Applications and Applied Mathematics: An International Journal (AAM)

Volume 16 | Issue 1 Article 5

6-2021

Analysis of Means (ANOM) Concepts and Computations

Kalanka P. Jayalath
University of Houston – Clear Lake

Jacob Turner
Stephen F. Austin State University

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Applications and Applied Mathematics:
An International Journal
(AAM)

Vol. 16, Issue 1 (June 2021), pp. 75 – 96

Analysis of Means (ANOM) Concepts and Computations

[1*]Kalanka P. Jayalath and [2]Jacob Turner

[1]Department of Mathematics and Statistics
University of Houston – Clear Lake
Houston, Texas 77058, USA
[1]jayalath@uhcl.edu

[2]Department of Mathematics and Statistics
Stephen F. Austin State University
Nacogdoches, Texas 75962, USA
[2]turnerja2@sfasu.edu

* Corresponding Author

Received: September 17, 2020; Accepted: January 17, 2021

Abstract

The classical Analysis of Means (ANOM) is a statistical inferencing procedure and visualization tool to analyze means from experiments with fixed effects. It can serve as an alternative to the Analysis of Variance (ANOVA) procedure that has distinct advantages when determining which effects contributed to an overall test's significant result. ANOM has been extended to handle numerous situations including robust procedures involving ranks. More recent advancements of this procedure allow one to handle both random, and mixed effect models. In this work, we discuss the recent developments on ANOM methods that are useful in practice, provide examples that illustrate their effectiveness, and discuss logistical issues with post hoc testing.

Keywords: Analysis of Means; Analysis of variance; Fixed Effect; Multiple Comparison; Random Effect

MSC 2010 No.: 62K99, 62A01, 62P30

1. Introduction

The analysis of means (ANOM) is a graphical testing procedure similar to Shewhart control charts that can be used for multiple group comparisons. The ANOM has several advantages when compared to traditional analysis of variance (ANOVA). At times, introducing ANOVA to beginners and practitioners with lack of statistical knowledge is challenging. The notion of splitting variability present in data into possible sources to compare and contrast means seems conceptually awkward when advancing from one and two sample t-tests. Additionally, making inferences on treatment means using ANOVA is a two-stage process. First, an overall F-test is conducted to see whether there is at least one treatment mean significantly different from the rest. If concluded to be true, then a post-hoc test is applied to identify which specific means caused the observed significance.

An unfortunate consequence of post hoc testing in ANOVA is that they do not always correspond to the decision of the overall F-test. For example, the overall F-test might be rejected, but all pairwise comparisons from a Tukey or Bonferroni procedure yield no significant differences. Alternatively, ANOM is a simple graphical testing procedure that simultaneoulsy compares each individual mean to their overall mean. The visual comparisons are derived so that the family wise error rate is controlled and correspond directly to the overall ANOM testing procedure.

Due to its integration of statistical inference within a visual frame work, the ANOM procedure has become increasingly popular among practitioners and data scientists. For instance, ANOM can easily be used as a pretest to test the randomness of sampled data by plotting systematic group means at early stages in survey sampling problems. ANOM has also been applied in many areas of research including medicine, health care, quality management, and environmental sciences (Mohammed and Holder (2012); Prokeš et al. (2017); Delvoye et al. (2009); Homa (2007); Murthy et al. (2018)). In terms of software accessibility, it has been implemented in many statistical packages including SAS (PROC ANOM), MINITAB, JMP, and R 'ANOM' package (Pallmann and Hothorn (2016)).

The ANOM procedure was first introduced by Ott (1967) and later appeared in Schilling (1973) and Ott et al. (1975). The contributions of Nelson (1982), Nelson (1983), and Nelson (1988) can be considered substantial due to the development of exact ANOM critical values and its introduction to factorial treatment structures. The critical values for unbalanced designs were reported in Stoline and Ury (1979) and Ury et al. (1980). A complete set of critical value tables for many designs including sample size calculations can be found in Nelson et al. (2005). Use of ANOM in one-way factorial experiments and analyzing proportions and counts were discussed in Ramig (2016). A non-parametric alternative using ranks (ANOMR) was introduced by Bakir (1989) and may be considered as another milestone in ANOM literature. Analysis of means for variances (ANOMV) proposed by Wludyka and Nelson (1997) allows testing homogeneity of variances among groups graphically. The ANOM procedure was later extended to accompany heteroscedastic data (HANOM) by Nelson and Dudewicz (2002). Recently, Jayalath and Ng (2018) and Jayalath and Ng (2020) developed the ANOMQ procedure (ANOM using studentized range or q distribution) to facilitate random factors in common statistical designs.

One particular disadvantage of ANOM compared to ANOVA is the implementation of the procedure in more complex study designs. The difficulty is mainly due to deriving its sampling distribution. The ANOM procedure was primarily developed to analyze fixed factor effects, where the levels of the factor consist of all possible levels. When the levels of a factor is a random sample from a larger population of possible levels, the factor is referred to as a random effect and an adjustment to the sampling distribution is required. For instance, when testing means in a fixed factor design, ANOM utilizes the equi-correlated multivariate t-distribution. When the factor is a random effect, an adjustment should be made using ANOMQ that relies on the studentized range distribution (Jayalath and Ng (2018)). However, regardless of the factor being fixed or random, the ANOVA procedure uses the univariate F-distribution.

Mendeş and Yiğit (2013) conducted a comprehensive simulation study to compare and contrast ANOVA and ANOM with regards to their type-I error and statistical power in single fixed factor experiments. They concluded that for homogeneous data, both ANOVA and ANOM tests have similar type-I error rates but both tests were negatively affected by the degree of heterogeneity of the variances. They also note that for homogeneous data, the power of the tests was equally affected by the combination of sample sizes and their variance ratios. For unbalanced designs, they concluded that the ANOVA test is somewhat more powerful than the ANOM test. Further, they noticed that the size of the group mean with respect to its variance also plays a role in the discrepancy of power between the two procedures.

Many of the classical ANOM procedures such as ANOM, ANOMV, HANOM and ANOMR are readily available in Nelson et al. (2005) and this work focuses on the latest developments on ANOM methods that are useful in practice. Therefore the main focus of this paper is on the computation aspect of ANOM charts for both balanced and unbalanced completely randomized designs with fixed and random factor effects.

In Section 2, we review the current ANOM, ANOMQ, and ANOMR methods and extend the ANOMQ procedure to analyze data from unbalanced designs. Section 3 focuses on applications and we exhibit various situations where these methods can appropriately be applied by emphasizing their calculations and inferential aspects when compared to ANOVA. Section 4 provides a discussion on the ANOM procedures general qualities and reflects on some issues the procedures have in regards to post hoc testing. In Section 5, we provide our concluding remarks.

2. Analysis of Means Approach

In this section, we discuss the ANOM procedures for both balanced and unbalanced single factor designs in the presence of fixed and random factors. That is, we exhibit ANOM, ANOMQ, and ANOMR procedures. The discussion is then extended for two-factor factorial treatment structure.

2.1. Balanced Design

Let us consider a single factor completely randomized design (CRD) with t treatment levels each with n observations. The corresponding fixed effects model can be written as

$$y_{ij} = \mu + \tau_i + \epsilon_{ij},\ i = 1, 2, \ldots, t,\ j = 1, 2, \ldots, n, \tag{1}$$

where $N = nt$ is the total number of observations and $\tau_i = \mu - \mu_i$ is the fixed effect of the ith treatment level with respect to the overall mean μ. We also assume that the random errors $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

For this model, observed treatment average $\bar{y}_{i\bullet} = \sum_{j=1}^{n} y_{ij}/n$ is an unbiased estimator for μ_i. Highly disperse $\bar{y}_{i\bullet}$'s would indicate significant mean effects. On the other hand, if none of the effects is significantly different from the others, then the $\bar{y}_{i\bullet}$'s must be close to the overall average $\bar{y}_{\bullet\bullet} = \sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij}/N$. In summary, the ANOM tests to see if at least one of the μ_i's is significantly different from the overall mean μ and graphically identifies exactly which treatment means cause the observed significance. That is, in ANOM we plot the $\bar{y}_{i\bullet}$'s along with $\bar{y}_{\bullet\bullet}$ in a decision chart similar in appearance to a control chart to visualize the mean effects. Nelson (1981) indicates that the joint distribution of the absolute mean differences $|\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}|$ becomes an equi-correlated t-dimensional t-distribution with correlation $\rho = -1/(t-1)$ with degrees of freedom $N - t$. This distribution is also known as the Studentized maximum absolute deviation distribution or the h-distribution. Its probability calculations are reported in Nelson (1993) and Nelson et al. (2005).

For this single-factor model, the following upper and lower decision limits, denoted as UDL and LDL, respectively, will be used to make appropriate decisions about the mean effects:

$$UDL = \bar{y}_{\bullet\bullet} + h(\alpha; t, N-t)\sqrt{(t-1)MSE/N}, \tag{2}$$

$$LDL = \bar{y}_{\bullet\bullet} - h(\alpha; t, N-t)\sqrt{(t-1)MSE/N}, \tag{3}$$

where $MSE = \sum_{i=1}^{t} s_i^2/t$, s_i is the sample standard deviation of the i-th treatment level, $N - t$ is the degrees of freedom for the MSE, and t is the number of means being compared. The critical value $h(\alpha, t, N-t)$ is the upper 100α percentage point of the h-distribution.

When the factor τ in the model in Equation (1) becomes random, we write the random effect model as below.

$$y_{ij} = \mu + a_i + \epsilon_{ij},\ i = 1, 2, \ldots, t,\ j = 1, 2, \ldots, n. \tag{4}$$

Here we assume that the factor $a_i \overset{i.i.d.}{\sim} N(0, \sigma_a^2)$ and it is independent from the ϵ_{ij}. Note that, when the factor is fixed, it affects on the mean of y_{ij}, i.e., $E(y_{ij}) = \mu + \tau_i$ and $Var(y_{ij}) = \sigma^2$, but when the factor is random, it affects on the variability of y_{ij}, i.e, $E(y_{ij}) = \mu$ and $Var(y_{ij}) = \sigma^2 + \sigma_a^2$.

Therefore, in order to test the effect of the random effect we should consider the hypotheses $H_0 : \sigma_a^2 = 0$ vs $H_a : \sigma_a^2 \neq 0$. To test the significant dispersion among the treatment means, Jayalath and Ng (2018) suggested using the ANOMQ procedure that relies on the Studentized-

range (q) distribution. That is, plot $\bar{y}_{i\bullet}$ along with following UDL and LDL.

$$UDL = \hat{\omega} + \frac{1}{2} q_{(\alpha;t,N-t)} \sqrt{MSE/n}, \tag{5}$$

$$LDL = \hat{\omega} - \frac{1}{2} q_{(\alpha;t,N-t)} \sqrt{MSE/n}, \tag{6}$$

where $\hat{\omega} = (\bar{y}_{[1]\bullet} + \bar{y}_{[t]\bullet})/2$ is the mid-range estimate, and $\bar{y}_{[i]\bullet}$ is the ith ordered treatment average.

Jayalath and Ng (2018) indicates that the widths of the ANOM charts are generally wider than that of ANOMQ for more practical cases. However, they reported a few instances where this finding is inconsistent. For instance, widths of both decision limits become approximately equal when only two treatments means ($t = 2$) are compared at $\alpha = 0.01, 0.05$ and 0.10 levels and the widths of ANOMQ decision limits become wider when $t \geq 15$ and $N - t = 2$ at $\alpha = 0.01$ and $t \geq 16$ and $N - t = 3$ at $\alpha = 0.001$. This indicates that the decision limits in the ANOM procedure should be selected depending on the random and fixed nature of the factors in the model. However, the standard ANOVA does not exhibit such a change in its p-value calculations as it uses the F-distribution to test both fixed and random factor effects.

2.2. Unbalanced Design

We reconsider the model given in Equation (1) with t treatment levels but each with varying number of observations. That is, let $i = 1, 2, \ldots, t$, and $j = 1, 2, \ldots, n_i$. Under the regular assumptions, Nelson (1989) suggested the following decision limits to test means.

$$UDL = \bar{y}_{\bullet\bullet} + m(\alpha; t, N-t)\sqrt{(N-n_i)MSE/Nn_i}, \tag{7}$$

$$LDL = \bar{y}_{\bullet\bullet} - m(\alpha; t, N-t)\sqrt{(N-n_i)MSE/Nn_i}, \tag{8}$$

where $N = \sum_{i=1}^{t} n_i$, $MSE = \sum_{i=1}^{t} (n_i - 1)s_i^2/(N-t)$ and $m(\alpha; t, N-t)$ values are given in Nelson (1989) and Nelson et al. (2005). As one expects, the limits for each treatment will depend on their sample size (n_i).

Analyzing random factor effects using ANOMQ for unbalanced designs needs careful attention. The treatment average of the ith treatment can be written as $\bar{y}_{i\bullet} = \mu + a_i + \bar{\epsilon}_{i\bullet}$. This indicates that $Var(\bar{y}_{i\bullet}) = \sigma_a^2 + \sigma^2/n_i$. Therefore, using the same arguments in Jayalath and Ng (2018), under the null hypothesis $H_0 : \sigma_a^2 = 0$, the ratio

$$\frac{\bar{y}_{[t]\bullet} - \bar{y}_{[1]\bullet}}{\hat{\sigma}\sqrt{\frac{1}{2}\left(\frac{1}{n_{[t]}} + \frac{1}{n_{[1]}}\right)}} \sim q_{(t,N-t)},$$

where $\hat{\sigma} = \sqrt{MSE}$ and $\bar{y}_{[i]\bullet}$ is the ith ordered treatment average with sample size $n_{[i]}$.

To test the significance of the dispersion among treatment means we plot $\bar{y}_{i\bullet}$ along with the sample mid-range estimate $\hat{\omega} = (\bar{y}_{[1]\bullet} + \bar{y}_{[t]\bullet})/2$. That is, we may declare the overall variability is significant

if at least one $\bar{y}_{i\bullet}$ falls outside the following ANOMQ control limits:

$$UDL = \hat{\omega} + \frac{1}{2} q_{(\alpha;t,N-t)} * \sqrt{\frac{MSE}{2}\left(\frac{1}{n_{[1]}} + \frac{1}{n_{[t]}}\right)}, \tag{9}$$

$$LDL = \hat{\omega} - \frac{1}{2} q_{(\alpha;t,N-t)} * \sqrt{\frac{MSE}{2}\left(\frac{1}{n_{[1]}} + \frac{1}{n_{[t]}}\right)}. \tag{10}$$

2.3. Robust ANOM Charts

The ANOM procedure is fairly robust to the departure of the normality assumption. However, Mendeş and Yiğit (2013) indicates that its statistical power depends on how well the homogeneity of the variance assumption is satisfied. Data transformation is preferred when data are non-normal and heterogeneous. In the cases where transformation is infeasible or inadequate, non-parametric methods such as rank-based tests may be preferred. Analysis of means using ranks (ANOMR) proposed by Bakir (1989) is a useful candidate for such data. The ANOMR procedure is a graphical alternative for the popular Kruskal-Wallis (K-W) test.

To explain the ANOMR procedure we reconsider the single fixed factor unbalanced design discussed in Section (2.2) and assume that t populations of interest have a similar shape and at most different in their location parameters. Like in the K-W test, we first assign ranks r_{ij} ($i = 1, 2, \ldots, t$, and $j = 1, 2, \ldots, n_i$) for all the responses y_{ij} ignoring their group memberships in the combined sample of size $N = \sum_{i=1}^{t} n_i$. Let $\bar{R}_{i\bullet}$ denotes ith treatment rank average, that is, $\bar{R}_{i\bullet} = \sum_{j=1}^{n_i} r_{ij}/n_i$ and the overall average of the ranks is $\bar{R}_{\bullet\bullet} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} r_{ij}/N = (N+1)/2$.

Then, to test the null hypothesis that all t populations have exact same location parameter, the calculated rank averages $\bar{R}_{i\bullet}$ are plotted along with the following upper and lower decision limits to conduct the ANOMR test.

$$UDL = \bar{R}_{\bullet\bullet} + C(\alpha, t, n_1, n_2, ..., n_t), \tag{11}$$

$$LDL = \bar{R}_{\bullet\bullet} - C(\alpha, t, n_1, n_2, ..., n_t), \tag{12}$$

where $C(\alpha, t, n_1, n_2, ..., n_t)$ is a constant that satisfies $P\left(\underset{1\le i\le t}{Max}|\bar{R}_{i\bullet} - \bar{R}_{\bullet\bullet}| \ge C\right) = \alpha$ under the null hypothesis. For equal sample sizes, it is recommended to apply Bonferroni adjustment to obtain suitable decision limits using $C^*(\alpha, t; n) = w(\frac{\alpha}{2t}, n(t-1), n)/n - (N+1)/2$ in place of $C(\alpha, t; n_1, n_2, ..., n_t)$, where $w(\frac{\alpha}{2t}, n(t-1), n)$ is the upper $100(\frac{\alpha}{2t})\%$ percentile point of the Wilcoxon rank sum statistic with sample sizes $n(t-1)$ and n. However, obtaining exact critical values $C(\alpha, t, n_1, n_2, ..., n_t)$ become computationally expensive for even relatively moderate t and n_i values. Bakir (1989) provided a limited set of exact critical values for a few specific significance levels.

Due to unavailability of exact critical values for moderate sample sizes, Bakir (1989) suggested using asymptotic procedures. Further, they clarified that the asymptotic joint distribution of the $|\bar{R}_{i\bullet} - \bar{R}_{\bullet\bullet}|$ is the same as that of the $|\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}|$. Therefore, it is recommended using the ANOM

decision limits given in Equations (2), (3), (7), and (8) by replacing the observed data y_{ij} by their ranks r_{ij}.

2.4. Factorial Treatment Structure

The ANOM procedure can be used to analyze data from completely randomized designs with factorial treatment structures. Analysis of the fixed effects in the factorial structure is well developed in ANOM and analysis of random effects can be found in Jayalath and Ng (2018).

Let us consider a model with two fixed factors (say A and B) in a factorial treatment structure

$$y_{ijl} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijl},\ i = 1, 2, \ldots, t,\ j = 1, 2, \ldots, k,\ l = 1, 2, \ldots, n, \tag{13}$$

where $\epsilon_{ijl} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $N = ntk$.

To test the interaction effect, Nelson (1988) suggested to plot $z_{j(ii')} = x_{j(ii')} - \bar{x}_{\bullet(ii')}$ along with the following decision limits, where $x_{j(ii')} = \bar{y}_{ij\bullet} - \bar{y}_{i'j\bullet}$ are the interaction slopes.

$$UDL = 0 + g(\alpha, (t, k), tk(n-1))\sqrt{MSE}\sqrt{2(k-1)/kn}, \tag{14}$$

$$LDL = 0 - g(\alpha, (t, k), tk(n-1))\sqrt{MSE}\sqrt{2(k-1)/kn}, \tag{15}$$

where $MSE = \sum_i \sum_j \sum_l \frac{(y_{ijl} - \bar{y}_{ij\bullet})^2}{tk(n-1)}$ and $g(\alpha, (t, k), tk(n-1))$ values are given in Nelson (1988) and Nelson et al. (2005). When one of the factors has only two levels, say $t = 2$, we plot slopes $x_{j(12)} = \bar{y}_{1j\bullet} - \bar{y}_{2j\bullet}$ along with the following decision limits:

$$UDL = \bar{x}_{\bullet 12} + h(\alpha, k, 2k(n-1))\sqrt{MSE}\sqrt{2(k-1)/kn}, \tag{16}$$

$$LDL = \bar{x}_{\bullet 12} - h(\alpha, k, 2k(n-1))\sqrt{MSE}\sqrt{2(k-1)/kn}, \tag{17}$$

where $\bar{x}_{\bullet 12} = \sum_{j=1}^{k} x_{j(12)}/k$. In either case, when the interaction is insignificant, one can continue to test the main effects as follows.

Let f represent the number of levels for the factor of interest. That is, in our stetting $f = t$ for factor A and $f = k$ for factor B. Then, to test the significance of the main effects, the treatment averages are suggested to plot along with the following limits:

$$UDL = \bar{y}_{\bullet\bullet\bullet} + h(\alpha, f, tk(n-1))\sqrt{MSE}\sqrt{(f-1)/N}, \tag{18}$$

$$LDL = \bar{y}_{\bullet\bullet\bullet} - h(\alpha, f, tk(n-1))\sqrt{MSE}\sqrt{(f-1)/N}, \tag{19}$$

For random factor analysis, we consider the following model:

$$y_{ijl} = \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijl},\ i = 1, 2, \ldots, t,\ j = 1, 2, \ldots, k,\ l = 1, 2, \ldots, n, \tag{20}$$

where $a_i \overset{i.i.d.}{\sim} N(0, \sigma_a^2)$, $b_j \overset{i.i.d.}{\sim} N(0, \sigma_b^2)$, $(ab)_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_{ab}^2)$ and $\epsilon_{ijl} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $a_i, b_j, (ab)_{ij}$ and ϵ_{ij} are mutually independent. In this model all three factors are considered to be random.

Jayalath and Ng (2018) considered the same model and discussed a case where factor A has only two levels, $t = 2$. That is, they suggested plotting slopes $x_{j(12)} = \bar{y}_{1j\bullet} - \bar{y}_{2j\bullet}$ along with the following decision limits to test the random interaction effect, $H_0 : \sigma^2_{ab} = 0$.

$$UDL = m_{\bar{ab}} + \frac{1}{2} q_{(\alpha;k,tk(n-1))} \sqrt{2MSE/n}, \tag{21}$$

$$LDL = m_{\bar{ab}} - \frac{1}{2} q_{(\alpha;k,tk(n-1))} \sqrt{2MSE/n}, \tag{22}$$

where $m_{\bar{ab}} = (x_{[k](12)} + x_{[1](12)})/2$ is the mid-range estimate and in which $x_{[j](12)}$ is the *j*th order statistic.

Then, one can continue to test main effects for factors A and B. The following decision limits should be used for testing random factor effect for B.

$$UDL = m_{\bar{b}} + \frac{1}{2} q_{(\alpha;k,(t-1)(k-1))} \sqrt{\hat{\sigma}^2_{ab}/nt}, \tag{23}$$

$$LDL = m_{\bar{b}} - \frac{1}{2} q_{(\alpha;k,(t-1)(k-1))} \sqrt{\hat{\sigma}^2_{ab}/nt}, \tag{24}$$

where $m_{\bar{b}} = (\bar{y}_{\bullet[k]\bullet} + \bar{y}_{\bullet[1]\bullet})/2$ in which $\bar{y}_{\bullet[j]\bullet}$ is the *j*th order statistic and $\hat{\sigma}^2_{ab} = \sum_i \sum_j \frac{(y_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2}{(t-1)(k-1)}$ is the mean squares due to AB interaction.

The following decision limits should be used for testing random factor effect for A.

$$UDL = m_{\bar{a}} + \frac{1}{2} q_{(\alpha;t,(t-1)(k-1))} \sqrt{\hat{\sigma}^2_{ab}/nk}, \tag{25}$$

$$LDL = m_{\bar{a}} - \frac{1}{2} q_{(\alpha;t,(t-1)(k-1))} \sqrt{\hat{\sigma}^2_{ab}/nk}, \tag{26}$$

where $m_{\bar{a}} = (\bar{y}_{[t]\bullet\bullet} + \bar{y}_{[1]\bullet\bullet})/2$.

3. Illustrative Examples

Example 3.1.

Wine Quality Data: This data set contains wine measurements that include 11 physicochemical variables and one quality variable, an aggregate score from multiple wine judges, taken on 4898 white wines reported in Cortez et al. (2009). The goal of the original study was to model wine quality based on physicochemical variables. In spite, we would like to test whether the pH values of white wines depend on the quality of the wines. The pH values of these wines range from 2.72 through 3.82 and the quality score ranged from three to nine. Thus, for our purpose, the quality score will be viewed as an ordinal categorical factor with larger quality scores corresponding to better quality.

Table 1. The highest ten pH values of wines from each quality score

Quality	3	4	5	6	7	8
pH	3.55	3.72	3.79	3.81	3.82	3.59
	3.53	3.65	3.77	3.80	3.76	3.57
	3.44	3.63	3.77	3.80	3.70	3.56
	3.42	3.53	3.74	3.76	3.66	3.55
	3.37	3.53	3.69	3.75	3.65	3.55
	3.31	3.52	3.67	3.75	3.64	3.55
	3.24	3.51	3.66	3.74	3.64	3.53
	3.24	3.49	3.66	3.72	3.61	3.47
	3.24	3.49	3.66	3.72	3.60	3.46
	3.23	3.49	3.63	3.69	3.59	3.45
$\bar{y}_{i\bullet}$	3.357	3.556	3.704	3.754	3.667	3.528
s_i	0.124	0.081	0.058	0.039	0.073	0.050

To illustrate a balanced design setting for our first example, we select samples that contains the highest 10 pH values from each of the quality groups. We discarded the data from the highest quality score group (Quality 9) as it has only five observations. The selected data set, including its summary statistics, is provided in Table 1. Upon examination, this data exhibit a slight deviation from normality and the data from quality 3 show higher variability than the data from the remaining quality scores. However, as both ANOVA and ANOM are robust against such departures of the assumptions, we continue with the current data without seeking any transformation.

The sampled data follow the single fixed factor model described in Section (2.1). Therefore, we first apply the ANOVA test and its results are shown in Table 2. The *p*-value from the ANOVA F-test (< 0.001) indicates that there is at least two wine quality scores that have significantly different average pH values at a 5% significance level. However, the ANOVA table does not indicate which pair is significant or how many significant pairs among the six factor levels without further testing but rather quantifies the overall significance by its *p*-value.

Table 2. ANOVA output for the highest ten pH values of wines from each quality score

	Df	Sum Sq	Mean Sq	F value	p-value
Quality	5	1.0500	0.20999	36.40	< 0.001
Residuals	54	0.3115	0.00577		

On the other hand, to exhibit how ANOM could answer those questions, we reanalyze this data applying the single factor ANOM chart using Equations (2) and (3). As there are $t = 6$ quality groups each with $n = 10$ observations, and $N = 60$, the balanced ANOM critical value $h(0.05; 6, 54) = 2.72$. Using the standard deviations of each group, we calculate $MSE = (0.124^2 + 0.081^2 + 0.058^2 + 0.039^2 + 0.073^2 + 0.050^2)/6 = 0.0058$. As a result, the margin

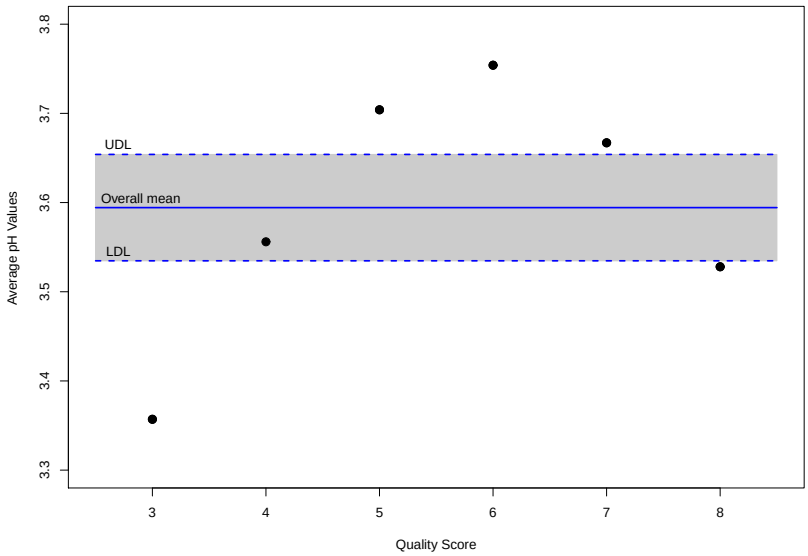

Figure 1. ANOM chart for the highest ten pH values of vines from each quality score

of error of the decision limits becomes $h(\alpha; t, N-t)\sqrt{(t-1)MSE/N} = 0.0596$. Therefore, the lower and upper decision limits for this data are $LDL = 3.5943 - 0.0596 = 3.5347$, $UDL = 3.5943 + 0.0592 = 3.6540$, respectively. To test the equality of mean pH among quality groups, the average pH values for each quality score ($\bar{y}_{i\bullet}$'s) are plotted along with the decision limits as shown in Figure 1.

Similar to ANOVA, as the averages fall outside the decision limits of the ANOM chart (Figure 1), it rejects the null hypothesis of equality of means at the 5% significance level. Moreover, this ANOM chart indicates that the average of the top 10 pH values for wines with the lowest quality score (quality 3) and the highest (quality 8) are significantly lower than the overall average and the wines scoring the middle (quality 5, 6, and 7) show significantly higher averages. Wines with quality 4 seem to show an insignificant departure from the overall average. Further, this ANOM chart indicates that wine quality may not be able to be uniquely determined by their highest pH values because of the evident parabolic curve. If this were the case, one would expect the mean pH against quality to be more linear if the ranks are ignored. Indeed, the ANOM helps visualizing existing patterns in the data where other standard tests such as ANOVA may not naturally reveal as much information.

Example 3.2.

Wine Quality Data-Random Sampling: In this example, we reconsider the same wine database from Example 3.1. However, in this case, we randomly sample 20% of the data from each group of quality except from quality 3 and 9 due to their low counts. For those two groups, we select all the available data. The resulting data summaries are shown in Table 3.

Table 3. Summary of the sampled wine data (20%) from each quality score

Quality	3	4	5	6	7	8	9
$\bar{y}_{i\bullet}$	3.188	3.181	3.145	3.192	3.220	3.205	3.308
s_i	0.210	0.150	0.130	0.160	0.163	0.154	0.083
n_i	20	33	290	440	176	35	5

This is an unbalanced data set with seven fixed factor levels ($t = 7$). To analyze this data, we apply the single factor unbalanced design discussed in Section (2.2). The data do not indicate any serious departure from normality. The assumption of variance homogeneity is violated due to comparatively small variability of the data in quality 9, but the other groups show similar standard deviations. Also note, that the small variability for quality 9 seems to coincide with its smaller sample size. However, we apply both ANOM and robust ANOMR to offer comparisons between the two procedures.

Table 4. ANOVA output for randomly sampled wines (20%) from each quality score

	Df	Sum Sq	Mean Sq	F value	p-value
Quality	6	0.783	0.13053	5.597	< 0.001
Residuals	992	23.134	0.02332		

The ANOVA table for this data is shown in Table 4. The ANOVA result indicates that the null hypothesis of equality of means is rejected at the 5% significance level.

For this data, $N = 999$ and the unbalanced ANOM critical value $m(0.05; 7, 992) = 2.68$. Using the overall average $\bar{y}_{\bullet\bullet} = 3.184$ and the $MSE = 0.0233$, we calculate the upper and lower ANOM decision limits given in Equations (7) and (8) for each wine quality score separately. As expected, for the unbalanced design, the decision limits for each quality score are varying commensurate with their sample sizes (Figure 2).

Based on the decision limits shown in Figure 2, this ANOM test rejects the null hypothesis of equality of means providing the same conclusion as of ANOVA. Besides, it indicates that wines with a quality scores 5 and 7 cause this significance. In practice, it may be of immediate interest to further investigate the causes for such departures in these two groups. It is interesting to note that though the average pH of quality 7 ($\bar{y}_{7\bullet} = 3.220$) is somewhat close to the overall mean ($\bar{y}_{\bullet\bullet} = 3.184$) it becomes significant, and on the other hand, though the average pH of wines with the highest quality ($\bar{y}_{9\bullet} = 3.308$) shows the highest departure it is not significantly different from the overall average. This is mainly due to the corresponding higher and lower sample sizes of those two groups ($n_7 = 176, n_9 = 5$). This indicates how one can use ANOM charts to further investigate practical significance of the observed results questioning observed statistically significant and insignificant results via visualization.

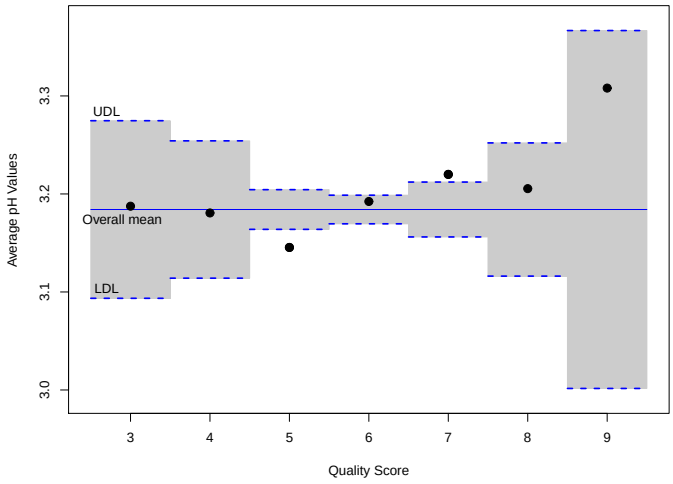

Figure 2. ANOM chart for pH values from randomly sampled wines (20%) from each quality score

As this data indicated some departures from variance homogeneity between quality scores, it may be interesting to see how well the ANOMR method handles this data. As explained in Section (2.3), to apply the ANOMR procedure, we first combined all the data ignoring the quality score groupings and ranked them numerically from one through 999. The resulting rank summaries for each group of wine quality are reported in Table 5.

Due to large sample sizes, it is impossible to obtain the exact critical values for ANOMR and therefore we rely on the asymptotic ANOMR decision limits. That is, we use the exact ANOM critical value $m(0.05; 7, 992) = 2.68$ in this ANOMR decision chart. For the ranked pH values, the overall average $\bar{R}_{\bullet\bullet} = 500.00$ and the $MSE = 80740.94$. Then, using Equations (7) and (8), we calculate upper and lower decision limits for the average ranked pH values for each quality score and plotted them in Figure 3.

Table 5. Summary of the ranked data

Quality	3	4	5	6	7	8	9
$\bar{R}_{i\bullet}$	405.750	490.636	458.048	483.809	606.318	570.857	558.400
$s(R)_i$	285.773	314.273	280.061	285.521	277.625	299.408	312.616
n_i	20	33	290	440	176	35	5

Similar to both ANOVA and ANOM, this test also indicates a significantly different central locations among the quality scores at the 5% significance level. Also, in both the ANOM and ANOMR analyses quality 5 and 7 wines consistently become significant. It is interesting to note that this ANOMR chart (Figure 3) separated low and high quality wines into two separate groups by the overall mean. However, we may need further analyses to conclude the significance of this particular grouping. That is, similar to ANOVA we need additional tests to compare specific contrasts of interest though its visual display may shed some light in certain contrast comparisons.

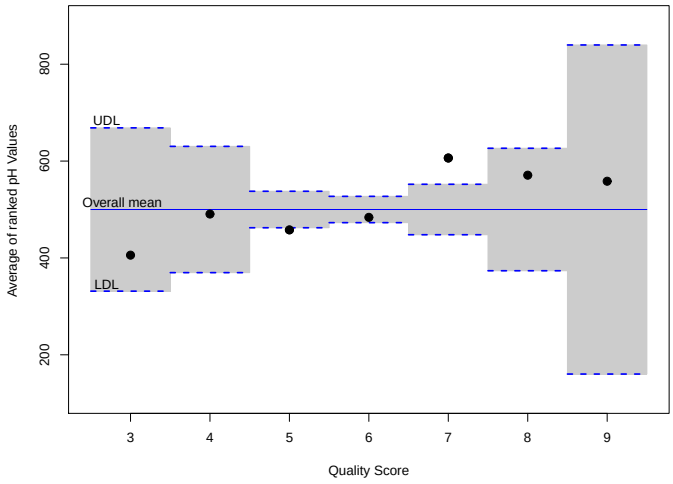

Figure 3. ANOMR chart for ranked pH values from randomly sampled wines (20%) from each quality score

In summary, since the raw data showed evidence against the homogeneity of the variance assumption, we may favor ANOMR findings when compared to ANOM. Even though all three procedures provide a consistent conclusion about the main hypothesis of equality of central location of pH values among quality scores, graphical testing provides virtuous information compared to the *p*-value reported from the ANOVA F-test.

Example 3.3.

Telephone Ringing: In this example, we consider a hypothetical telephone ringing experiment reported in Magezi (2015) where a researcher is interested in how quickly human listeners can detect a telephone ringing in the presence of concurrent speech.

For our discussion, we discard some of the factors they originally reported. However, the response variable of interest is the average reaction time (RT) in milliseconds and the design factor is the listeners. As the listeners is a random sample of a set of possible listeners, it is considered as a random factor. The resulting data and summary statistics are reported in Table 6.

Table 6. Telephone Ringing data

	L1	L2	L3	L4	L5	L6	L7	L8	L9
	296.82	1203.90	1373.48	676.91	1196.12	1643.68	708.04	1390.64	1946.32
	285.55	1159.17	1336.72	709.46	1293.70	1723.51	715.77	1358.31	1904.58
	300.03	1161.27	1373.52	580.08	1249.96	1689.71	757.79	1283.64	1987.01
	264.63	786.23	1018.99	532.64	929.54	1190.55	673.56	1047.35	1455.13
	234.56	869.69	999.06	577.31	953.03	1236.43	610.81	1015.54	1501.93
	178.34	922.95	987.12	566.63	943.91	1264.81	480.25	891.52	1472.95
	128.99	552.84	756.14	354.54	684.76	753.43	454.00	581.74	943.28
	153.88	551.63	747.83	407.98	648.58	796.68	424.30	712.05	993.63
	65.11	539.44	657.57	287.05	615.26	856.15	431.35	556.07	943.28
$\bar{y}_{i\bullet}$	211.990	860.791	1027.826	521.400	946.096	1239.439	583.986	981.873	1460.901
s_i	84.121	273.913	279.790	143.142	260.352	384.492	136.089	321.648	427.981

To analyze this data we first apply the standard ANOVA by employing the single random factor model given in Equation (4). The resulting ANOVA table is shown in Table 7 and it indicates there is significant variability among the listeners at the 5% significance level.

Table 7. ANOVA output for Telephone Ringing data

	Df	Sum Sq	Mean Sq	F value	p-value
Listeners	8	10487254	1310907	16.82	< 0.001
Residuals	72	5613115	77960		

On the other hand, since the factor is random we employ the balanced ANOMQ procedure discussed in Section (2.1) to conduct the graphical testing. For this data, $t = 9, n = 9, N = 81$, and the critical value is $q(0.05; 9, 72) = 4.523$. The mid-range estimate for the average reaction time is $\hat{\omega} = (1460.901 + 211.990)/2 = 836.446$ and the $MSE = 77959.94$. Applying this information in Equations (5) and (6), we calculated the upper and lower ANOMQ decision limits shown in Figure 4.

The ANOMQ chart visualizes each listeners' direct contribution to the overall variability and it indicates that listeners L1, L4, L6, L7, and L9 deviate significantly from the mid-range estimate. Therefore, we concluded that the variability of the average reaction time among listeners is significant and this result is consistent with the conclusion arrived at the ANOVA test. However, the main interest is to estimate the variability due to listeners, and not so much on listeners' averages and their differences unless the selected listeners are identified with specific group characteristics. The graphical test of ANOMQ helps to identify main causes of the observed significant variability due to the listener. Since the listeners represent a random sample and perhaps from a large population, we may generalize this finding to the entire population. Based on the available information and the results obtained from the ANOMQ chart, there are no additional conclusions regarding the nature, trend, or cause of the observed significance can be provided.

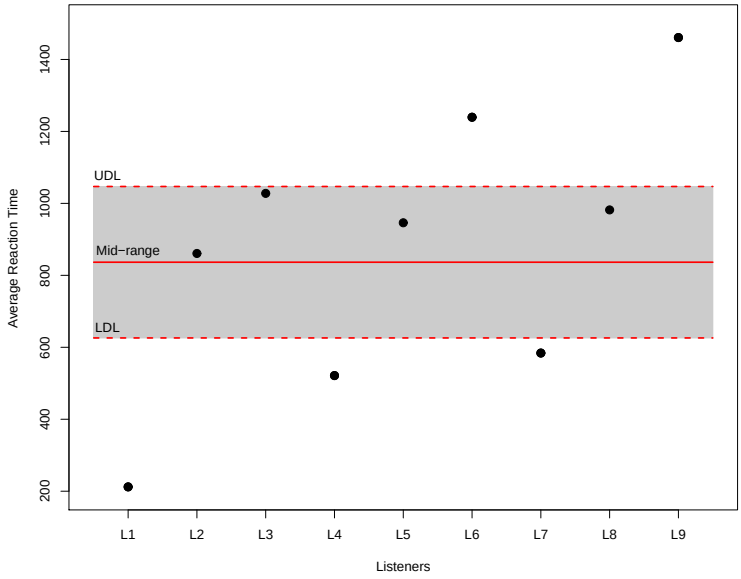

Figure 4. ANOMQ chart for Telephone Ringing data

Example 3.4.

Genome Size: In this example, we focus on a genome data set that represents the genome size, measured in picograms of DNA per haploid cell, in several large groups of crustaceans that were reported in Gregory (2014). However, McDonald (2014) indicated that the closely related species are likely to have similar genome sizes because they recently descended from a common ancestor and therefore data from closely related species would not be independent. Hence, they randomly chose one species from each family to represent the genome sizes. The resulting sampled data are reported in Table 8. In this example, we want to use a graphical test to answer a biological question that whether some groups of crustaceans have different genome sizes than others.

Table 8. Genome sizes of Crustaceans

	Amphipods	Barnacles	Branchiopods	Copepods	Decapods		Isopods	Ostracods
Genom size	0.74	0.67	0.19	0.25	1.60	4.66	1.71	0.46
	0.95	0.90	0.21	0.25	1.65	4.70	2.35	0.70
	1.71	1.23	0.22	0.58	1.80	4.75	2.40	0.87
	1.89	1.40	0.22	0.97	1.90	4.84	3.00	1.47
	3.80	1.46	0.28	1.63	1.94	5.23	5.65	3.13
	3.97	2.60	0.30	1.77	2.28	6.20	5.70	
	7.16		0.40	2.67	2.44	8.29	6.79	
	8.48		0.47	5.45	2.66	8.53	8.60	
	13.49		0.63	6.81	2.78	10.58	8.82	
	16.09		0.87		2.80	15.56		
	27.00		2.77		2.83	22.16		
	50.91		2.91		3.01	38.00		
	64.62				4.34	38.47		
					4.50	40.89		
					4.55			
$\bar{y}_{i\bullet}$	15.447	1.377	0.789	2.264	8.757		5.002	1.326
s_i	20.406	0.671	0.979	2.352	10.102		2.745	1.075

This data fail to satisfy both the assumptions of normality and the homogeneity of variances. Therefore, we applied the natural log transformation and found that the transformed data satisfies the assumptions. Summary of the log-transformed data is given in Table 9.

Table 9. Summary of the ln(Genome size) data

	Amphipods	Barnacles	Branchiopods	Copepods	Decapods	Isopods	Ostracods
$\bar{y}_{i\bullet}$	1.878	0.229	-0.743	0.256	1.634	1.453	0.051
s_i	1.450	0.463	0.956	1.206	0.956	0.617	0.739
n_i	13	6	12	9	29	9	5

Since the species were randomly selected from each family, we consider the factor species as random and employ the random factor unbalanced design discussed in Section (2.2). Table 10 shows the ANOVA output for log transformed genome data. The results indicate that there is significant species specific variability among the crustaceans at the 5% significance level.

Table 10. ANOVA output for Genome sizes of Crustaceans

	Df	Sum Sq	Mean Sq	F value	p-value
Crustaceans	6	72.93	12.155	11.72	< 0.001
Residuals	76	78.80	1.037		

For the summary shown in Table 9, $n_{[1]} = 12$, $n_{[7]} = 13$, $MSE = 1.037$, and the mid-range estimate $\hat{\omega} = (-0.743 + 1.878)/2 = 0.568$. The critical value for the decision limits is $q(0.05, 7, 76) = 4.283$. Then, using Equations (9) and (10) the decision limits are calculated and plotted along with the treatment averages as shown in the ANOMQ chart in Figure 5.

As it is shown in Figure 5, the average genome sizes of Amphipods, Branchiopods, Decapods, and Isopods have highly deviated from the mid-range estimate indicating a significant variation in median genome size among these seven taxonomic groups of crustaceans. Also, this ANOMQ chart suggests that there are three different sets of genome sizes that may cause the significant variability. That is, a group with large genome size that includes Amphipods, Decapods, and Isopods, a group with middle genome size that includes Barnacles, Copepods, and Ostracods, and a group with small genomes that includes only Branchiopods. Apart from the evident groupings, it is clear from this analysis that some groups of crustaceans have different genome sizes than others and that causes the observed significant result. This analysis further clarifies the advantage of using graphical tests such as ANOMQ in explaining sources of variability in an experiment for the researches and scientists with minimal statistical knowledge when compared to standard ANOVA approach.

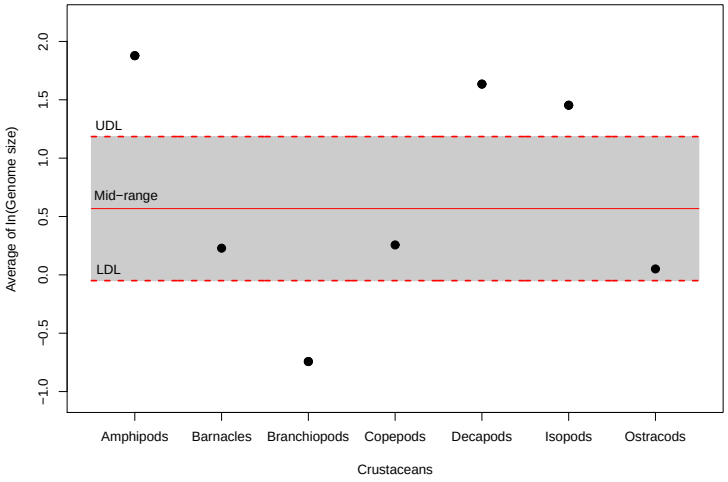

Figure 5. ANOMQ chart for Genome sizes of Crustaceans

Example 3.5.

Transconductance Tube: Gupta et al. (2020) reported results from a factorial experiment that was performed to study the effect of plate temperature and filament lighting on transconductance of a certain type of tube. That consists with two levels of plate temperature (T1 and T2) and four levels of filament lighting current L1, L2, L3, and L4; three repeats were made for each combination of plate temperature and filament current. The transconductance measurements are given in Table 11. In this example, we assume both the factors are fixed and apply both ANOVA and ANOM procedures.

Table 11. Transconductance tube data

	T1			T2			$\bar{y}_{\bullet j \bullet}$
L1	3774	4364	4374	4216	4524	4136	4231.333
L2	4710	4180	4514	3828	4170	4180	4263.667
L3	4176	4140	4398	4122	4280	4226	4223.667
L4	4540	4530	3964	4484	4332	4390	4373.333
$\bar{y}_{i \bullet \bullet}$	4305.333			4240.667			4273

We analyze this data by employing the two-factor factorial design with an interaction effect. The resulting ANOVA table is shown in Table 12 and it indicates that all three effects are insignificant at the 5% significance level.

Table 12. ANOVA output for Transconductance tube data

	Df	Sum Sq	Mean Sq	F value	p-value
Current	3	85943	28648	0.556	0.652
Temperature	1	25091	25091	0.487	0.495
Interaction	3	253668	84556	1.640	0.220
Residuals	16	825075	51567		

Then, in order to conduct the ANOM, we first calculate the interaction slope terms $x_{j(12)} = \bar{y}_{1j\bullet} - \bar{y}_{2j\bullet}$ for $j = 1, 2, 3, 4$. That yields, $x_{1(12)} = -121.333, x_{2(12)} = 408.667, x_{3(12)} = 28.667$, and $x_{4(12)} = -57.333$. These values are plotted along with the appropriate decision limits given in Equations (16) and (17) as shown in Figure 6(a). It is clear that none of the slope estimates falls outside the decision limits indicating an insignificant interaction effect.

With no significant interaction, we continue to test the main effects using the decision limits given in Equations (18) and (19) in where we plot the treatment averages shown in Table 11 along with overall average as shown in Figures 6(b) and (c). As observed in the ANOVA, the ANOM method also indicates that both the main effects are insignificant. However, the ANOM tests help visualize the underlying structure of the insignificance factors. Additionally, it helps identify near significance effects such as the interaction effect due to the factor level combination $L1(T1, T2)$ shown in Figures 6(a) that may be useful to the researches in evaluating practical significance.

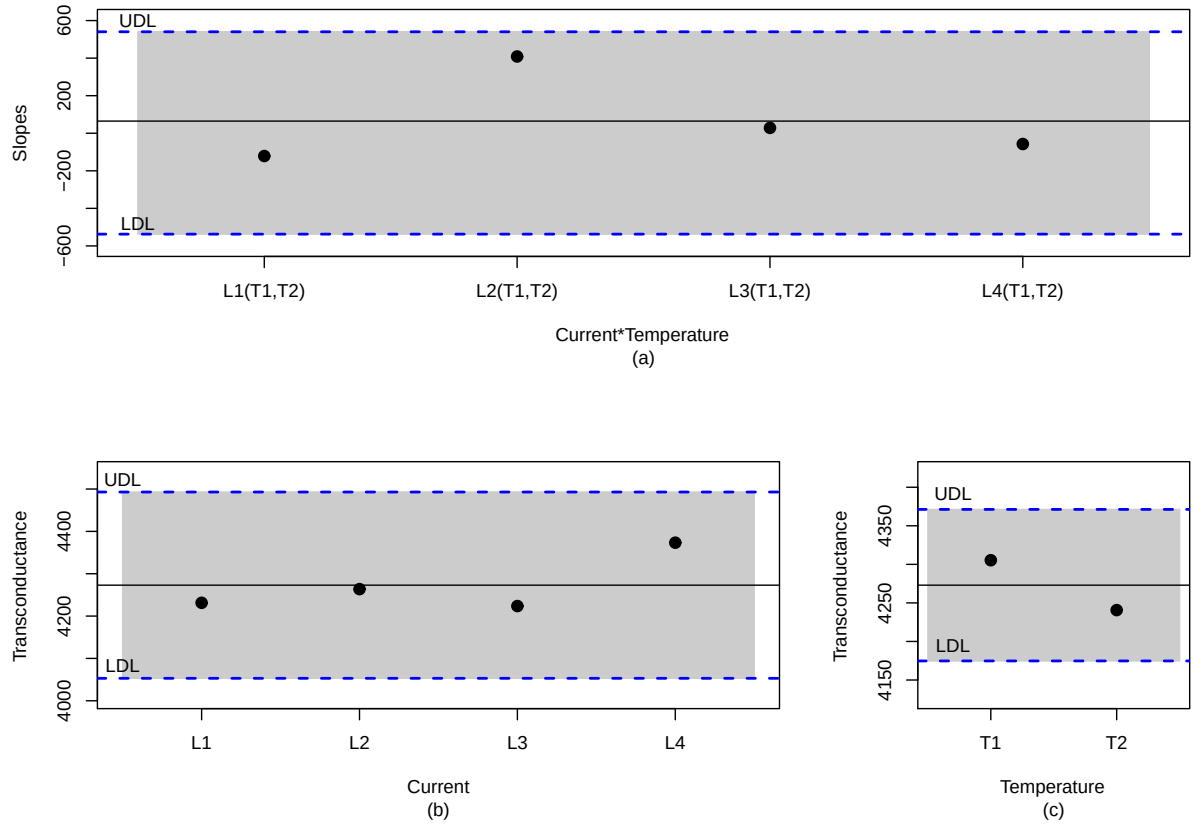

Figure 6. ANOM chart for Transconductance tube data

4. Discussion

The advantages of using analysis of means procedures in a vast array of design settings have been illustrated by the examples provided in this manuscript. The graphical representation of the tests allows for straight forward explanation of significant central tendencies and sources of variability that can especially help users with minimal statistical knowledge.

An important note on the ANOM procedure, which can also create some confusion when drawing comparisons to ANOVA, is on the issue of post hoc testing. ANOM provides a specific set of post hoc tests comparing individual effects to the overall average. These comparisons directly correspond to the overall ANOM test and are "free" in the sense that they do not create any need for additional post hoc testing adjustments. The ANOVA procedure does not have this ability. Often times, users of ANOVA conceptually confound the overall ANOVA F-test with the additional post hoc testing procedures used to provide additional insight to the rejection of the overall test.

It is redundant to compare ANOM and ANOVA procedures in terms of post hoc testing because the same testing procedures will often be used for both procedures when specific comparisons are needed. As illustrated in the balanced fixed factor analysis of Example 3.1, the individual

comparisons to the global mean produced by ANOM allows for one to investigate patterns and determine which levels of the wine quality score contributed to the significant result. However, if one was interested in direct comparison of means, say all pairwise comparison of the levels of quality score, the ANOM post hoc tests cannot answer these questions directly. A post hoc testing procedure such as Tukey's honest significant difference (HSD) would need to be applied regardless of using ANOM or ANOVA as the global test.

5. Conclusion

The purpose of this work is to provide a comprehensive study of analyzing data from single-factor and two-factor designs using graphical tests while highlighting their advantages to an ANOVA alternative. We considered both fixed and random factor analyses highlighting differences in their methodological developments and inferential procedures. The examples throughout this document provide technical insight to various real situations such as handling both balanced and unbalanced designs in the fixed or random effects settings. The ANOMR applications were also demonstrated as robust alternatives to the general ANOM procedure.

Due to its exceptional data visualization ability and ease of interpretation, the ANOM procedures should be considered a strong alternative for the ANOVA. In addition to interpretation advantages, we also believe that introduction of ANOM procedures in the early stages of statistical education may help students and researches to understand fundamental concepts in statistical testing and further enhance the interpretation of the p-value for statistical versus practical significance.

Acknowledgment:

The authors thank the anonymous referees, editors and the editor-in-chief for their constructive comments and suggestions on an earlier manuscript which led to a substantial improvement.

REFERENCES

Bakir, S. T. (1989). Analysis of means using ranks, Communications in Statistics–Simulation and Computation, Vol. 18, No. 2, pp. 757–776.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems, Vol. 47, No. 4, pp. 547–553.

Delvoye, P., Guillaume, C., Collard, S., Nardella, T., Hannecart, V. and Mauroy, M.-C. (2009). Pre-conception health promotion: analysis of means and constraints, The European Journal of Contraception & Reproductive Health Care, Vol. 14, No. 4, pp. 307–316.

Gregory, T.R. (2014). Animal genome size database, http://www.genomesize.com

Gupta, B., Guttman, I. and Jayalath, K. (2020). *Statistics and Probability with Applications for Engineers and Scientists Using MINITAB, R and JMP*, 2nd Ed., John Wiley & Sons.

Homa, K. (2007). Analysis of means used to compare providers' referral patterns, Quality Management in Healthcare, Vol. 16, No. 3, pp. 256–264.

Jayalath, K. P. and Ng, H.K.T. (2018). Analysis of means approach for random factor analysis, Journal of Applied Statistics, Vol. 45, No. 8, pp. 1426–1446.

Jayalath, K. P. and Ng, H.K.T. (2020). Analysis of means approach in advanced designs, Applied Stochastic Models in Business and Industry, Vol. 36, No. 3, pp. 501–520.

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (lmmgui), Frontiers in Psychology, Vol. 6, No. 2.

McDonald, J. H. (2014). *Handbook of Biological Statistics*, 3rd Ed., Baltimore, MD: Sparky House Publishing.

Mendeş, M. and Yiğit, S. (2013). Comparison of ANOVA-f and ANOM tests with regard to Type I error rate and test power, Journal of Statistical Computation and Simulation, Vol. 83, No. 11, pp. 2093–2104.

Mohammed, M. A. and Holder, R. (2012). Introducing analysis of means to medical statistics, BMJ Qual Saf, Vol. 21. No. 6, pp. 529–532.

Murthy, K. N., Saravana, R. and Rajendra, P. (2018). Critical comparison of north east monsoon rainfall for different regions through analysis of means technique, MAUSAM, Vol. 69, No. 3, pp. 411–418.

Nelson, L. (1983). Exact critical-values for use with the analysis of means, Journal of Quality Technology, Vol. 15, No. 1, pp. 40–44.

Nelson, P. R. (1981). Numerical evaluation of an equicorrelated multivariate non-central t distribution, Communications in Statistics–Simulation and Computation, Vol. 10, No. 1, pp. 41–50.

Nelson, P. R. (1982). Exact critical points for the analysis of means, Communications in Statistics–Theory and Methods, Vol. 11, No. 6, pp. 699–709.

Nelson, P. R. (1988). Testing for interactions using the analysis of means, Technometrics, Vol. 30, No. 1, pp. 53–61.

Nelson, P. R. (1989). Multiple comparisons of means using simultaneous confidence intervals, Journal of Quality Technology, Vol. 21, No. 4, pp. 232–241.

Nelson, P. R. (1993). Additional uses for the analysis of means and extended tables of critical values, Technometrics, Vol. 35, No. 1, pp. 61–71.

Nelson, P. R. and Dudewicz, E.J. (2002). Exact analysis of means with unequal variances, Technometrics, Vol. 44, No. 2, pp. 152–160.

Nelson, P.R., Wludyka, P.S. and Copeland, K.A.F. (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions*, Vol. 18, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Ott, E. R. (1967). Analysis of means – a graphical procedure, Industrial Quality Control, Vol. 24, No. 2, pp. 101–109.

Ott, E. R., Schilling, E.G. and Neubauer, D.V. (1975). *Process Quality Control: Troubleshooting and Interpretation of Data*, 4th ed., Milwaukee, WI: American Society for Quality, Quality Press.

Pallmann, P. and Hothorn, L. A. (2016). Analysis of means (ANOM): A generalized approach using R, Journal of Applied Statistics, Vol. 43, pp. 1541–1560.

Prokeš, L., Hegrová, J. and Kanickỳ, V. (2017). Analysis of means (ANOM) as a tool for comparison of sample treatment methods: Testing various mineralization procedures for selenium determination in biological materials, Journal of AOAC International, Vol. 100, No. 1, pp. 236–240.

Ramig, P. F. (1983). Applications of the analysis of means, Journal of Quality Technology, Vol. 15, No. 1, pp. 19–25.

Schilling, E. G. (1973). A systematic approach to the analysis of means – Part I. Analysis of treatment effects, Journal of Quality Technology, Vol. 5, No. 3, pp. 93–108.

Stoline, M.R. and Ury, H.K. (1979). Tables of the studentized maximum modulus distribution and an application to multiple comparisons among means, Technometrics, Vol. 21, No. 1, pp. 87–93.

Ury, H.K., Stoline, M.R. and Mitchell, B.T. (1980). Further tables of the studentized maximum modulus distribution: Further tables of the studentized, Communications in Statistics-Simulation and Computation, Vol. 9, No. 2, pp. 167–178.

Wludyka, P. S. and Nelson, P. R. (1997). Analysis of means type tests for variances using subsampling and jackknifing, American Journal of Mathematical and Management Sciences, Vol. 17, No. 1-2, pp. 31–60.