively.

Under an unbalanced design with k treatment groups, each degree of freedom are varying corresponding to their t distribution; however, this joint t distribution may cause the difficulties in solving the critical values due to the asymmetry. In this paper, we applied Chen and Chen (1998) and Chen's (2001) single-stage and modified single-stage sampling procedure, respectively, to find critical values and to solve the HANOM in the one-way layout with an unbalanced design. In Sec. 2, a single-stage sampling procedure with a minor correction to HANOM has been introduced. We use a grid-searching method to determine the required sample size with a prespecified level of HANOM test power. In Sec. 3, a modified single-stage sampling procedure to HANOM is introduced. In Sec. 4, simulation studies for empirical type I error rates are considered. In Sec. 5, one numerical example is given. Finally, a conclusion and discussion is given in Sec. 6.

2. Single-stage sampling procedure $\mathcal{P}_1$

When the sample size are unequal, we used Chen and Chen (1998) single-stage sampling procedure to solve the HANOM. This procedure is referred to as $\mathcal{P}_1$ and the details are as follows:

$\mathcal{P}_1$: Suppose we have k treatment groups and n_i observations among i-th group. Let X_{ij} $(i = 1, ..., k; j = 1, ..., n_i)$ be random samples from a normal distribution with unknown mean, μ_i, and unknown and unequal variance, σ_i^2. Bishop (1976) firstly introduced the idea of splitting a sample into two portions in the ANOVA problem. Lam (1987) applied this method to a single-stage subset selection procedure under heteroscedasticity. Then Chen and Lam (1989) used it to interval estimation of the largest means and a series of follow-up studies (Wen, Chen, and Chen 2007; Chen, Wen, and Wang 2009; Chen, Wen, and Chuang 2011; Wen, Huang, and Zhong 2017; Wen, Wen, and Wang 2022). Among each i-th $(i = 1, ..., k)$ population, we randomly select $n_i - 1$ observations to define the sample mean and the sample variance, respectively, by

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

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

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

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

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

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

where $S_{[k]}^2$ is the maximum of $\{S_1^2, \ldots, S_k^2\}$. The advantage of Equation (2) implies $\sum_{j=1}^{n_i-1} U_i + V_i = \sum_{j=1}^{n_i} \omega_{ij} = 1$ that makes the weighted sample mean, $\tilde{X}_i$ in (1), an unbiased estimate of the treatment mean, μ_i, and Equation (3) implies $\sum_{j=1}^{n_i} \omega_{ij}^2 S_i^2 = S_{[k]}^2/n_i$ to eliminate the influence of the unknown and unequal variances according to Stein (1945). Based on the Equations (2) and (3), the solutions of the weights of U_i and V_i are derived as

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

and

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

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

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

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

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

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 t distribution with degrees of freedom $n_i - 2$ (Wen and Chen 1994; Wen, Wen, and Wang 2022).

2.1. The one-way layout model

For 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 of the i-th treatment, ε_{ij} are assumed to be normally and independently distributed with mean 0 and unknown group-specific variance σ_i^2. The objective here is to test the null hypothesis that the population means, μ_i, $i = 1, \ldots, k$, are all equal, that is, $H_0 : \mu_1 = \ldots = \mu_k$. In contrast to the alternative hypothesis in which one or more means are different from the overall mean.

The main characteristic in HANOM facilitates decisions were made by a HANOM decision chart (Nelson and Dudewicz 2002). HANOM can also indicate which treatment mean differs

significantly from the overall mean when there are differences among the treatment means. Three lines constitute a HANOM decision chart:

i. Centerline

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

which is an overall mean of all weighted sample means.

ii. Lower Decision Line (LDL)

$$\overline{\tilde{X}} - h^* \cdot \frac{S_{[k]}}{\sqrt{n_i}} \tag{7}$$

iii. Upper Decision Line (UDL)

$$\overline{\tilde{X}} + h^* \cdot \frac{S_{[k]}}{\sqrt{n_i}} \tag{8}$$

where $h^* = h^*\left(\frac{\alpha}{2}, k, n_1 - 2, \ldots, n_k - 2\right)(> 0)$ is the critical value depending on the level of significance, α, k treatment groups, and the degrees of freedom $(n_i - 2,$ where $i = 1, \ldots, k)$.

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

2.2. Determination of critical values

Similar to Nelson (1982), we define the rejection region of treatment i as

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

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

$$\begin{aligned} 1 - \alpha &= P\left(\left| \frac{\tilde{X}_i - \overline{\tilde{X}}}{S_{[k]}/\sqrt{n_i}} \right| \le h^*,\ i = 1, \ldots, k \right) \\ &\overset{H_0}{=} P\left(-h^* \le \frac{\tilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{1}{k} \sum_{z=1}^{k} \frac{\tilde{X}_z - \mu_z}{S_{[k]}/\sqrt{n_i}} \le h^*,\ i = 1, \ldots, k \right) \qquad (9) \\ &= P\left(-h^* \le \left(\frac{k-1}{k} \right) T_i - \sqrt{n_i} \left(\frac{1}{k} \sum_{z \ne i}^{k} \frac{T_z}{\sqrt{n_z}} \right) \le h^*,\ i = 1, \ldots, k \right) \\ &= P\left(-h^* \le \tilde{T}_i \le h^*,\ i = 1, \ldots, k \right) \qquad (10) \end{aligned}$$

where $\tilde{T}_i = \left(\frac{k-1}{k}\right) T_i - \sqrt{n_i}\left(\frac{1}{k}\sum_{z \ne i}^{k} \frac{T_z}{\sqrt{n_z}}\right)$, $i = 1, \ldots, k$. The detailed proof is given in Appendix A. From Equation (6), T_i and T_z in Equation (9) are t distributions with degrees of freedom $n_i - 2\,(i = 1, \ldots, k)$ and $n_z - 2\,(z = 1, \ldots, k)$, respectively. Here, we take $\tilde{T}_{[1]}$ is the minimum of $\left\{\tilde{T}_1, \ldots, \tilde{T}_k\right\}$ and $\tilde{T}_{[k]}$ is the maximum of $\left\{\tilde{T}_1, \ldots, \tilde{T}_k\right\}$. The degrees of freedom of $\tilde{T}_{[1]}$ and $\tilde{T}_{[k]}$ are different, so the critical values of $\tilde{T}_{[1]}$ and $\tilde{T}_{[k]}$ are different. We rewrote Equation (10) as follows.

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

Therefore, we select the critical value h^* to satisfy the following condition:

$$\max\left\{P\left(\tilde{T}_{[1]} < -h^*\right), P\left(\tilde{T}_{[k]} > h^*\right)\right\} \le \frac{\alpha}{2}. \tag{11}$$

Monte-Carlo simulation method is used to find out the approximate sampling distributions of $\tilde{T}_{[1]}$ and $\tilde{T}_{[k]}$. k Student's t distributions with specific degrees of freedom are required to generate and calculate $\tilde{T}_i$, $i = 1, ..., k$. Afterwards, we sort these $\tilde{T}_i$, $i = 1, ..., k$, from small to large, and find the minimum value $\tilde{T}_{[1]}$ and the maximum value $\tilde{T}_{[k]}$ at every single run. After 100,000 independent runs, the probability of $\alpha/2$ of the distributions of $\tilde{T}_{[1]}$ and $\tilde{T}_{[k]}$ can be approximated by selecting the cumulated number from the ranking list of the 100,000 values of $\tilde{T}_{[1]}$ and $\tilde{T}_{[k]}$, respectively. We take the $100(1-\alpha/2)$-*th* percentile of the approximate sampling distribution of $\tilde{T}_{[k]}$ and the $100(\alpha/2)$-*th* percentile of the approximate sampling distribution of $\tilde{T}_{[1]}$, and select the larger critical value as h^* according to Equation (11).

2.3. Computing the p-value

The p-value for HANOM can be identified as a value that the weighted sample mean, $\tilde{X}_i$, is the greatest distance away from the weighted overall mean, $\overline{\tilde{X}}$ (Nelson, Wludyka, and Copeland 2005). More precisely, the p-value of the procedure $\mathcal{P}_1$ for HANOM is determined by

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

So, p-value is obtained by solving the equation above. We increase p by 0.001 from 0 to 1. After running the HANOM procedure, we try to find the $\left(100\frac{p}{2}\right)$ percentile of h^* distribution and could find the minimized p as the p-value of the $\mathcal{P}_1$ for HANOM that is

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

Therefore, we can obtain control chart, critical value, and p-value from the above procedures.

2.4. Determination of power and sample size

Define the power function for the HANOM is as follows.

$$\begin{aligned}\text{power}(\boldsymbol{\mu}) &= 1 - P\left(-h^* \le \frac{\tilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n_i}} - \frac{\overline{\tilde{X}} - \bar{\mu}}{S_{[k]}/\sqrt{n_i}} + \frac{\mu_i - \bar{\mu}}{S_{[k]}/\sqrt{n_i}} \le h^*,\ i = 1, ..., k\right)\\ &= 1 - P\left(-h^* \le \tilde{T}_i + \frac{\mu_i - \bar{\mu}}{S_{[k]}/\sqrt{n_i}} \le h^*,\ i = 1, ..., k\right),\end{aligned} \tag{12}$$

where $\boldsymbol{\mu}$ is a mean vector and the power depends on the particular configuration of the $\mu_i's$. The power is determined using the least favorable configuration (LFC) of the means on the subspaces

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

It is shown in Nelson and Dudewicz (2002) that on M_δ the LFC for HANOM is a configuration of the form

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

Therefore, the power function from Equation (12) will be rewritten by

$$\begin{aligned} \text{power}(\boldsymbol{\mu}) &= 1 - P\left(-h^* \leq \tilde{T}_i^* \leq h^*,\ i = 1, ..., k\right) \\ &\geq 1 - P\left(\tilde{T}_{[1]}^* < -h^*\right) - P\left(\tilde{T}_{[k]}^* > h^*\right) \end{aligned} \tag{13}$$

where $\tilde{T}_i^* = \tilde{T}_i + \frac{\mu_i - \bar{\mu}}{S_{[k]}/\sqrt{n_i}}, i = 1, ..., k,$ while $\tilde{T}_{[1]}^*$ and $\tilde{T}_{[k]}^*$ are the minimum and the maximum of $\left\{\tilde{T}_1^*, ..., \tilde{T}_k^*\right\}$, respectively. The power function given the level of significance, α, number of treatments, k, degrees of freedom, $n_i - 2$ where $i = 1, ..., k,$ the largest standard deviation, $S_{[k]}$, and the effect size δ could be calculated in a similar method to the critical values. We developed an interface by *R* Shiny to obtain these values. The URL of the interface can accessed at https://weiming3524.shinyapps.io/hanom_ub/.

For a given level of significance, α, and a desired power of HANOM, $1 - \beta$, the sample size n_i could be determined. However, the test statistic, critical value, and power of the single-stage sampling procedure for HANOM depend on the sample sizes of each treatment. The required sample size of each treatment cannot be directly obtained. We use a grid-searching method to obtain the required sample size as follows:

i. Take an initial sample size $n_{0i} = 3, i = 1, ..., k$. The critical value for HANOM can be obtained from Equation (11) by a given level of significance, α.
ii. If the power value for HANOM is calculated by Equation (13) under the initial sample size meets the desired power, the required sample size is n_{0i}. If not, we increase n_{0i} by 1 until it achieves the desired power.
iii. Iteratively run step (i) and (ii) until the power achieves the desired $1 - \beta$. At this time, the sample size n_i is the minimum required size to guarantee the given power.

Of note, sample size and power calculation can both be obtained through the *R* shiny interface we created.

3. Modified single-stage sampling procedure $\mathcal{P}_2$

Instead of using the $n_i - 1$ observations from each population in the $\mathcal{P}_1$ procedure, Chen (2001) developed another single-stage sampling procedure by employing the n_0 observations to test the equality of means in the ANOVA setting with unknown and unequal variances. We applied this procedure to solve the HANOM with an unbalanced design. The modified single-stage sampling procedure is referred to as $\mathcal{P}_2$:

1. For each i population of X_{ij} $(i = 1, ..., k; j = 1, ..., n_i)$, take the initial sample n_0 to be the minimum of $n_i - 1$ $(i = 1, ..., k)$, and calculate the initial sample mean $\bar{X}_{i,n_0}$ and the sample variance S_{i,n_0}^2.

2. For each population, calculate the weighted sample mean

$$\hat{X}_i = U_i \sum_{j=1}^{n_0} X_{ij} + V_i \sum_{j=n_0+1}^{n_i} X_{ij}. \tag{14}$$

The weights of U_i and V_i are as

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{n_i - n_0}{n_0}\left[n_i z^*/S^2_{i,n_0} - 1\right]} \tag{15}$$

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{\frac{n_0}{n_i - n_0}\left[n_i z^*/S^2_{i,n_0} - 1\right]} \tag{16}$$

which satisfy the following two conditions as in $\mathcal{P}_1$ procedure

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

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

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

3. The weighted sample mean, $\hat{X}_i$ $(i = 1, ..., k)$, shows that the random variables $T_i^* = (\hat{X}_i - \mu_i)/\sqrt{z^*}$, $i = 1, ..., k$, are distributed independently following the Student's t distribution with degrees of freedom $n_0 - 1$. The rigorous proof can be referred to Wen and Chen (1994).

3.1. HANOM decision chart for the one-way layout model

Similar to procedure $\mathcal{P}_1$, the centerline of the HANOM decision chart of the $\mathcal{P}_2$ is $\bar{\hat{X}} = \sum_{i=1}^{k} \hat{X}_i/k$ which is an overall mean of all weighted sample means. In addition, the upper and the lower decision lines are $\bar{\hat{X}} \pm d^* \cdot \sqrt{z^*}$, respectively. The critical value, $d^* = d^*\left(\frac{\alpha}{2}, k, n_0 - 1\right)(> 0)$, depends on the level of significance, α, the k treatment groups, and the degrees of freedom, $n_0 - 1$. Given the level of significance, α, the critical value, d^*, for HANOM satisfy

$$\alpha = P\left(\left|\frac{\hat{X}_i - \bar{\hat{X}}}{\sqrt{z^*}}\right| > d^*,\ i = 1, ..., k\right)$$
$$\overset{H_0}{=} P\left(\left|\widehat{T}_i\right| > d^*,\ i = 1, ..., k\right)$$

where $\widehat{T}_i = T_i^* - \bar{T}^*$. And, T_i^*, $i = 1, ..., k$, following the t distribution with degrees of freedom $n_0 - 1$, while $\bar{T}^*$ is the mean of the T_i^*. Here, we take $P(\hat{T}_{[1]} < -d^*) + P(\hat{T}_{[k]} > d^*) \leq \alpha$. Since $(\widehat{T}_1, ..., \widehat{T}_k)$ is symmetric by Wu and Chen (1998) and the degrees of freedom are the same, we let

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

Similarly, finding out the exact joint distribution of $(\widehat{T}_1, ..., \widehat{T}_k)$ as $(\tilde{T}_1, ..., \tilde{T}_k)$ is also difficult. Hence, Monte-Carlo simulation is also used to obtain the approximate sampling distribution of the maximum order statistic of $(\widehat{T}_1, ..., \widehat{T}_k)$, namely $\hat{T}_{[k]}$. Same as the procedure $\mathcal{P}_1$, we took the $100(1 - \alpha/2)$-th percentile of the approximate sampling distribution of $\hat{T}_{[k]}$ by 100,000 independent runs, that is, $P(\hat{T}_{[k]} < d^*) = 1 - \alpha/2$. The critical value, d^*, is obtained.

The p-value of the procedure $\mathcal{P}_2$ for HANOM can be obtained by the greatest distance away from the centerline; that is,

Table 1. Experimental conditions.

	Number of treatment groups		
	$k=3$	$k=4$	$k=5$
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

Note: k is the number of the treatment groups; n1 to n3 are the equal sample size allocations and n4 to n7 are the unequal sample size allocations; V0 is the homogeneity and V1 to V3 are the heterogeneity situations.

$$1-p=P\left(\max_i\left|\frac{\hat{X}_i-\bar{\hat{X}}}{\sqrt{z^*}}\right|\le d^*\left(\frac{p}{2},k,n_0-1\right)\right).$$

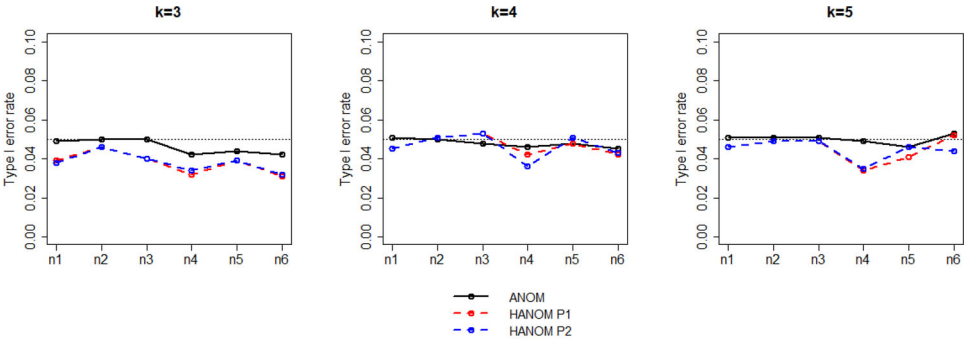
Hence, we can also obtain control chart, critical value, and p-value from the $\mathcal{P}_2$ procedures. In order to use the procedures easily, the web application using *R* Shiny was developed to simulate both $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM. The URL of the interface can be accessed at https://weiming3524.shinyapps.io/hanom_ub/.

4. Empirical type I error rate simulations

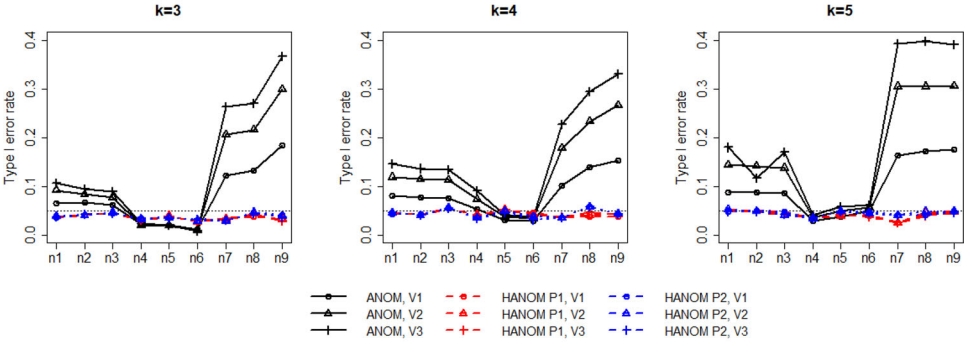
In the simulation studies, classical ANOM test, $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM are compared with respect to the type I error rates (α) for validating their quality. Based on the previous studies (Mendeş and Yiğit 2013), we chose $k=3$, 4, 5 treatments under the normal distributions in order to compare theirs associated empirical type I error rates. The empirical type I error rates were estimated by determining the total number of null hypotheses that were incorrectly rejected. We ran the total number of 5000 trials, and the empirical type I error rate was the number of times that we rejected the null hypothesis which is divided by the total run. Several experimental conditions were given in detail (Table 1). Scenarios n1 to n3 were the equal sample size and n4 to n9 were the unequal sample size conditions. Among the unequal sample size conditions n4 to n6, the larger sample sizes were associated with larger variances (direct pairing). Among the n7 to n9 conditions, the larger sample sizes were associated with smaller variances (inverse pairing). And, the scenarios included the homogeneity conditions (V0) and the heteroscedasticity situations (V1 to V3).

Figure 1 shows the relative performances of the classical ANOM test, $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM with respect to empirical type I error rates corresponding to the different scenarios. The numerical results such as empirical type I rate were shown in Appendices (Table B1–B2).

Figure 1a contains the results when the variances are homogeneity. The results reveal that ANOM tests displayed similar to type I error rates and were not greatly influenced by making the sample size unequal in groups. $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM are slightly conservative, especially the fewer number of group ($k=3$) and the small samples size (n4) for $k=4$ and $k=5$. When the samples sizes are equal, the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures are equivalent. There was no



(a) Type I error rate when variance are homogeneous



(b) Type I error rate when variance are heteroscedasticity

Figure 1. Type I error rate of the ANOM and the HANOM procedures under different sample size allocations and variance scenarios.

Table 2. Strength of reinforcing bars 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

obvious difference between the $\mathcal{P}_1$ and the $\mathcal{P}_2$ procedures under unequal sample size conditions (n4 to n6).

Figure 1b contains the results when the variances are heteroscedasticity. It was observed that the ANOM tests were more influenced by the heterogeneity of variances than the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for the HANOM. When $k = 3$, the type I error rates of the ANOM tests exceeded the nominal level of 5% under an equal sample size (n1 to n3). Under unequal sample size, the type I error rates of the ANOM tests were below 5% (conservative) when sample sizes and variances were directly paired (n4 to n6). The ANOM tests were markedly exceeded 5% (liberal) when sample sizes and variances were inversely paired (n7 to n9). In addition, the increase in variance ratios (V2 and V3) further affected the results. The $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM are more robust than the ANOM test under these conditions. This is the same as the homogeneity conditions. There is no obvious difference between the $\mathcal{P}_1$ and the $\mathcal{P}_2$ procedures. As noticed, the type I error rates of ANOM and HANOM displayed similar trends from increasing the number of groups. The difference is that the ANOM tests markedly exceeded the level 5% as the increases in the number of groups. Empirical α can even reach up to near 40% in the certain cases ($k = 5$).

Table 3. The results of the multiple comparisons test.

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

The $\mathcal{P}_1$ procedures displayed the type I error rates between $0.026 \leq \alpha \leq 0.056$ and the $\mathcal{P}_2$ procedures were between $0.029 \leq \alpha \leq 0.059$ in all conditions.

Overall, the ANOM tests generally performed type I error rates around 5% under homogeneity of variances. We also showed that the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for HANOM are robust under heterogeneous variances.

5. A numerical example

Weerahandi (2004) gave an experiment in engineer for comparing four brands of reinforcing bars. The purpose of the experiment is to test the mean strength of reinforcing bars for equality. The data are given in Table 2. Shapiro-Wilk test was used to check the normality of the data and Levene's test was used 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 $\geq$ 0.121 and 0.001, respectively).

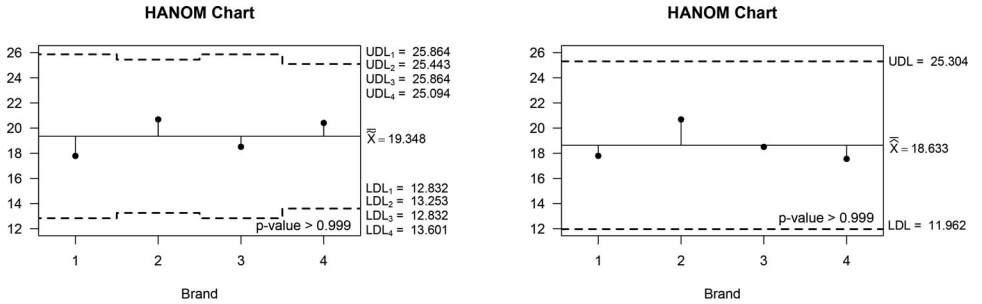
First, we used the traditional Brown-Forsythe test (1974) to test the equality of means in the case of small samples under heteroscedasticity. The p-value produced by the Brown-Forsythe test is 0.232, which means we did not reject the equality of the population means. However, if we use the Games-Howell post-hoc test (1976) to perform the multiple multiple comparison tests, the results show a significant difference in brand mean between groups 3 and 4 (p-value = 0.016, Table 3). There is a contradictory result between the two tests, such as Brown-Forsythe and the Games-Howell post-hoc, and consequently the results are difficult to interpret. In addition, we applied the generalized fiducial test and parametric bootstrap test of the generalized p-value based tests were developed by Weerahandi and Krishnamoorthy (2019) to this example. The p-value of the fiducial test and parametric bootstrap are 0.035 and 0.026, respectively, which both are less than the 0.05 significance level and we reject the null hypothesis. After rejecting the null hypothesis that the treatment means are equal, we apply the post-hoc multiple comparison procedures. These two procedures and pairwise comparisons are available by the package *onewaytests* (Dag et al. 2021a) of the statistical software *R*. The results in the Table 3 presented no statistical difference between all pairwise comparisons. It is, again, difficult to interpret the contradictory result between the global mean test and the multiple comparisons. Alternatively, we applied the HANOM to this example. The HANOM provides global information and additionally allows for the local inference that itemizes which specific treatment means differ from the grand mean (Pallmann and Hothorn 2016).

We first applied the single-stage sampling procedure $\mathcal{P}_1$ for the HANOM to this example. We calculated the initial sample mean $\bar{X}_i, i = 1, ..., 4$, and the sample standard deviation S_i with the first $n_i - 1$ observations. Then, we used Equation (1) to calculate the weighted sample mean $\tilde{X}_i$ by the weights of U_i and V_i from Equations (4) and (5) for brand i. The summary statistics are shown in Table 4.

The upper and the lower decision lines are $\overline{\tilde{X}} \pm h^* \cdot \frac{S_{[k]}}{\sqrt{n_i}}$, respectively. According to the Table 4, the centerline of HANOM decision chart is as follows:

Table 4. Summary statistics for $\mathcal{P}_1$ and $\mathcal{P}_2$ of the strength of reinforcing bars data.

$\mathcal{P}_1$	Brand 1	Brand 2	Brand 3	Brand 4
n_i	7	8	7	9
$\bar{X}_i$	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
$\tilde{X}_i$	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
$\bar{X}_{i,n_0}$	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
$\hat{X}_i$	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



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

Figure 2. HANOM chart by procedure $\mathcal{P}_1$ and $\mathcal{P}_2$.

$$\bar{\tilde{X}} = \frac{17.785 + 20.688 + 18.511 + 20.407}{4} = 19.348$$

At the 0.05 level, the larger critical value $h^* = 3.001$ (the standard error is 0.017), where $h^* = h^*\left(\frac{\alpha}{2} = 0.025, k = 4, df_1 = 5, df_2 = 6, df_3 = 5, df_4 = 7\right)$. And, $S_{[k]} = \max\{S_1, S_2, S_3, S_4\} = 5.745$. The lower decision line (LDL) and the upper decision line (UDL) can be calculated from Equations (7) and (8), and the results are shown in Table 4 and Figure 2a. From the results of the HANOM chart (Figure 2a), there is no weighted sample mean, $\tilde{X}_i$, falling outside of LDLs or UDLs, which indicates the strength in bars would not be affected by different brands.

Furthermore, we applied the modified single-stage sampling procedure $\mathcal{P}_2$ to this example. The sample size n_i of the brand i are $n_1 = 7, n_2 = 8, n_3 = 7,$ and $n_4 = 9$. n_0 is the minimum of $\{n_1 - 1, ..., n_4 - 1\}$ that is $n_0 = 6$. We calculated the initial sample mean and the sample standard deviation with the first 6 observations to obtain $\bar{X}_{i,n_0}$ and S_{i,n_0}, respectively. Then, we used Equation (14) to calculate the weighted sample mean $\hat{X}_i$ by the weights of U_i and V_i from Equations (15) and (16) for brand $i, i = 1, ..., 4$. The summary statistics are shown in Table 4.

The upper and the lower decision lines are $\bar{\hat{X}} \pm d^* \cdot \sqrt{z^*}$, respectively. According to the Table 4, The overall weighted sample mean $\bar{\hat{X}}$ is equal to 18.633. At the 0.05 level, the critical value $d^* = d^*\left(\frac{\alpha}{2} = 0.025, k = 4, df = 5\right) = 3.135$ (the standard error is 0.015). 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 line (UDL) and the lower decision

line (LDL) are $18.633 \pm 3.135 \times \sqrt{4.527}$, and the results are also shown in Table 4 and Figure 2b. According to the results of the HANOM chart (Figure 2b), no weighted sample mean fell outside of LDL or UDL. We thus conclude that the effect will not be affected by different treatments. This conclusion is the same as the single-stage sampling procedure $\mathcal{P}_1$. Four treatment means in $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures do not significant differ from the grand mean and produce a p-value greater than 0.999 in both procedures. The reason for this conservative results is the relatively small sample sizes. If we used the ANOM test in this numerical example, we also get a large and non-significant p-value. As per theory, an initial sample of size greater than or equal to 3 ($n_i \geq 3$) will work, but 10 or more would be more suitable for a better result according to Bishop and Dudewicz (1978).

6. Conclusion and discussion

Analysis of means (ANOM) is an alternative test to the analysis of variance (ANOVA). From the traditional test equality of several means, we find that the result of the overall means test easily contradicts the post hoc pairwise comparison. Statistical significance under the ANOVA model does not guarantee the consistent result of multiple comparisons between groups, and vice versa. On the other hand, ANOM can be displayed a clear graphical procedure which is similar to control chart with a center line and upper and lower decision limits. Once the treatment means fall outside of the decision limits, these certain treatment means are considered to be statistically different from the overall mean. In this paper, we used a single-stage sampling procedure ($\mathcal{P}_1$) and a modified single-stage sampling procedure ($\mathcal{P}_2$) to deal with the ANOM under heteroscedasticity (HANOM) to compare k population means.

Two single-stage sampling procedures ($\mathcal{P}_1$ and $\mathcal{P}_2$) for HANOM test were compared with the classical ANOM test with regard to type I error rate under different experimental conditions, and this similar comparison with ANOVA can be referred to Mendeş and Yiğit (2013). The simulation results showed that when the variances are homogeneous, ANOM tests have generally performed around 5% type I error rates, regardless of the sample size combinations. However, this is not valid in the case that the variances are not homogeneous. Type I error rates with regard to the ANOM tests were found to lie far from 5% criteria under heteroscedasticity. Therefore, under these conditions, the ANOM test for comparing treatment means is not recommended. We suggest the proposed $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures for the HANOM test is more suitable under the heteroscedasticity and various sample size conditions. Under all sample size combinations in the heteroscedastic situations, the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures are robust. Under the homogeneity situations, the $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures are more conservative with a fewer number of groups and smaller sample sizes. This result is similar to the traditional method of heteroscedasticity such as Brown-Forsythe test, Welch's F test or generalized variable based tests. Related comparisons can be referred to Brown and Forsythe (1974), Li et al. (2012), and Wang et al. (2019). However, the better simulation results could be obtained when there are more groups or at least 10 observations per treatment group. Bishop and Dudewicz (1978) also suggested that the sample size of the first stage in the two-stage sampling procedure contains at least 10 observations giving better results in practical studies. Therefore, we suggested that the initial sample of size n_i in the single-stage sampling procedure should be chosen at least 10 observations. In fact, as long as the number of groups is greater than 4 and the size of each group is greater than 5 observations, we can obtain the relatively good results.

Weerahandi and Krishnamoorthy (2019) and Cavus and Yazıcı (2020) both showed that the parametric bootstrap can control the type I error rates relatively well. Particularly, according to the simulation study by Cavus and Yazıcı (2020), the parametric bootstrap test controls the Type I error rates under various scenarios in fairly well results, whereas the fiducial test tends to be somewhat conservative. In the numerical example analysis, the difference in results has been

shown among the generalized fiducial test, generalized parametric bootstrap tests, and the single-stage sampling procedure for HANOM. We have found that both $\mathcal{P}_1$ and $\mathcal{P}_2$ procedures can easily identify the means having a significant difference from the overall mean, which is a convenient way to compare several population means. HANOM can avoid the contradiction between the overall means test and a post hoc pairwise comparison. The comparison between generalized tests and HANOM tests can be carried out in future studies.

From the simulation and the numerical example analysis above, despite the same results of the two procedures, we make some suggestions to choose a suitable procedure. If there is a considerable disparity in the sample sizes among treatments, we recommend using the procedure $\mathcal{P}_1$. Since the lower and upper decision lines of $\mathcal{P}_1$ depend on the different sample sizes, it is more efficient to identify which means are significantly different from the overall mean. On the other hand, if there is a disparity in the variances among treatments, we recommend using procedure $\mathcal{P}_2$. The reason is that the standard error of HANOM, $\sqrt{z^*}$, is the square root of the maximum of the S_i^2/n_i in which $\mathcal{P}_2$ is a conservative way.

In this paper, we focus on single-stage and modified single-stage sampling procedures for HANOM model in a one-way layout. For the design with more than one factor, two- or higher-way interaction should be considered carefully. In the recent studies, two-way ANOVA under heteroscedasticity were previously developed by Ananda, Dag, and Weerahandi (2022). Heteroscedastic two-way ANOVA are also available in *R* package *twowaytests* (Dag et al. 2021b). Therefore, HANOM in a two-way layout or higher-way model are encouraged to be developed under heteroscedastic variance case.

Acknowledgments

The authors are sincerely grateful to referees for their insightful comments and suggestions which helped to improve the quality of this paper.

Funding

The research was partially supported by the Ministry of Science and Technology, Taiwan, MOST 109-2118-M-006-004-MY2.

References

Ananda, M. M. A., O. Dag, and S. Weerahandi. 2022. Heteroscedastic two-way ANOVA under constraints, to appear in. *Communications in Statistics - Theory and Methods* :1–16. doi: 10.1080/03610926.2022.2059682.

Bishop, T. A. 1976. Heteroscedastic anova, manova., and multiple-comparisons. PhD diss. The Ohio State University.

Bishop, T. A., and E. J. Dudewicz. 1978. Exact analysis of variance with unequal variances: Test procedures and tables. *Technometrics* 20 (4):419–30. doi: 10.1080/00401706.1978.10489696.

Brown, M., and A. Forsythe. 1974. Small sample behavior of some statistics which test the equality of several means. *Technometrics* 16 (1):129–32. doi: 10.1080/00401706.1974.10489158.

Castillo-Cuenca, J. C., Á. Martínez-Moreno, J. M. Diaz-Cao, A. Entrena-García, J. Fraga, P. C. Arias, S. Almería, and I. García-Bocanegra. 2021. Seroprevalence of Toxoplasma gondii and associated risk factors in domestic pigs raised from Cuba. *Parasitology Research* 120 (8):2897–903. doi: 10.1007/s00436-021-07245-1.

Cavus, M., and B. Yazıcı. 2020. Testing the equality of normal distributed and independent groups' means under unequal variances by doex package. *The R Journal* 12 (2):134–54. doi: 10.32614/RJ-2021-008.

Chen, H. J., and K. Lam. 1989. Single-stage interval estimation of the largest normal mean under heteroscedasticity. *Communications in Statistics - Theory and Methods* 18 (10):3703–18. doi: 10.1080/03610928908830118.

Chen, H. J., M. J. Wen, and C. J. Chuang. 2011. On testing the equivalence of treatments using the measure of range. *Computational Statistics & Data Analysis* 55 (1):603–14. doi: 10.1016/j.csda.2010.06.002.

Chen, H. J., M. J. Wen, and A. M. Wang. 2009. On testing the bioequivalence of several treatments using the measure of distance. *Statistics* 43 (5):513–30. doi: 10.1080/02331880802497447.
Chen, S. Y. 2001. One-stage and two-stage statistical inference under heteroscedasticity. *Communications in Statistics: Simulation and Computation* 30 (4):991–1009. doi: 10.1081/SAC-100107792.
Chen, S. Y., and H. J. Chen. 1998. Single-stage analysis of variance under heteroscedasticity. *Communications in Statistics - Simulation and Computation* 27 (3):641–66. doi: 10.1080/03610919808813501.
Chen, S. Y., H. J. Chen, and H. F. Chang. 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. doi: 10.1081/SAC-120028433.
Dag, O. A. Dolgun, N. M. Konar, S. Weerahandi, M. Ananda, and M. O. Dag. 2021a. Package 'onewaytests'. https://cran.r-project.org/web/packages/onewaytests/index.html
Dag, O. A. Dolgun, S. Weerahandi, and M. Ananda. 2021b. Package 'twowaytests'. https://cran.r-project.org/web/packages/twowaytests/index.html
Games, P. A., and J. F. Howell. 1976. Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics* 1 (2):113–25.
Kiani, M., J. Panaretos, and S. Psarakis. 2008. A new procedure to monitor the mean of a quality characteristic. *Communications in Statistics - Simulation and Computation* 37 (9):1870–80. doi: 10.1080/03610910802296588.
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.
Li, X., L. Tian, J. Wang, and J. R. Muindi. 2012. Comparison of quantiles for several normal populations. *Computational Statistics & Data Analysis* 56 (6):2129–38. doi: 10.1016/j.csda.2012.01.002.
Lopez-Mejia, M., and E. Roldan-Valadez. 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: The Official Journal of National Stroke Association* 25 (3):515–22. doi: 10.1016/j.jstrokecerebrovasdis.2015.10.033.
Mendeş, M., and S. Yiğit. 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):2093–104. doi: 10.1080/00949655.2012.679942.
Nelson, P. R. 1982. Exact critical points for the analysis of means. *Communications in Statistics - Theory and Methods* 11 (6):699–709. doi: 10.1080/03610928208828263.
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* II:97–114.
Nelson, P. R., and E. J. Dudewicz. 2002. Exact analysis of means with unequal variances. *Technometrics* 44 (2): 152–60. doi: 10.1198/004017002317375109.
Nelson, P. R. P. S. Wludyka, and K. A. Copeland. 2005. *The analysis of means: A graphical method for comparing means, rates, and proportions*. Philadelphia, PA/Alexandria, VA: SIAM/ASA, 25.
Ott, E. R. 1967. Analysis of means - a graphical procedure. *Industrial Quality Control* 24 (2):101–9.
Pallmann, P., and L. A. Hothorn. 2016. Analysis of means: A generalized approach using *R*. *Journal of Applied Statistics* 43 (8):1541–60. doi: 10.1080/02664763.2015.1117584.
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):243–58. doi: 10.1214/aoms/1177731088.
Wang, Y., T. Pham, D. Nguyen, E. S. Kim, Y. H. Chen, J. Kromrey, Z. Yi, and Y. Yin. 2019. 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.
Weerahandi, S. 2004. *Generalized inference in repeated measures: Exact methods in MANOVA and mixed models*, 43. N.J.: John Wiley & Sons.
Weerahandi, S., and K. Krishnamoorthy. 2019. A note reconciling ANOVA tests under unequal error variances. *Communications in Statistics - Theory and Methods* 48 (3):689–93. doi: 10.1080/03610926.2017.1419264.
Wen, M. J., and H. J. Chen. 1994. Single-stage multiple comparison procedures under heteroscedasticity. *American Journal of Mathematical and Management Sciences* 14 (1-2):1–48. doi: 10.1080/01966324.1994.10737368.
Wen, M. J., S. Y. Chen, and H. J. Chen. 2007. On testing a subset of regression parameters under heteroskedasticity. *Computational Statistics & Data Analysis* 51 (12):5958–76. doi: 10.1016/j.csda.2006.11.018.
Wen, M. J., L. C. Huang, and J. Zhong. 2017. Single-stage sampling procedure of the t best populations under heteroscedasticity. *Communications in Statistics - Theory and Methods* 46 (18):9265–73. doi: 10.1080/03610926.2016.1206935.
Wen, M. J., C. C. Wen, and W. M. Wang. 2022. Single-stage sampling procedure for heteroscedasticity in multiple comparisons with a control. *To Appear in Communications in Statistics-Simulation and Computation* :1–10. doi: 10.1080/03610918.2022.2053863.
Wu, S. F., and H. J. Chen. 1998. Multiple comparisons with the average for normal distributions. *American Journal of Mathematical and Management Sciences* 18 (1-2):193–218. doi: 10.1080/01966324.1998.10737459.