Penultimate version. If citing, please refer instead to the published version in Journal of Applied Statistics, DOI: 10.1080/02664763.2015.1117584.

Analysis of means (ANOM): A generalized approach using R

Philip Pallmann and Ludwig A. Hothorn

Abstract

Papers on the analysis of means (ANOM) have been circulating in the quality control literature for decades, routinely describing it as a statistical stand-alone concept. Therefore we clarify that ANOM should rather be regarded as a special case of a much more universal approach known as multiple contrast tests (MCTs). Perceiving ANOM as a grand-mean-type MCT paves the way for implementing it in the open-source software R. We give a brief tutorial on how to exploit R's versatility and introduce R package `ANOM` for drawing the familiar decision charts. Beyond that, we illustrate two practical aspects of data analysis with ANOM: firstly, we compare merits and drawbacks of ANOM-type MCTs and ANOVA F-test and assess their respective statistical powers, and secondly, we show that the benefit of using critical values from multivariate t-distributions for ANOM instead of simple Bonferroni quantiles is oftentimes negligible.

ANOVA F-test, multiple contrast test, multivariate t-distribution, control chart, industrial quality assessment

1 Introduction

The analysis of means (ANOM) is a common statistical procedure in quality assurance for comparing several treatment means against an overall mean (grand mean) in a variety of experimental design and observational study situations. It is basically a graphical method, yielding control charts that allow to draw conclusions and interpret results easily with respect to both statistical and practical significance.

ANOM has been in use for more than 40 years now. Resting upon basic ideas outlined by Halperin and colleagues Halperin et al. (1955), Ott published his pioneering ANOM paper (Ott, 1967, reprinted in Ott (1983)) in which he coined the phrase 'analysis of means'. Various extensions have been proposed since then, and ANOM has been adopted to match with a variety of experimental designs, randomization structures, and data types (e.g., Schilling, 1973a,b; Enrick, 1976; Subramani, 1992). Well-written and concise overviews of ANOM methodology are available e.g., Ramig's synopsis of applications Ramig (1983), Rao's review article Rao (2005), and the textbook by Nelson Nelson et al. (2005). Applications of ANOM in health care Homa (2007) and medicine Mohammed and Holder (2012) were describe recently.

Researchers and users in quality control often treat ANOM as if it were a stand-alone method. In fact, ANOM can be considered as a special case of a much broader statistical concept known as multiple contrasts tests (MCTs) Mukerjee et al. (1987); Bretz et al. (2001). The family of MCTs unifies a number of well-established multiple comparison procedures such as Tukey's all-pairwise comparisons Tukey (1994), Dunnett's comparison of several means against a control Dunnett (1955), and Williams' test on trend Williams (1971, 1972); Bretz (2006), among others. In a similar fashion, we propose a generalized approach for ANOM using the concept of MCTs, specifically comparisons to the grand mean. MCTs usually involve few groups with relatively large sample sizes whereas in ANOM it is not unusual to compare a relatively large number of groups each of which has only small sample size, but these are just marginal differences. In principle, ANOM and grand-mean-type MCTs are equivalent (except that MCT results are usually not presented as control charts).

Commercial software for ANOM has been available for quite some time e.g., the homonymous SAS procedure SAS Institute Inc. (2012). Perceiving ANOM as grand-mean-type MCT enables us to carry out data analyses using the freely available software R R Core Team (2013), which offers a flexible framework that can handle most diverse data scenarios.

This article pursues two basic goals: first, we present a practitioner-friendly comprehensive treatment of ANOM in the context of MCTs, and second, we explain how to do ANOM in R, using thoroughly worked examples of real-world datasets.

Beyond that, we explore two more aspects of practical relevance: we assess the *de facto* benefit of complex methods for calculating critical values compared to a simple Bonferroni correction, and we investigate applicability and statistical power of ANOM-type MCTs and ANOVA F-test. Many practitioners habitually perform their statistical analysis with ANOVA, even in situations when this is suboptimal. The one-way

ANOVA gives a global answer to the question whether any of the group means is different from any other group mean. However, in most applications, the statement that 'at least one of the group means is different from at least any other group mean' is of little practical relevance. People rather want to know which groups differ e.g., which of the group means are significantly higher or lower than the grand mean. ANOM provides this kind of information on local hypotheses and additionally a global test decision together with a control chart.

The remainder of this article is structured as follows. In Section 2 we reformulate ANOM as a multiple contrast test and summarize several extensions. Section 3 compares performance and power of ANOM and ANOVA for situations where both methods might be adequate. The benefit of computing quantiles from multivariate t-distributions over a Bonferroni adjustment is quantified in Section 4. Section 5 is dedicated to an illustration of R's vast functionality for ANOM with a strong focus on worked examples. A brief conclusion is given in Section 6.

2 ANOM as a multiple contrast test

2.1 General methodology

Assume a normally distributed random variable $Y_{ij} = \mu_i + \epsilon_{ij}$ in a randomized one-way layout where μ_i is the mean of the ith group, $i = 1, \ldots, k$, containing individuals $j = 1, \ldots, n_i$, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The column vector of group means is $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$, the grand mean of all groups is denoted by $\bar{\mu}$, and $\boldsymbol{\theta} = \boldsymbol{\mu}/\sigma$ is the vector of scaled expectations. The null hypothesis of all group means being equal to the grand mean is to be tested against the alternative that at least one group mean differs from the grand mean:

$$\begin{aligned} H_0 &: \ \mu_i = \bar{\mu} \ \forall \ i \in \{1, \ldots, k\} \\ H_A &: \ \exists \ i : \ \mu_i \neq \bar{\mu}, \ i \in \{1, \ldots, k\} \end{aligned}$$

The goal is to control the familywise error rate (FWER) i.e., the probability of one or more false-positive findings among all k hypotheses, at a pre-defined level α, traditionally 5%.

The hypotheses of interest can be formulated using contrasts. A contrast is a linear combination of group means with coefficients $\mathbf{c} = (c_1, \ldots, c_k)$, and we can write it up as

$$c_1\mu_1 + \cdots + c_k\mu_k = \mathbf{c}\boldsymbol{\mu}$$

with $\sum_{i=1}^{k} c_i = 0$. Several contrast vectors $\mathbf{c}_l = (c_{l1}, \ldots, c_{lk}), \ l = 1, \ldots, q$ can be combined in a $q \times k$ contrast matrix $\mathbf{C} = (\mathbf{c}_1', \ldots, \mathbf{c}_q')'$ that defines the entity of comparisons in an MCT procedure. Rewriting ANOM as an MCT requires a grand mean contrast matrix $\mathbf{C}_{GM}$ of dimension $q \times k$:

$$\mathbf{C}_{GM} = \begin{pmatrix} 1 - \frac{n_1}{N} & -\frac{n_2}{N} & \ldots & -\frac{n_k}{N} \\ -\frac{n_1}{N} & 1 - \frac{n_2}{N} & \ldots & -\frac{n_k}{N} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{n_1}{N} & -\frac{n_2}{N} & \ldots & 1 - \frac{n_k}{N} \end{pmatrix}$$

where $N = \sum_{i=1}^{k} n_i$ is the total sample size. In fact, the dimension of $\mathbf{C}_{GM}$ is $k \times k$ because the number of contrasts, q, equals the number of groups, k; this is not the case for MCTs in general.

A contrast test statistic is a standardized contrast:

$$T_l = \frac{\sum_{i=1}^{k} c_{li}\mu_i}{s\sqrt{\sum_{i=1}^{k} c_{li}^2/n_i}}$$

with s the square root of the common variance estimate. T_l follows a central t-distribution with $\nu = n_i - 1$ degrees of freedom. Correspondingly, the vector of all q test statistics $\mathbf{T} = (T_1, \ldots, T_q)$ is jointly q-variate t-distributed Bretz et al. (2001) with $\nu = N - k$ degrees of freedom and a common correlation matrix $\mathbf{R}$ whose elements under H_0 are

$$\rho_{ll'} = Corr(T_l, T_{l'}) = \frac{\sum_{i=1}^{k} c_{li}c_{l'i}/n_i}{\sqrt{(\sum_{i=1}^{k} c_{li}^2/n_i)(\sum_{i=1}^{k} c_{l'i}^2/n_i)}}$$

where $1 \leq l \neq l' \leq k$.

In the case of two-sided hypotheses, the ith local null hypothesis $H_0^{(i)} : \mu_i = \bar{\mu}$ is rejected in favor of $H_A^{(i)} : \mu_i \neq \bar{\mu}$ if

$$|T_l| \geq t_{two,q,1-\alpha,\nu,\mathbf{R}}$$

where $t_{two,q,1-\alpha,\nu,\mathbf{R}}$ is the two-sided $1-\alpha$ equicoordinate quantile from the central q-variate t-distribution with ν degrees of freedom and correlation matrix $\mathbf{R}$. Similarly, the global null hypothesis is rejected if

$$\max\{|T_1|, \ldots, |T_q|\} \geq t_{two,q,1-\alpha,\nu,\mathbf{R}}.$$

In opposition to the the ANOVA F-test, MCTs (and thus ANOM) are not confined to two-sided inference. If a practical question suggests one-sided hypotheses, we may easily assess them using ANOM. We reject $H_0^{(i)} : \mu_i \leq \bar{\mu}$ in favor of $H_A^{(i)} : \mu_i > \bar{\mu}$ if $T_l > t_{one,q,1-\alpha,\nu,\mathbf{R}}$. Correspondingly, we reject $H_0^{(i)} : \mu_i \geq \bar{\mu}$ and accept $H_A^{(i)} : \mu_i < \bar{\mu}$ if $T_l < t_{one,q,\alpha,\nu,\mathbf{R}}$.

Computation of adjusted p-values and simultaneous confidence intervals (SCIs) is detailed e.g., in Bretz et al. (2001) and Hothorn et al. (2008). In the context of ANOM, we are particularly interested in decision limits DL_i which represent the minimum significant effect. For two-sided alternative hypotheses in a one-way layout, we obtain

$$DL_i = \bar{y} \pm s \; t_{two,q,1-\alpha,\nu,\mathbf{R}} \sqrt{\frac{N-n_i}{Nn_i}}$$

as upper and lower boundary, respectively, where $\bar{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$. In the case of one-sided alternative hypotheses, we use

$$UDL_i = \bar{y} + s \; t_{one,q,1-\alpha,\nu,\mathbf{R}} \sqrt{\frac{N-n_i}{Nn_i}}$$

as an upper limit or

$$\begin{aligned} LDL_i &= \bar{y} - s \; t_{one,q,1-\alpha,\nu,\mathbf{R}} \sqrt{\frac{N-n_i}{Nn_i}} \\ &= \bar{y} + s \; t_{one,q,\alpha,\nu,\mathbf{R}} \sqrt{\frac{N-n_i}{Nn_i}} \end{aligned}$$

as a lower limit. If the one-way design is balanced, the term $\sqrt{\frac{N-n_i}{Nn_i}}$ simplifies to $\sqrt{\frac{k-1}{N}}$. Extensions for proportions and rates (using a normal approximation), two-way and higher-order layouts, will be the topic of the next section.

Notice that, as long as they refer to comparisons with the grand mean, SCIs can be converted into ANOM decision limit via

$$\begin{aligned} LDL_i &= \bar{\mu} - |\mu_i - \bar{\mu} - upr_i| \\ UDL_i &= \bar{\mu} + |\mu_i - \bar{\mu} - lwr_i| \end{aligned}$$

where lwr_i and upr_i denote lower and upper SCI bounds whose computation is explained in Hothorn et al. (2008). This notion eludicates that ANOM decision limits are basically SCIs shifted by the group effects (i.e., the differences between group means and the grand mean).

2.2 Extensions

R's basic ANOM functionality is illustrated with a data example of water filters in Section 5.1. Beyond that, ANOM can be extended to a multitude of specialized applications and non-trivial data scenarios. We discuss a number of common challenges in the following.

2.2.1 ANOM under variance heterogeneity

When variances cannot be assumed equal across groups, a heteroscedastic version of ANOM is expedient. The so-called HANOM was introduced in the past decade Nelson and Dudewicz (2002); Dudewicz and Nelson (2003). In the context of MCTs, two distinct approaches dealing with heterogeneous variances are available.

One approximate solution is the heteroscedastic generalization of MCTs described in Hasler and Hothorn (2008); it involves computation of group-specific variance estimates s_i^2 which are plugged into the correlation matrix. In conjunction with contrast-specific Satterthwaite approximations to the degrees of freedom ν_i, we end up with separate critical values (i.e., quantiles from a multivariate t-distribution) for each group comparison to the grand mean.

A different approach was proposed by Herberich et al. (2010); it is based on so-called sandwich variance estimators Huber (1967), primarily the heteroscedasticity-consistent HC3 estimator

$$\tilde{s}_i^2 = \frac{n_i}{n_i - 1} s_i^2$$

that was developed in MacKinnon and White (1985) and recommended by Long and Ervin (2000). The estimates $\tilde{s}_i^2$ are plugged into the correlation matrix but 'naive' degrees of freedom $\nu = N - k$ are used so that each test statistic is compared to the same critical value.

Simulations have shown that the former approach is more robust in the presence of small sample sizes Hasler (2014). How to implement ANOM with heteroscedastic data in R is illustrated in Section 5.2.

2.2.2 ANOM for the ratio to the grand mean

Instead of assessing each group's *difference* to the grand mean, one can also make inference for relative changes. This leads to convenient interpretations in terms of percentages, which may be more appropriate in some applications. For multiple endpoints to be monitored simultaneously (e.g., the single strand break factor and weight of a textile fiber in the bivariate quality control problem of Yeh et al. (2004)), ratios to the grand mean are more intuitive than differences, especially when the endpoints are measured on different scales.

ANOM for ratios of parameters was described by Djira and Hothorn (2009) already. A heteroscedastic version of this method based on the approach of Hasler and Hothorn (2008) can be found straightforwardly and is exemplified in Section 5.3.

2.2.3 ANOM for counts and proportions

A common way of analyzing count data is by fitting a Poisson generalized linear model (GLM) involving a logarithmic link function. Subsequent multiple comparisons with the grand mean apply to the parameters estimated in the model.

For binomial data (proportions), we draw a distinction regarding the structure of data:

i) If only one proportion $\pi_i = \frac{x_i}{n_i}$ of events x_i (successes) over samples n_i (trials) is available for each of the k groups, the data are usually summarized in a $2 \times k$ table. Simultaneous inference for differences of binomial proportions was discussed in Schaarschmidt et al. (2008). We show a practical example in Section 5.5.

ii) If proportions are available for several independent units per group, then the variance of proportions observed in the same group i can be estimated, and hence we can fit a binomial GLM involving a logit link, followed by comparisons of model parameter estimates to their grand mean Hothorn et al. (2008).

2.2.4 Nonparametric ANOM

ANOM-type data analysis may also be carried out nonparametrically e.g., with ranks as inputs (ANOMR) Bakir (1989, 1994) or as a randomization test based on permutations (PANOM). Nonparametric MCTs are discussed by Konietschke et al. (2012); they express the hypotheses in terms of relative effects, which in turn are estimated based on global rankings. The estimators are unweighted, meaning they are independent of the groups' sample sizes ('pseudo-ranks').

The relative effect of two independent random variables X_1 and X_2 following some distributions F_1 and F_2 is generically defined as

$$p = P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2).$$

Therefore p is the probability that X_1 takes smaller values than X_2 (plus half the probability of taking equal values). Hence when $p < \frac{1}{2}$, X_1 is stochastically more likely to take larger values than X_2, and *vice versa* for $p > \frac{1}{2}$.

The relative effect of the ith group (in a one-way layout) is estimated as

$$\hat{p}_i = \frac{1}{N}(\bar{R}_i - \frac{1}{2})$$

where $\bar{R}_i$ is the mean of ranks belonging to group i. Besides the multivariate t-approximation to the distribution of test statistics, a range-preserving Fisher transform may be employed to ensure the decision limits lie within $[-1, 1]$. The nonparametric MCT-based ANOM procedure does not act on the assumption of homogeneous variances; hence no separate extension to heteroscedastic cases is required. Since no continuous probability distribution is assumed whatsoever, tied values or ordered categorical scores can be analyzed appropriately.

2.2.5 ANOM with replicated designs using linear mixed-effects models

There are various types of trials involving repeated measurements; two designs widespread in quality control are simple technical replicates (e.g., triplicates) and block designs. Replicates occur when each test object is measured a number of times in order to account for measurement error. Measurements obtained from the same object are usually correlated. An example of a block design is when there are several test persons (the 'blocks'), and each person assesses each test object once. This is a strategy to remedy subjectivity when quality is judged by means of individual ratings. Measurements by the same person are correlated, as are measurements on the same object. ANOM-type data analysis using a mixed-effects model in R is illustrated by a dataset of ergonomic stools in Section 5.4.

2.2.6 ANOM for variances

Suggestions for comparing variances with ANOM-like procedures have been made in abundance Wludyka and Nelson (1997a,b); Rao and Krishna (1997); Kumar and Rao (1998); Wludyka and Nelson (1999); Bernard and Wludyka (2001); Wludyka and Sa (2004). ANOM for variances can also be conducted within an MCT framework by employing robust Levene residuals i.e., absolute deviations from the median Brown and Forsythe (1974)

$$z_{ij} = |y_{ij} - \tilde{y}_i|$$

where $\tilde{y}_i$ denotes the ith group median Pallmann et al. (2014). We evaluate quality control data of springs using ANOM for variances in Section 5.6.

2.2.7 Further generalizations

MCTs in general parametric models were treated in Hothorn et al. (2008), extending the accessibility of multiple comparison procedures to a wide range of (semi-)parametric models such as regression and AN(C)OVA models, GLMs, mixed-effects models, and censored event-time models (e.g., the Cox proportional hazard model for survival times). This framework also covers models with covariates as well as two-way and higher-order layouts. There are only two basic requirements for MCTs to be feasible: the parameter estimates of interest must be (at least asymptotically) multivariate normal, and the corresponding covariance estimates must be consistent. That being given, we can easily compute ANOM decision limits for a variety of continuous or discrete endpoints, single- or multi-factor designs, balanced or unbalanced datasets, equal or unequal variances, differences or ratios of parameters, and variances.

3 ANOM vs. ANOVA F-test

3.1 General characteristics

Practitioners sometimes use ANOM and one-way ANOVA as if they were exchangeable. In fact, both of them can be applied to similar testing problems, and therefore we may consider them as competitors, which makes it reasonable to compare e.g., their power behavior under certain configurations. However, it is important to notice that ANOM and ANOVA provide substantially different information:

1. Although the null hypothesis

   $$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k$$

   is the same for both procedures, their alternatives differ:

   $$H_A^{ANOM} : \ \exists\, i : \ \mu_i \neq \bar{\mu}, \ 1 \leq i \leq k$$
   $$H_A^{ANOVA} : \ \exists\, (i, i') : \ \mu_i \neq \mu_{i'}, \ 1 \leq i \neq i' \leq k$$

   Or, in words: H_A^{ANOM} states that at least one group mean differs from the grand mean whereas H_A^{ANOVA} implies that there is at least one pair of groups whose means differ.

2. The ANOVA F-test is a purely global test (stating whether any two group means differ) whereas ANOM provides global information and additionally allows for local inference (itemizing which specific group means differ from the grand mean).

3. The F-test has a quadratic test statistic with numerator $\sum(y_i - \bar{y})^2$ whereas ANOM has a linear test statistic with numerator $\max(y_i - \bar{y})$.

4. The F-test is restricted to differences whereas ANOM can be defined either for differences or ratios to the grand mean.

5. ANOVA provides neither SCIs nor decision limits for quality control charts whereas ANOM yields adjusted p-values, SCIs, and decision limits.
6. Only two-sided alternatives can be formulated for the F-test whereas ANOM is suited for two- as well as directional one-sided alternatives.
7. The power is different for the same patterns of group means μ_i, group sample sizes n_i, and numbers of groups k, as will be explored in the following.

Simultaneous inference for *all possible* contrasts (i.e., the ANOVA set of hypotheses) is feasible by the method of Scheffé Scheffé (1959); however, this leads to extremely conservative decisions and is unrewarding for our problem.

3.2 Power

The power of ANOM has been studied Nelson (1985); Wludyka et al. (2001) and compared to the power of the ANOVA F-test Nelson (1983b); He et al. (2001); Chang et al. (2010); Mendeş and Yiğit (2013). It must be clearly emphasized, however, that the latter is a comparison of apples and oranges; the crucial point here is indeed their difference in alternative hypotheses. ANOVA is known to be the uniformly most powerful unbiased invariant test for the problem treated here, meaning that it has maximum power among all unbiased and invariant tests *for the entity of possible alternatives*. However, if the set of alternatives is narrowed down (e.g., by defining grand-mean contrasts), an MCT-type procedure like ANOM is more powerful than ANOVA for *some* configurations of the alternative (precisely those suiting the alternative space defined by the contrasts) but less powerful for all other configurations.

Our interest lies in the question: what is the price to pay for the benefits gained by the ANOM (i.e., local test decisions, SCIs, decision limits) in comparison to the ANOVA F-test? We want to elucidate this for some practically relevant cases involving small to moderate sample sizes. To this end we investigate the respective statistical powers of ANOM and ANOVA for different configurations of the alternative. To ensure a somewhat 'fair' comparison, we follow the approach of Hayter and Liu (1990) and Konietschke et al. (2013), who identified the so-called least favorable configurations (LFCs).

Recall that $\boldsymbol{\theta} = (\frac{\mu_1}{\sigma}, \ldots, \frac{\mu_k}{\sigma})$ is the vector of scaled expectations and $\bar{\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta_i$ the average of its elements. To keep things simple here, we limit our considerations to balanced ($n_1 = \cdots = n_k = n$) and homoscedastic ($\sigma_1 = \cdots = \sigma_k = \sigma$) one-way layouts. Our goal is to find the configuration (or 'shape') of the θ_i's for which the power functions of ANOM and ANOVA are minimized. This is the definition of the LFC, and we purpose to detect it separately for each of the two conditions

$$b_1(\boldsymbol{\theta}) = \max_{1\leq i\leq k}|\theta_i - \bar{\theta}| \geq b$$

and

$$b_2(\boldsymbol{\theta}) = \max_{1\leq i,j\leq k}|\theta_i - \theta_j| \geq b.$$

Under the first condition, we have $b_1(\boldsymbol{\theta}^*) = b \geq 0$, and the power function is minimized for

$$\boldsymbol{\theta}^* = (0, \ldots, 0, \frac{bk}{k-1}).$$

Similarly, under the second condition, we have $b_2(\boldsymbol{\theta}^*) = b \geq 0$, and the power function is minimized for

$$\boldsymbol{\theta}^* = (-\frac{b}{2}, 0, \ldots, 0, \frac{b}{2}),$$

adopting the notation of Konietschke et al. (2013). Happily, it turns out that these are the LFCs of both ANOM and ANOVA Hayter and Liu (1990).

The power of the ANOM-type MCTs can be calculated using the R package `MCPAN` Schaarschmidt et al. (2013). We are particularly interested in the any-pairs power i.e., the probability of rejecting at least one elementary null hypothesis for which the corresponding comparison is *a priori* known to be under $H_A^{(i)}$. We define the subset of (two-sided) grand mean comparisons which are truly under the alternative hypothesis as $\mathcal{S} = \{i : H_A^{(i)} : \mu_i \neq \bar{\mu}\}$. Then the any-pairs power is given by

$$P(\exists\, i : |T_i| \geq t_{k,1-\alpha,\nu,\mathbf{R}} | H_A^{(i)}) = 1 - P(-t_{k,1-\alpha,\nu,\mathbf{R}} < T_i < t_{k,1-\alpha,\nu,\mathbf{R}}, \forall\, i \in \mathcal{S}).$$

We assess the powers of ANOM and ANOVA under both LFCs for $0 \leq b \leq 3$ and $k = 10$ groups with sample sizes $n = \{3, 5, 10\}$ per group. For convenience, we fix $\sigma = 1$ so that $\boldsymbol{\theta} = \boldsymbol{\mu}$. Besides the two alternatives detected with minimum power, we include several other configurations of the alternative in our power study. The resulting power curves are displayed in Figure 1.

ANOM has higher power than ANOVA for the LFC under b_1, and their powers are about identical for the LFC under b_2; this is in line with the findings of Konietschke et al. (2013). Beyond that, we want to get a general idea of the behavior under non-LFC conditions. So if the θ_i's are split 'fifty-fifty' (as in Figure 1, alternatives 1 and 2), ANOM's power is clearly inferior. On the contrary, ANOM outperforms the F-test if one group is distinctly different from the remaining θ_i's (see alternatives 3 and 4 in Figure 1), as ANOM is particularly sensitive to this configuration of the alternative (by its very own definition). For various other shapes of the θ_i's, the power discrepancies between ANOM and ANOVA are often only marginal (e.g., alternatives 5 through 8 shown in Figure 1).

This illustrates the knottiness of the matter: the powers of ANOM and ANOVA are governed by a non-trivial interplay of sample sizes n_i and the configuration of the θ_i's. Moreover, varying the number of groups k and allowing for unbalanced designs would further increase complexity. And this is still not the full story. Once again: any direct comparison of the respective powers of ANOM and ANOVA is *per se* intricate as they test the same H_0 against different alternative hypotheses. This being the case, we urge the reader to interpret our results with caution.

As a consequence, we will not give any universal—and probably misleading—recommendation on whether to use ANOM or ANOVA. What we want to point out is that one should not obviate ANOM simply for fear of losing power. Quite the contrary, one should not hesitate to use ANOM instead of ANOVA if the former provides a more accurate, better interpretable answer to the question of subject-matter interest. The decision whether to analyze one's data with ANOM or ANOVA (or any other method) should always be guided by practical considerations. If scientific interest is targeted on whether a factor influences the outcome, the F-test is an apt choice. By contrast, if the matter is to unravel which specific levels of a factor differ from the grand mean of all factor levels, ANOM is perfectly suitable, and there is no general need to perform an ANOVA prior to ANOM. Last but not least, the unfavorable 'fifty-fifty' configurations of the alternative are rather unlikely to occur in quality control; hence the use of ANOM can be supported also from a power perspective.

4 Multivariate t vs. Bonferroni quantiles

There are various approaches for multiplicity adjustment with ANOM-type comparisons of group means versus a grand mean. The simplest choice is a Bonferroni correction i.e., the confidence level α is divided by the number of groups k. On the contrary, comparing test statistics to critical points from a multivariate t-distribution is a rather sophisticated method. Great effort has been devoted to obtaining critical points for ANOM (e.g., Nelson, 1982, 1983a, 1993; Guirguis and Tobias, 2004), and they are frequently preferred to Bonferroni's correction on the grounds that they account for the correlation among comparisons, which makes the procedure less conservative.

This justification for applying involved methods, however, is often rather weak in practice. We present power comparisons of ANOM for two-sided hypotheses using quantiles from a multivariate t-distribution versus multiple comparisons to the grand mean with a Bonferroni adjustment in Figure 2. The profit of incorporating the correlation among test statistics is sizeable only with very few groups ($k < 5$) and tiny samples (e.g., $n_i = 3$). With $b = 2$, $n_i \geq 5$ and/or $k \geq 5$, the gain in power is at most 3–5%, irrespective of the configuration of the θ_i's. All power ratios approach unity with increasing b.

To understand why the impact of this rather complicated analysis is so small, it is helpful to have a look at the correlation among comparisons, which is a function of the number of groups. The pairwise correlations of contrasts are $-\frac{n_i}{N}$, so when there are 10 groups to be compared to their grand mean in a balanced setting, the absolute value of the correlation among individual tests is no more than $\frac{1}{k} = 0.1$.

To summarize, the benefit compared to a simple Bonferroni adjustment is considerable only with very few groups and very small samples but quickly vanishes as k and n_i increase. On top of that, the computational burden for obtaining quantiles from a multivariate t-distribution can be sizeable when the number of groups is large. Thus, we must recognize that additional time and effort are inversely linked to the profit earned, so we may feel free to apply Bonferroni in most practical scenarios without losing any notable amount of power. This might be different in settings where the group means are correlated e.g., in the presence of blocks or repeated measurements.

5 Using R for ANOM

The open-source software R is a popular and well-established tool in many areas of statistics and applied sciences. In quality control, however, R has been neglected for some reason or other. This is quite contrary to its vast amount of functions exceeding that in many commercial software products. The objective of this section is to illustrate R's versatility for ANOM-type data analysis by means of worked examples.

The R package `multcomp` Hothorn et al. (2008) provides a universal implementation of MCTs which are, as we have seen, just a generalization of ANOM. Several other packages contain functions for more specialized applications. Computation of quantiles from multivariate t-distributions is generally based on the Genz-Bretz algorithm Genz and Bretz (2009) and implemented in the package `mvtnorm` Genz et al. (2013). An elaborate description of multiple comparisons using R is given in Bretz et al. (2010).

What has been missing in R so far is a function plotting ANOM results in a decision chart. We provide such a function in our package `ANOM` Pallmann (2015), which is downloadable for free by executing this line in R:

```
install.packages("ANOM")
```

In the following hands-on guide to the usage of the package `ANOM` we take the reader's basic familiarity with R for granted.

5.1 Standard ANOM

Hsu (1984) presents data from a comparison of seven brands of water filters. Water samples were run through the filters and then they were incubated; the number of bacterial colonies growing on each filter was taken as a measure of its efficiency (good filters are littered with many colonies). The dataset is unbalanced as only two devices of brands 4 and 7 were tested but three of all other brands. The complete data is stored in the R package `ANOM`; we can access it via:

```
library(ANOM)
waterfilter
```

Assuming a normal distribution and homogeneous variances across filter brands, we can apply a grand mean MCT to assess whether any of the brands filter fewer bacteria than average at a multiple type I error level of 5%. The workflow in R is straightforward: after loading the package `multcomp`, we fit a linear model to the data and apply a generalized linear hypothesis test with grand mean contrasts:

```
library(multcomp)
wfmodel <- lm(colonies ~ brand, data=waterfilter)
wf <- glht(wfmodel, mcp(brand="GrandMean"), alternative="less")
```

We set the option `alternative="less"` because we only want to detect filter brands that are inferior to the grand mean, and hence testing one-sided hypotheses is appropriate. Adjusted p-values and corresponding SCIs are computed with the commands `summary(wf)` and `confint(wf)`. A concise presentation of the results is provided in a decision chart (Figure 3) obtained by

```
ANOM(wf)
```

We find that devices of brand 1 filter significantly fewer bacteria from water samples than average, which is underpinned by the small adjusted p-value of 0.002. The ANOVA F-test yields even smaller a p-value of 0.0006 but does not disclose which of the brands are worse than average.

This simple analysis is based on the doubtful assumption of equal group variances. Especially brands 4, 6 and 7 have distinctly smaller variances than the other groups, suggesting a modified ANOM procedure that can cope with heteroscedasticity. If fact, heterogeneous variances are rather the rule with real-world data than the exception.

5.2 ANOM with heterogeneous variances

One approach to take unequal variances into account is to replace the covariance matrix from the linear model with a heteroscedasticity-consistent (HC) sandwich estimate as proposed in Herberich et al. (2010). We redo our analysis with the help of the `sandwich` package Zeileis (2004) and get a decision chart with altered decision limits:

```
library(sandwich)
wf1 <- glht(wfmodel, mcp(brand="GrandMean"), alternative="less",
            vcov=vcovHC)
ANOM(wf1)
```

The modified covariance matrix changes the results: now we find that both filter brands 1 and 4 host significantly fewer bacteria than the grand mean of all filters (Figure 4). Brand 4 filters have 'gained' significance at the multiple 5% level because their mean is estimated with very small error.

A different way of acknowledging heteroscedasticity is to compute separate degrees of freedom and critical values for each of the contrasts Hasler and Hothorn (2008). This is implemented in the R package `SimComp` Hasler (2012); setting the option `covar.equal=F` invokes the desired heteroscedasticity adjustment:

```
library(SimComp)
wf2 <- SimCiDiff(data=waterfilter, grp="brand", resp="colonies",
                 type="GrandMean", alternative="less", covar.equal=F)
wf2p <- SimTestDiff(data=waterfilter, grp="brand", resp="colonies",
                    type="GrandMean", alternative="less", covar.equal=F)
ANOM(wf2, stdep=waterfilter$colonies, stind=waterfilter$brand, pst=wf2p)
```

The decision chart of the water filter data using this method (Figure 5) looks somewhat different from Figure 4, and the difference of filter brand 1 to the grand mean is no longer significant due to the large variance in this sample. Note that estimating group-specific variances may be problematic with sample sizes as small as in our example.

5.3 ANOM for ratios to the grand mean

Yet another option when analyzing the water filter data is to assess ratios to the grand mean instead of differences. This has the advantage that the effects can be nicely interpreted as the group means' percent deviations from the grand mean. The ratio functions from `SimComp` will compute SCIs and p-values, and again we use the heteroscedastic versions:

```
wfr <- SimCiRat(data=waterfilter, grp="brand", resp="colonies",
                type="GrandMean", alternative="less", covar.equal=F)
wfrp <- SimTestRat(data=waterfilter, grp="brand", resp="colonies",
                   type="GrandMean", alternative="less", covar.equal=F)
ANOM(wfr, stdep=waterfilter$colonies, stind=waterfilter$brand, pst=wfrp)
```

We get Figure 6, which looks identical to Figure 5 except the scale of the y-axis has been changed to percentages.

5.4 ANOM for clustered data

Four ergonomically designed stools were tested by nine different people who rated the effort to raise from these stools on the so-called Borg scale (it measures perceived exertion on a scale of 6–20) Wretenberg et al. (1993). This is a 'block' design with nine clusters of observations (the raters), and the experimental structure should be correctly reflected in the statistical analysis. Hence the method of choice is a linear mixed-effects model that can be fitted in R using the package `nlme` Pinheiro et al. (2013), which also contains the dataset. The stool types are modeled as fixed effects and the test persons are considered random:

```
library(nlme)
esmodel <- lme(effort ~ Type, random=~1|Subject, data=ergoStool)
```

Now we can use our mixed-effects model fit to perform MCTs with a grand mean contrast matrix and build the ANOM decision chart. The arguments `xlabel` and `ylabel` help us modify the axis labels to our taste:

```
es <- glht(esmodel, mcp(Type="GrandMean"), alternative="two.sided")
ANOM(es, xlabel="Stool Type", ylabel="Exertion (Borg Scale)")
```

We see that the effort to raise from stools of types 1 and 4 is significantly lower than average whereas stool type 2 causes exertion above the grand mean (Figure 7).

5.5 ANOM with a binomial endpoint

Nelson et al. (2005) present binomial data from a comparison of math achievements among ten elementary schools in a U.S. district: six conventional neighborhood schools (N1 through N6) and four alternative schools (A1 through A4). The performances of 563 fifth graders were assessed using a standardized test score, and the outcome was defined as the proportion of children scoring proficient. This dataset is stored as object `math` in the `ANOM` package.

We investigate the differences of the binomial proportions versus the grand mean using a so-called Add-2 adjustment i.e., two successes and two failures are added to each group Agresti and Caffo (2000). Corresponding R functions can be found in the package `MCPAN` Schaarschmidt et al. (2013):

```
library(MCPAN)
add2 <- binomRDci(n=math$enrolled, x=math$proficient, names=math$school,
                  alternative="two.sided", method="ADD2", type="GrandMean")
add2p <- binomRDtest(n=math$enrolled, x=math$proficient, names=math$school,
                     alternative="two.sided", method="ADD2", type="GrandMean")
ANOM(add2, xlabel="School", ylabel="Proportion Proficient", pbin=add2p)
```

We find that roughly 74 % of students overall scored proficient, but there are large discrepancies between schools. Figure 8 reveals that neighborhood schools N1, N3, and N4 had significantly higher proportions of successful students in comparison to the mean of all schools. On the other hand, neighborhood school N2 and the alternative schools A3 and A4 performed exceptionally lousy (and significantly below the grand mean).

5.6 ANOM for variances

Quality control inspectors examined the uniformity of springs of four brands; they collected six springs per brand and determined the weight needed to stretch each spring by 0.1 inches Nelson et al. (2005). Their goal was to track down brands whose stiffness is highly variable, which would probably make them unmarketable.

The dataset is called `spring` in the `ANOM` package, and we add two columns to it: each brand's median weight, and each spring's absolute deviation from the corresponding brand median (robust Levene residuals):

```
spring$median <- tapply(spring$weight, spring$brand, median)[spring$brand]
spring$absdev <- with(spring, abs(weight - median))
```

Then we can construct the ANOM chart as usual:

```
spmodel <- lm(absdev ~ brand, spring)
sp <- glht(spmodel, mcp(brand="GrandMean"), alternative="greater")
ANOM(sp, xlabel="Brand", "Absolute Deviation from Median")
```

Figure 9 shows that the stiffnesses of brands 1 and 4 are more variable than average but still quite far from being significant.

5.7 Further examples

A PDF guide with more worked examples, including ANOM for Poisson data, ANOM in a two-way layout, nonparametric ANOM, and an in-depth elaboration on ANOM with mixed-effects models, is embedded in the `ANOM` package and can be accessed through `vignette("ANOM", package="ANOM")`.

6 Discussion

The analysis of means has been applied in quality control for several decades and with numerous extensions e.g., for discrete endpoints, unbalanced data, heterogeneous variances, various experimental designs, and more. Nonparametric approaches have been proposed as well as tests for comparing variances. Usually ANOM has been treated as if it were a stand-alone method, but in truth it belongs to a much broader class of multiple comparison procedures known as multiple contrast tests. One advantage of perceiving ANOM as a grand mean-type MCT is that it makes ANOM available in the non-commercial software R that facilitates computing multivariate t quantiles, which are the key to obtaining adjusted p-values, SCIs, and the popular ANOM charts, which we implemented for the first time in R. Our package `ANOM` is open-source software and can be downloaded for free.

In many cases, however, adjusting for multiplicity with critical values from a multivariate t-distribution is over the top. The power gain compared to a simple Bonferroni correction is practically irrelevant. The clear advantage of Bonferroni's method is that it is blindingly easy to use and widely known among non-statisticians. Especially when there are lots of group means to be compared to the grand mean, a simple Bonferroni correction works fine, and the improvement of applying a more sophisticated method which acknowledges the correlation in the multivariate t-distribution is negligible.

The decision whether to analyze data with ANOM or ANOVA should be based on subject-matter grounds since both procedures provide distinctly different information. Whenever the question of interest is 'Which groups differ from the grand mean?', a multiple comparison procedure like ANOM yields more meaningful results than purely global inference using the F-test, and it can even be more powerful in finding differences.

Acknowledgements

The authors would like to thank the editor and two anonymous referees for a few helpful suggestions, and Frank Schaarschmidt for his comments on an earlier version of the manuscript.

Funding

The work of the second author was supported by the German Research Foundation [DFG HO-1687/9].

References

A. Agresti and B. Caffo. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Am. Stat.*, 54(4):280–288, 2000.

S. T. Bakir. Analysis of means using ranks. *Commun. Stat.-Simul. Comput.*, 18(2):757–776, 1989.

S. T. Bakir. Analysis of means using ranks for the randomized complete block design. *Commun. Stat.-Simul. Comput.*, 23(2):547–568, 1994.

A. J. Bernard and P. S. Wludyka. Robust i-sample analysis of means type randomization tests for variances. *J. Stat. Comput. Simul.*, 69(1):57–88, 2001.

F. Bretz. An extension of the Williams trend test to general unbalanced linear models. *Comput. Stat. Data Anal.*, 50(7):1735–1748, 2006.

F. Bretz, A. Genz, and L. A. Hothorn. On the numerical availability of multiple comparison procedures. *Biom. J.*, 43(5):645–656, 2001.

F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. Chapman & Hall/CRC, Boca Raton, FL, 2010.

M. B. Brown and A. B. Forsythe. Robust tests for equality of variances. *J. Am. Stat. Assoc.*, 69(346): 364–367, 1974.

C.-H. Chang, N. Pal, W. K. Lim, and J.-J. Lin. Comparing several population means: a parametric bootstrap method, and its comparison with usual ANOVA F test as well as ANOM. *Comput. Stat.*, 25(1):71–95, 2010.

G. D. Djira and L. A. Hothorn. Detecting relative changes in multiple comparisons with an overall mean. *J. Qual. Technol.*, 41(1):60–65, 2009.

E. J. Dudewicz and P. R. Nelson. Heteroscedastic analysis of means (HANOM). *Am. J. Math. Management Sci.*, 23:143–181, 2003.

C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.*, 50(272):1096–1121, 1955.

N. L. Enrick. An analysis of means in a three-way factorial. *J. Qual. Technol.*, 8(4):189–196, 1976.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Springer, Heidelberg, Germany, 2009.

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. *mvtnorm: Multivariate normal and t distributions*, 2013. URL `http://CRAN.R-project.org/package=mvtnorm`. R package version 0.9-9996.

G. H. Guirguis and R. D. Tobias. On the computation of the distribution for the analysis of means. *Commun. Stat.-Simul. Comput.*, 33(4):861–887, 2004.

M. Halperin, S. W. Greenhouse, J. Cornfield, and J. Zalokar. Tables of percentage points for the studentized maximum absolute deviate in normal samples. *J. Am. Stat. Assoc.*, 50(269):185–195, 1955.

M. Hasler. *SimComp: Simultaneous comparisons for multiple endpoints*, 2012. URL `http://CRAN.R-project.org/package=SimComp`. R package version 1.7.0.

M. Hasler. Heteroscedasticity: multiple degrees of freedom vs. sandwich estimation. *Stat. Pap.*, 2014. doi: 10.1007/s00362-014-0640-4.

M. Hasler and L. A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biom. J.*, 50(5): 793–800, 2008.

A. J. Hayter and W. Liu. The power function of the studentised range test. *Ann. Stat.*, 18(1):465–468, 1990.

Z. He, J. J. Xiong, and M. Zhong. The comparison of the analysis of means and the analysis of variance. In Q. E. Shi, editor, *New Trends of Industrial Engineering and Engineering Management in New Century*, pages 419–421. Electronics Industry, 2001.

E. Herberich, J. Sikorski, and T. Hothorn. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLoS One*, 5(3):e9788, 2010.

K. Homa. Analysis of means used to compare providers' referral patterns. *Qual. Manag. Health Care*, 16 (3):256–264, 2007.

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biom. J.*, 50 (3):346–363, 2008.

J. C. Hsu. *Design of Experiments: Ranking and Selection (Essays in Honor of Robert E. Bechhofer)*, chapter Ranking and selection and multiple comparisons with the best, pages 23–33. Marcel Dekker, New York, NY, 1984.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 221–233, 1967.

F. Konietschke, L. A. Hothorn, and E. Brunner. Rank-based multiple test procedures and simultaneous confidence intervals. *Electron. J. Stat.*, 6:738–759, 2012.

F. Konietschke, S. Bösiger, E. Brunner, and L. A. Hothorn. Are multiple contrast tests superior to the ANOVA? *Int. J. Biostat.*, 9(1):1–11, 2013.

M. P. Kumar and C. V. Rao. ANOM-type graphical method for testing the equality of several variances. *Commun. Stat.-Simul. Comput.*, 27(2):459–468, 1998.

J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *Am. Stat.*, 54(3):217–224, 2000.

J. G. MacKinnon and H. White. Some heteroscedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econom.*, 29(3):305–325, 1985.

M. Mendeş and S. Yiğit. Comparison of ANOVA-F and ANOM tets with regard to type I error rate and test power. *J. Stat. Comput. Simul.*, 83(11):2093–2104, 2013.

M. A. Mohammed and R. Holder. Introducing analysis of means to medical statistics. *BMJ Qual. Saf.*, 21 (6):529–532, 2012.

H. Mukerjee, T. Robertson, and F. T. Wright. Comparison of several treatments with a control using multiple contrasts. *J. Am. Stat. Assoc.*, 82(399):902–910, 1987.

L. S. Nelson. Exact critical values for use with the analysis of means. *J. Qual. Technol.*, 15(1):40–44, 1983a.

P. R. Nelson. Exact critical points for the analysis of means. *Commun. Stat.-Theory Methods*, 11(6):699–709, 1982.

P. R. Nelson. A comparison of sample sizes for the analysis of means and the analysis of variance. *J. Qual. Technol.*, 15(1):33–39, 1983b.

P. R. Nelson. Power curves for the analysis of means. *Technometrics*, 27(1):65–73, 1985.

P. R. Nelson. Additional uses for the analysis of means and extended tables of critical values. *Technometrics*, 35(1):61–71, 1993.

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

P. R. Nelson, P. S. Wludyka, and K. A. F. Copeland. *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions*. SIAM, Philadelphia, PA, and ASA, Alexandria, VA, 2005.

E. R. Ott. Analysis of means - a graphical procedure. *Ind. Qual. Control*, 24:101–109, 1967.

E. R. Ott. Analysis of means - a graphical procedure. *J. Qual. Technol.*, 15(1):10–18, 1983.

P. Pallmann. *ANOM: Analysis of means*, 2015. URL `http://CRAN.R-project.org/package=ANOM`. R package version 0.4.

P. Pallmann, L. A. Hothorn, and G. D. Djira. A Levene-type test of homogeneity of variances against ordered alternatives. *Comput. Stat.*, 29:1593–1608, 2014.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and nonlinear mixed effects models*, 2013. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-111.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org`.

P. F. Ramig. Applications of the analysis of means. *J. Qual. Technol.*, 15(1):19–25, 1983.

C. V. Rao. Analysis of means - a review. *J. Qual. Technol.*, 37(4):308–315, 2005.

C. V. Rao and S. H. Krishna. A graphical method for testing the equality of several variances. *J. Appl. Stat.*, 24(3):279–288, 1997.

SAS Institute Inc. *SAS/QC® 12.1 Users Guide*. SAS Institute Inc., Cary, NC, 2012.

F. Schaarschmidt, M. Sill, and L. A. Hothorn. Approximate simultaneous confidence intervals for multiple contrasts of binomial proportions. *Biom. J.*, 50(5):782–792, 2008.

F. Schaarschmidt, D. Gerhard, and M. Sill. *MCPAN: Multiple comparisons using normal approximation*, 2013. URL `http://CRAN.R-project.org/package=MCPAN`. R package version 1.1-15.

H. Scheffé. *The Analysis of Variance*. Wiley, New York, NY, 1959.

E. G. Schilling. A systematic approach to the analysis of means, part I. Analysis of treatment effects. *J. Qual. Technol.*, 5(3):92–108, 1973a.

E. G. Schilling. A systematic approach to the analysis of means, part II. Analysis of contrasts, part III. Analysis of non-normal data. *J. Qual. Technol.*, 5(4):147–159, 1973b.

J. Subramani. Analysis of means for experimental designs with missing observations. *Commun. Stat.-Simul. Comput.*, 21(7):2045–2057, 1992.

J. W. Tukey. The problem of multiple comparisons. In H. I. Braun, editor, *The Collected Works of John W. Tukey*, volume VIII. Chapman and Hall, New York, NY, 1994.

D. A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971. Correction: 31 (1975), 1019.

D. A. Williams. The comparison of several dose levels with a zero dose control. *Biometrics*, 28(2):519–531, 1972.

P. Wludyka and P. Sa. A robust i-sample analysis of means type randomization test for variances in unbalanced designs. *J. Stat. Comput. Simul.*, 74(10):701–726, 2004.

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

P. S. Wludyka and P. R. Nelson. Analysis of means type tests for variances using sub sampling and jacknifing. In T. Hayakawa, M. Aoshima, and K. Shimizu, editors, *Multivariate Statistical Inference (MSI) 2000: Multivariate Statistical Analysis in Honor of Professor Minoru Siotani on his 70th Birthday*, volume 17 of *American Journal of Mathematical and Management Sciences*, pages 31–60, 1997b.

P. S. Wludyka and P. R. Nelson. Two non-parametric, analysis-of-means-type tests for homogeneity of variances. *J. Appl. Stat.*, 26(2):243–256, 1999.

P. S. Wludyka, P. R. Nelson, and P. R. Silva. Power curves for the analysis of means for variances. *J. Qual. Technol.*, 33(1):60–65, 2001.

P. Wretenberg, U. P. Arborelius, and F. Lindberg. The effects of a pneumatic stool and a one-legged stool on lower limb joint load and muscular activity during sitting and rising. *Ergonomics*, 36(5):519–535, 1993.

A. B. Yeh, L. Huwang, and Y.-F. Wu. A likelihood-ratio-based EWMA control chart for monitoring variability of multivariate normal processes. *IIE Trans.*, 36(9):865–879, 2004.

A. Zeileis. Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw.*, 11(10): 1–17, 2004. URL `http://www.jstatsoft.org/v11/i10/`.

Figures

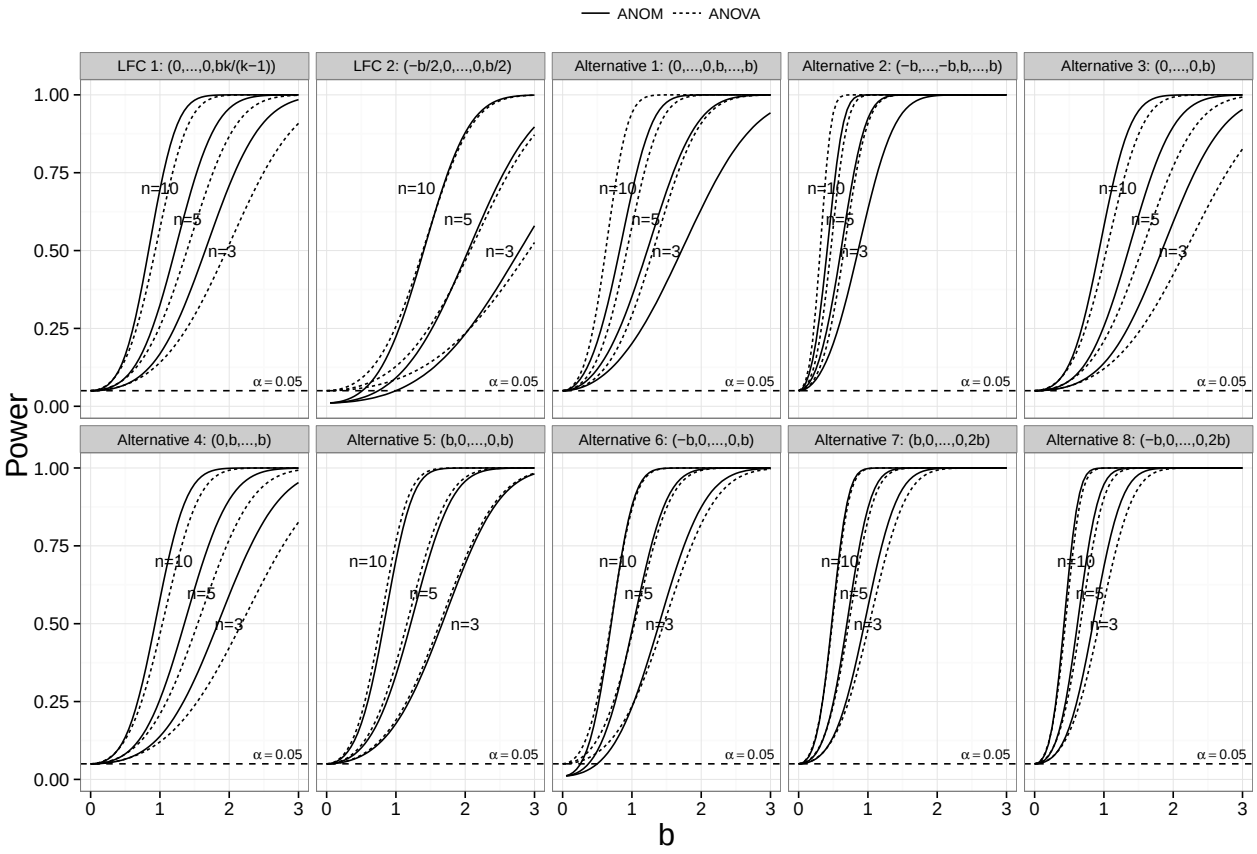

Figure 1: Power of ANOM and ANOVA for $k = 10$ groups, sample sizes $n_i = \{3, 5, 10\}$ per group (balanced one-way design), and $0 \leq b \leq 3$ under ten different configurations of the alternative, including both LFCs ($\alpha = 0.05$).

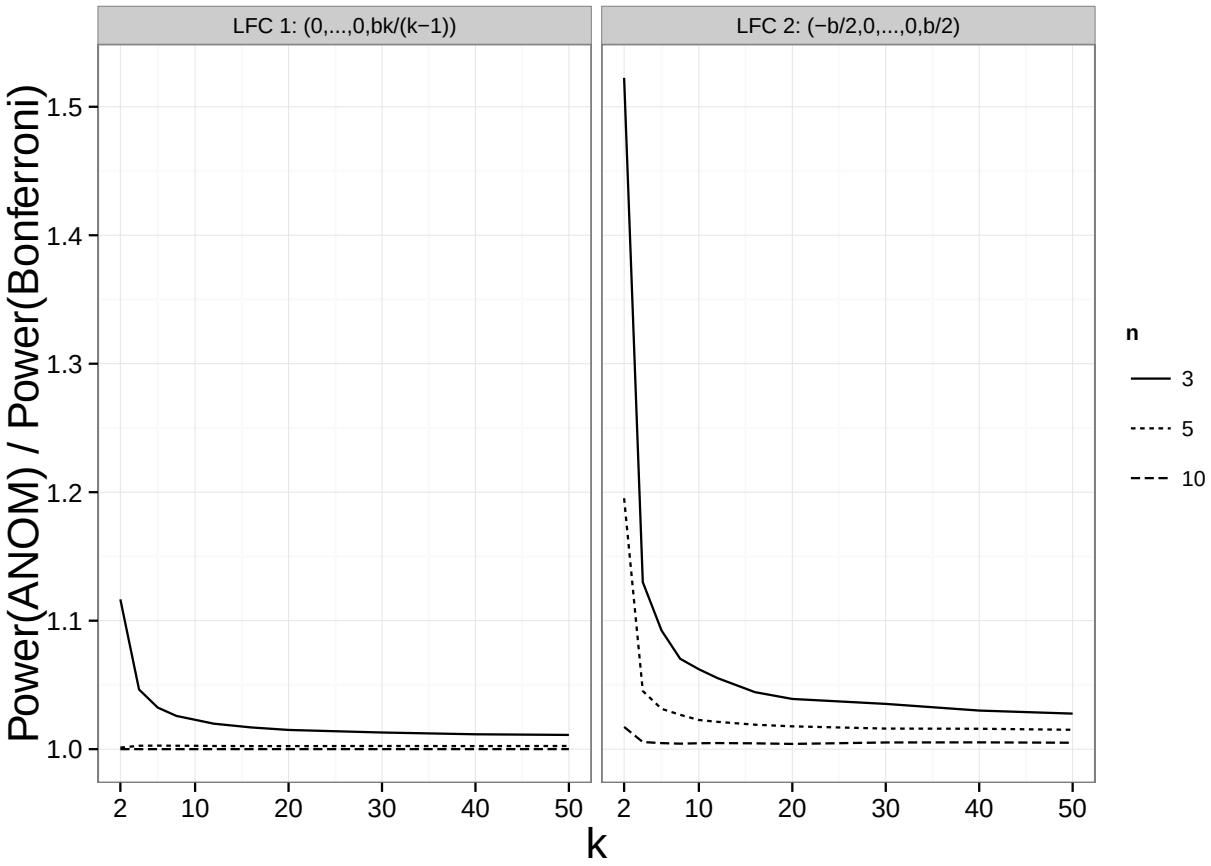

Figure 2: Power ratio of ANOM using critical points from a multivariate t-distribution and a simple Bonferroni adjustment for $2 \leq k \leq 50$ groups, sample size $n = \{3, 5, 10\}$ per group (balanced one-way design), and $b = 2$ under both LFCs of the alternative ($\alpha = 0.05$).

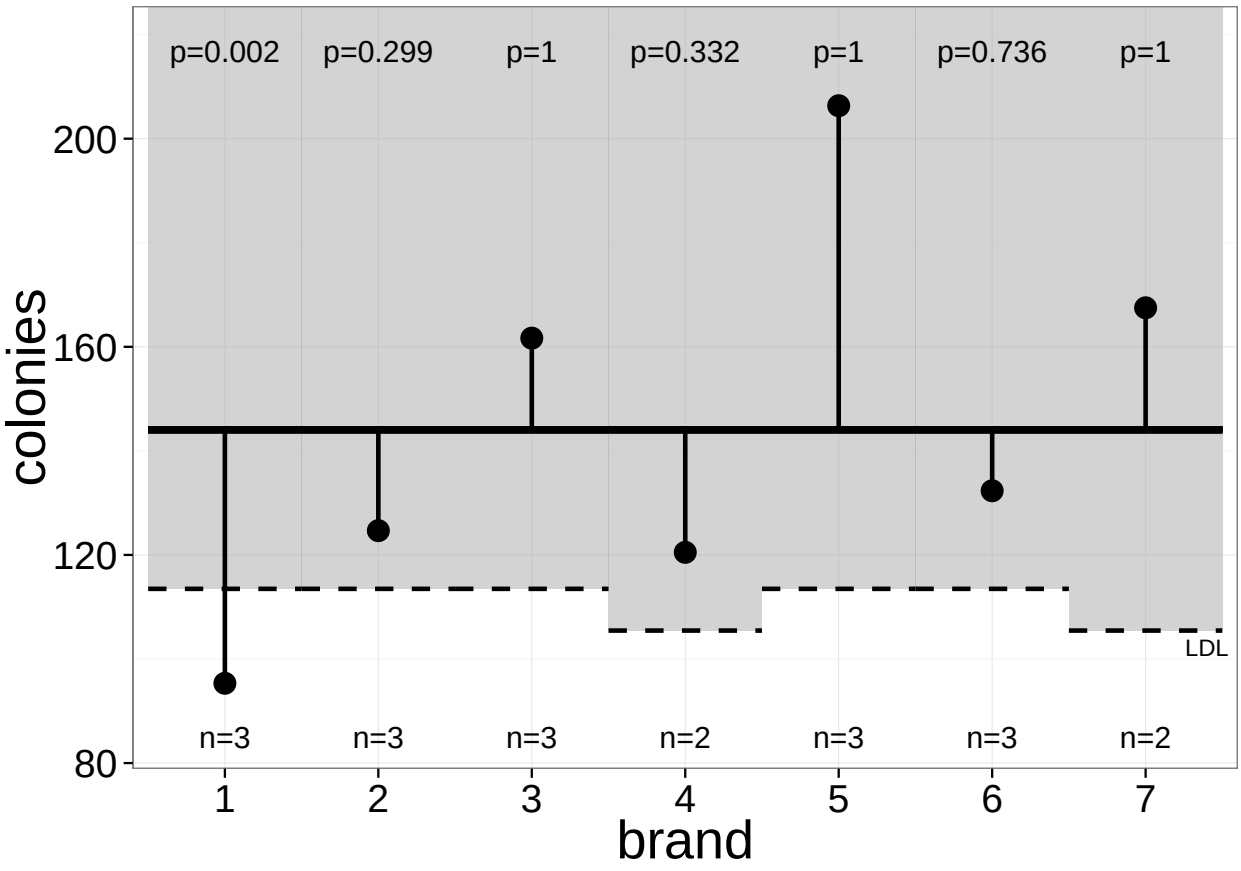

Figure 3: ANOM decision chart for the water filter data (assuming homogeneous variances).

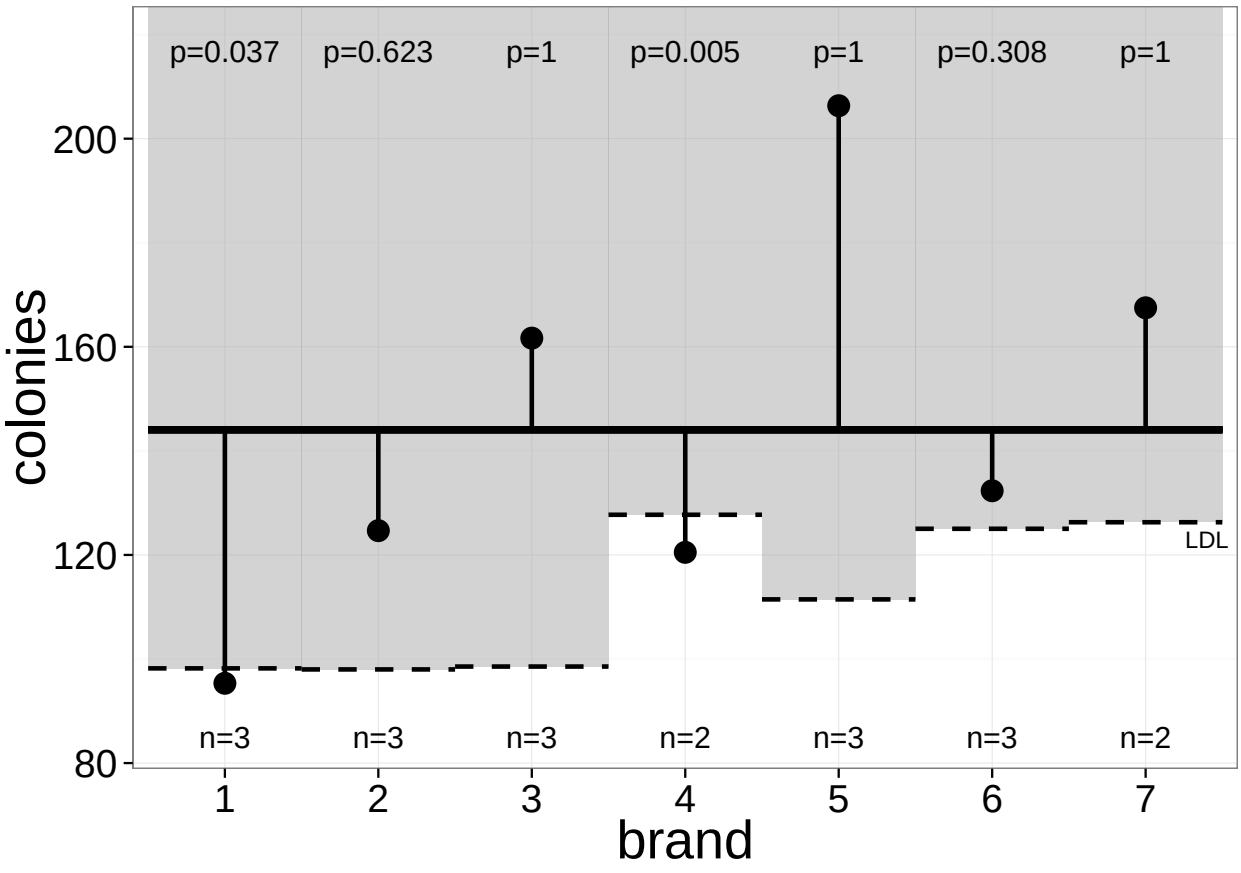

Figure 4: ANOM decision chart for the water filter data (accounting for heterogeneous variances via sandwich covariance estimation).

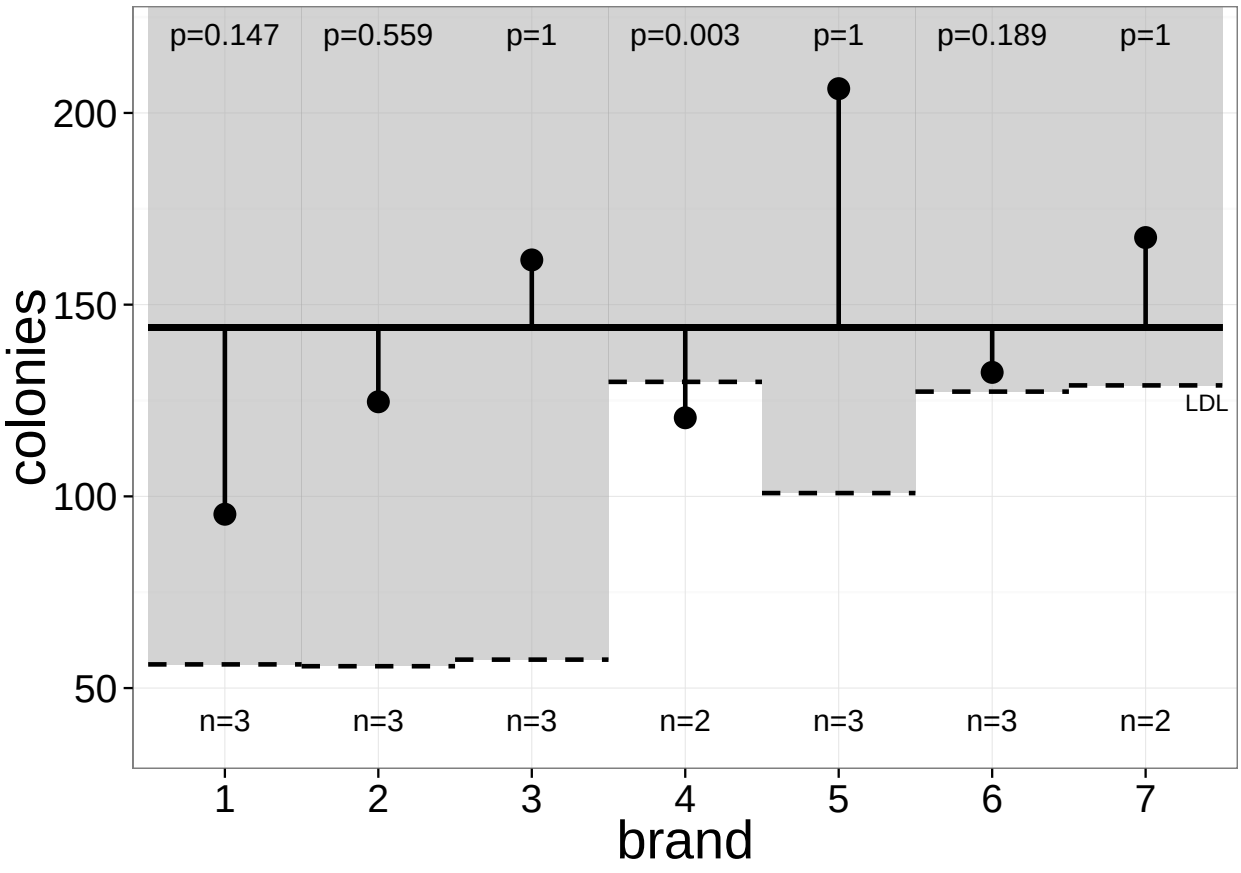

Figure 5: ANOM decision chart for the water filter data (accounting for heterogeneous variances using multiple degrees of freedom).

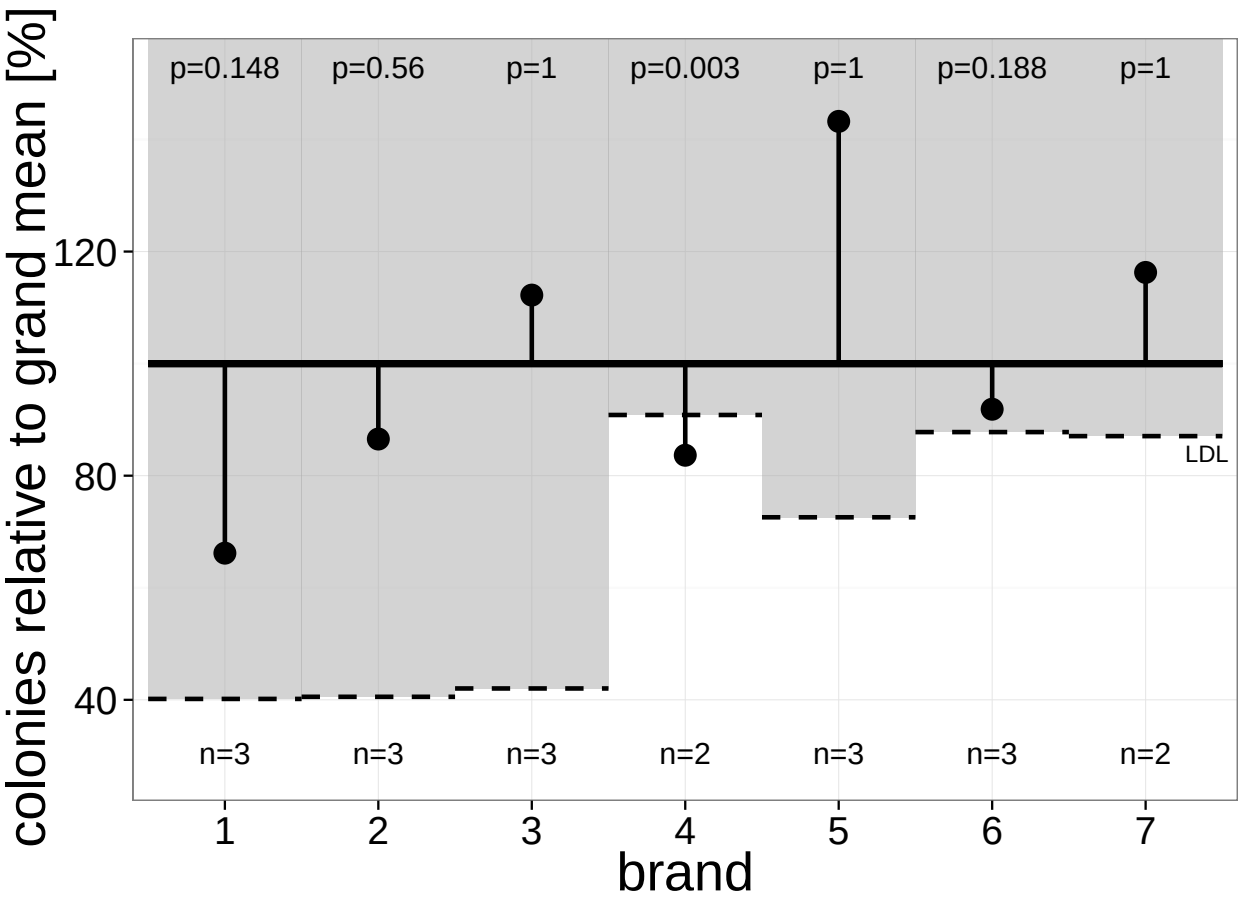

Figure 6: ANOM decision chart for the water filter data (accounting for heterogeneous variances using multiple degrees of freedom) in terms of percentages.

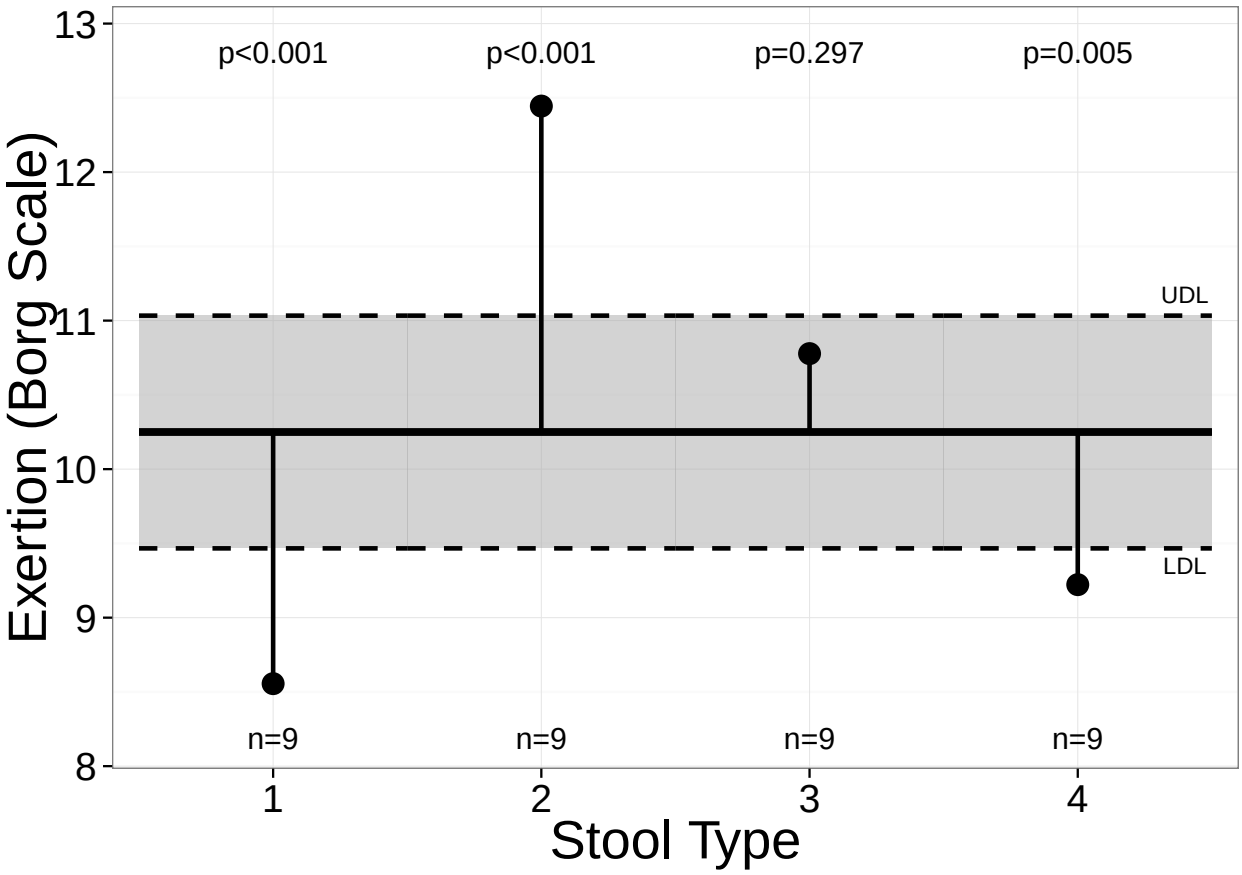

Figure 7: ANOM decision chart for the ergonomic stools data.

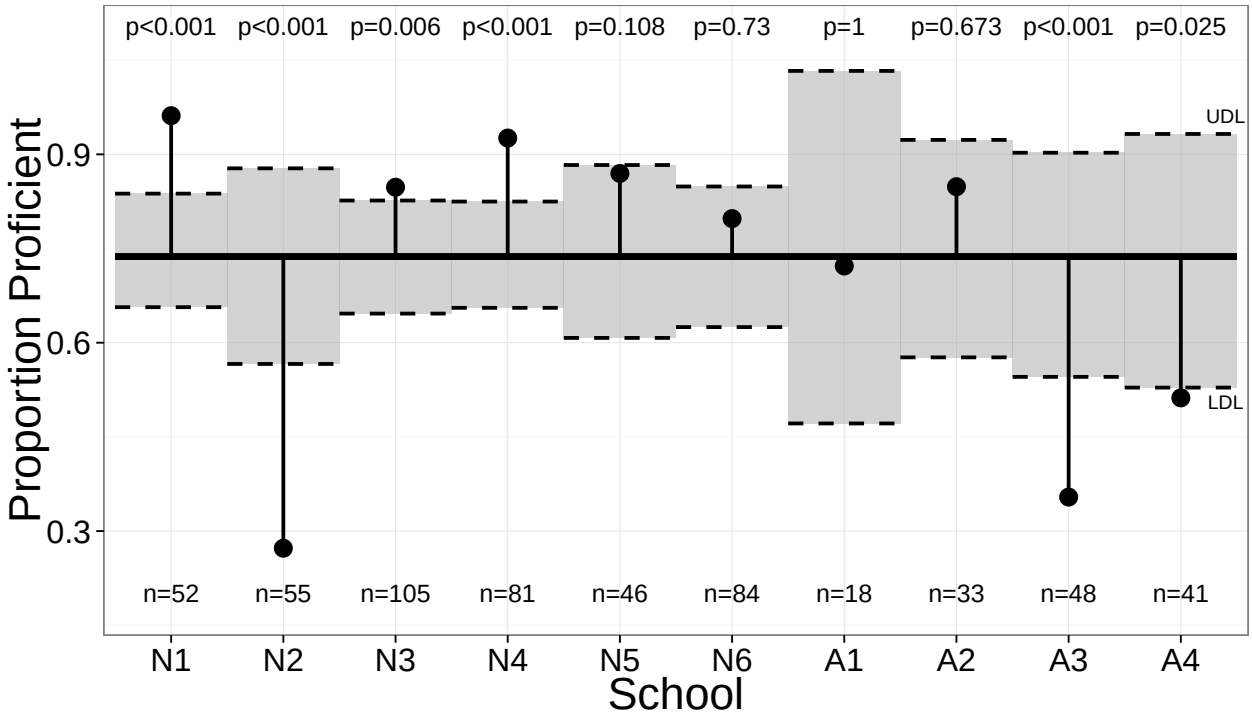

Figure 8: ANOM decision chart for the math proficiency test data.

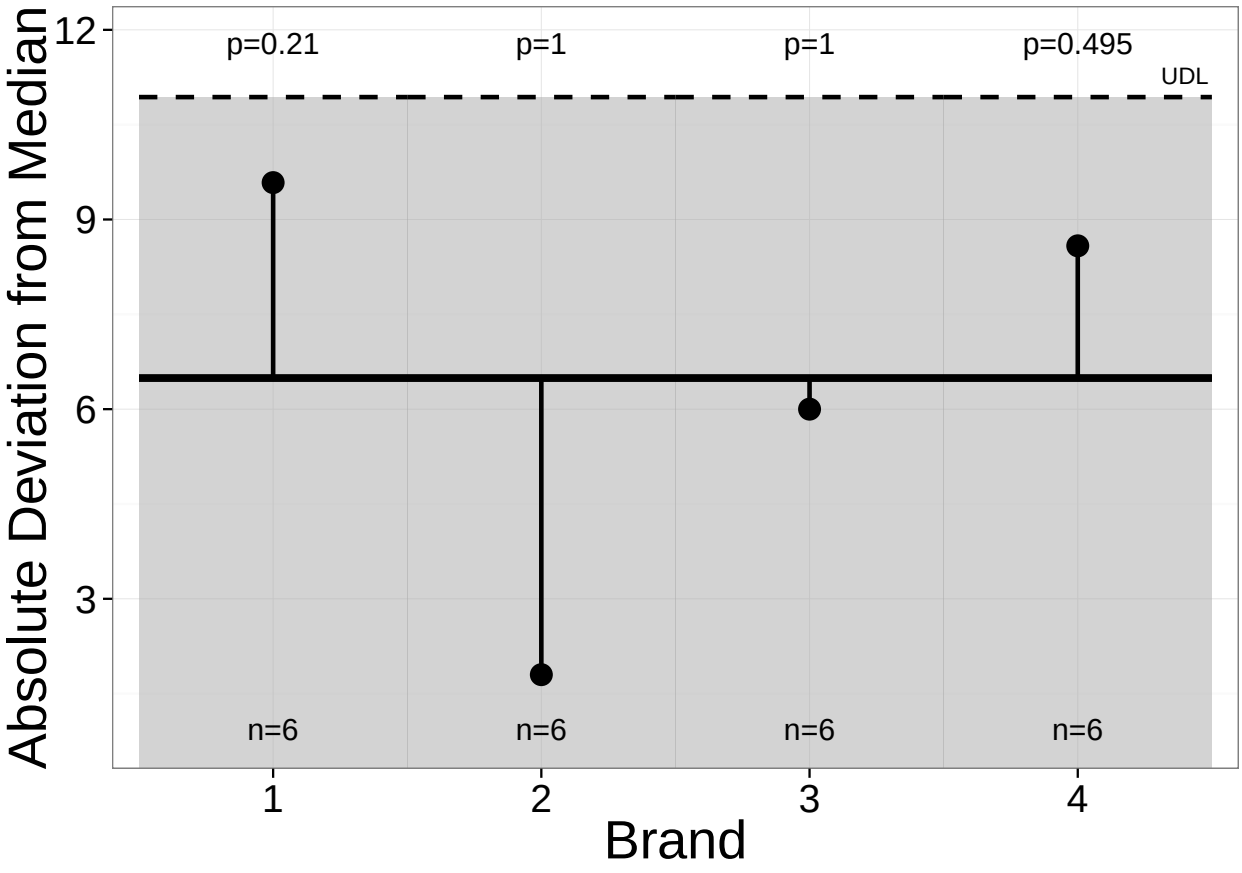

Figure 9: ANOM decision chart for the spring data.