Journal of Modern Applied Statistical Methods

Volume 12 | Issue 2 Article 6

11-1-2013

# The Impact of Continuity Violation on ANOVA and Alternative Methods

Björn Lantz
*Chalmers University of Technology, Gothenburg, Sweden*, bjorn.lantz@chalmers.se

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 105-120.

ISSN 1538 − 9472

# The Impact of Continuity Violation on ANOVA and Alternative Methods

**Björn Lantz**
Chalmers University of Technology
Gothenburg, Sweden

The normality assumption behind ANOVA and other parametric methods implies that response variables are measured on continuous scales. A simulation approach is used to explore the impact of continuity violation on the performance of statistical methods commonly used by applied researchers to compare locations across several groups.

*Keywords:* Continuity violation, ANOVA, Brown-Forsythe, Welch, Kruskal-Wallis

## Introduction

One of the standard research procedures to explore the effects of the violation of an assumption underlying a statistical method is to perform an experimental study using Monte Carlo simulation. The one-way ANOVA for comparing locations across three or more groups and alternative test procedures such as the Brown-Forsythe test, Welch test, and Kruskal-Wallis test have been subject to similar research since the 1970s (e.g., Glass et al., 1972; Bevan et al., 1974; Keselman et al., 1977), and continue to be studied today (e.g., Lantz, 2013; Cribbie et al., 2012; Cribbie et al., 2007). Some workers conclud the one-way ANOVA is relatively robust against violations of the homoscedasticity assumption as well as against violations of the normality assumption. However, textbooks in statistics (e.g., Lomax and Hahs-Vaughn, 2007; Ryan, 2007) often recommend the Brown-Forsythe and Welch tests when the data are characterised by apparent heteroscedasticity, particularly at unequal sample sizes, or the Kruskal-Wallis test when the data are clearly not mound-shaped.

Most research regarding the normality assumption on which the ANOVA relies focuses on continuous distributions that differ from the normal in terms of shape, skewness, or kurtosis (e.g., Ito, 1980; Khan and Rayner, 2003). The fact that the underlying distribution is assumed to be normal does not, however, imply

*Dr. Lantz is an Associate Professor in the Department of Technology Management and Economics, Chalmers University of Technology. Email him at: bjorn.lantz@chalmers.se.*

only a mound shape, zero skewness, and zero excess kurtosis; it also requires that the data be continuous by nature. In applied research, data subject to statistical analyses have often been collected using discrete scales. Assume, for example, that the subjects participating in a psychological experiment perform a certain task four times, and that the number of successful trials is recorded for each subject. In this case, an arbitrarily chosen subject will have zero, one, two, three, or four successful trials. Although means and standard deviations can be used to describe the locations of different groups of subjects in cases like this, the one-way ANOVA and parametric alternatives like the Brown-Forsythe test and the Welch test are, at least technically, invalidated as methodologies to compare means across groups. This is because the dependent variable is assumed to be continuous even though it actually is discretely distributed, with only a small number of possible values.

The impact of the relative violation of the continuity assumption emerges more strikingly when there are fewer possible values that the variable can take. Krieg (1999) derived equations for calculating the bias induced by coarse measurement scales, and showed that the bias is reduced as the number of scale points increases. Hence, one would assume that statistical comparisons of locations across groups should be relatively unproblematic even if data are discrete as long as the number of possible variable values is large. In contrast, it might be a problem when the number of possible variable values is small, or when the violation of continuity is more severe. However, explicit analyses on the violation of continuity are scarce in the literature, and most of the research in this area seems to be related to the scale coarseness issue (Symonds, 1924) rather than to continuity violation. Scale coarseness refers to the fact that Likert-type and similar ordinal-level scales are collapsed into discrete scale points to simplify the data collection process, even though the underlying constructs are assumed to be continuous. When respondents are faced with a scale that does not have a sufficient number of response options, information loss will occur. Continuity violation and scale coarseness are obviously related phenomena, but scale coarseness (see Symonds, 1924) is an issue primarily related to data collection, whereas violation of continuity (see Bevan et al., 1974) is an issue strictly related to data analysis.

Although there seems to be little research on how continuity violation affects the statistical methods commonly used to compare locations across groups, nevertheless some results can be found in the literature. Bevan et al. (1974) considered the appropriateness of ANOVA techniques when the response variable was discretely distributed and able to take three, five, or seven different values.

Their results suggested that the ANOVA was relatively robust to continuity violations with respect to Type I errors. However, Bevan et al. (1974) did not examine how power or alternative methods were affected by continuity violation. Gregoire and Driver (1987) tested the performance of selected parametric (including the F test) and nonparametric tests of location on the basis of sampling results from simulated Likert-type data and concluded that there was no clear-cut superiority for either type of test. It should be noted that their aim was to compare the methods rather than to explore the impact of scale discreteness. Rasmussen (1988) extended (and corrected) the analysis by Gregoire and Driver (1987), and demonstrated that the Type I and Type II error rates were not seriously compromised by the use of discrete data.

The impact of continuity violation on the significance and power of statistical methods commonly used to compare locations across several groups is explored in the one-way ANOVA layout and its robust alternatives, the Brown-Forsythe test, the Welch test, and the non-parametric Kruskal-Wallis test. The one-way ANOVA is based on the idea that the true means in groups are more likely to be equal if the variation between the groups is small compared to the variation within the groups. The Brown-Forsythe and Welch tests are considered robust compared with ANOVA, because their definitions of variation within groups are based on the relationships between the different sample sizes in the different groups, as opposed to a simple pooled variance estimate, which means that they become less sensitive to heteroscedasticity (see, e.g., Tomarken & Serlin, 1986). The Kruskal-Wallis test is considered robust because it is based on ranks rather than actual values, which means that the underlying distribution does not matter so long as the observed values can be ranked.

## Methodology

By definition, there is no discrete equivalent with only a few steps to the normal distribution, because a normally distributed random variable is unrestricted upward as well as downward, and can therefore take extreme values. Hence, it is technically impossible to make an exact evaluation of the impact of continuity violation on statistical methods that rest on the normality assumption. The best approximation compares results from a mound-shaped discrete distribution where the number of steps can be varied with results where the normality assumption holds, means and variances being equal. The binomial distribution is one such mound-shaped discrete distribution that exists for any number of steps, and it approaches the normal distribution when the number of steps becomes large

(Aczel and Sounderpandian, 2009). For example, Figure 1 displays the probability density function for the normal distribution with $\mu = 2$ and $\sigma^2 = 1$ and for the probability distribution for the binomial distribution with five possible outcomes, $\mu = 2$ and $\sigma^2 = 1$. Therefore, the binomial distribution is used in this study as an approximation of a continuity-violated normal distribution.

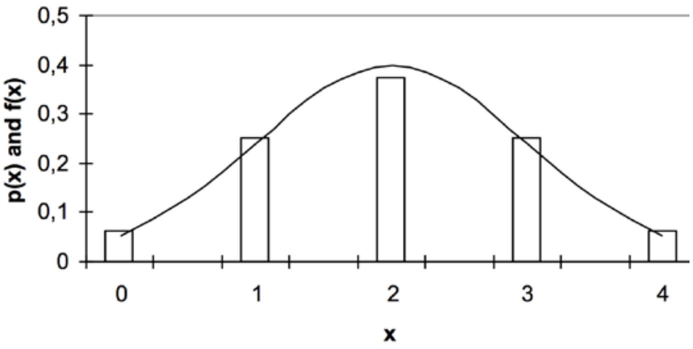

**Figure 1.** A binomial distribution and its corresponding normal distribution

An experimental design with three populations and four different combinations of small (defined as 5 observations) and large (defined as 25 observations) sample sizes was used. Discrete scales based on binomial distributions with two, three, four, five, and seven steps were used in each case. For each combination, the proportion of significant ANOVA, Brown-Forsythe, Welch, and Kruskal-Wallis (adjusted for ties) tests was compared to the proportion of significant tests when data was simulated from normal distributions with identical means and variances.

For each combination of sample sizes, test procedure, and number of steps, five different effect sizes were used. Table 1 shows the manner in which the values for the parameter *p* in the binomial distributions were varied for different values of *n* to achieve a suitable range of effect sizes (see Cohen, 1992), ranging from no effect ($f = 0.00$) to a very large effect ($f = 0.65$). For any individual combination of values of *p* and *n*, the distribution mean and variance could easily be calculated in order to obtain the corresponding normal distribution, because the mean is defined as *np* and the variance as *np*(1-*p*) for the binomial distribution (Aczel and Sounderpandian, 2009). For example, with five steps ($n = 4$) and $f = 0.25$, $p_1 = 0.424$, $p_2 = 0.500$, and $p_3 = 0.576$ because the mean and the variance then become 2.12 and 1.22 for group 1, 2.50 and 1.25 for group 2, and 2.88 and 1.22 for group 3, respectively, corresponding to the medium effect size $f = 0.25$. Hence, the simulated impact of continuity violation in this case is based on a comparison between the normal distributed random variables $X_1 \sim N(2.12, 1.22)$, $X_2 \sim N(2.50, 1.25)$, and $X_3 \sim N(2.88, 1.22)$ and the binomial distributed random variables $Y_1 \sim B(4, 0.424)$, $Y_2 \sim B(4, 0.500)$, and $Y_3 \sim B(4, 0.576)$.

**Table 1.** Values for the parameter *p* in the binomial distributions

| | | **Effect size (Cohen's *f*)** | | | | |
|---|---|---|---|---|---|---|
| **Steps** | **Group** | ***0.000*** | ***0.100*** | ***0.250*** | ***0.400*** | ***0.650*** |
| 2 | 1 | 0.500 | 0.439 | 0.351 | 0.273 | 0.166 |
| | 2 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | 3 | 0.500 | 0.561 | 0.649 | 0.727 | 0.834 |
| 3 | 1 | 0.500 | 0.457 | 0.394 | 0.334 | 0.245 |
| | 2 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | 3 | 0.500 | 0.543 | 0.607 | 0.667 | 0.756 |
| 4 | 1 | 0.500 | 0.465 | 0.413 | 0.346 | 0.285 |
| | 2 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | 3 | 0.500 | 0.535 | 0.587 | 0.654 | 0.715 |
| 5 | 1 | 0.500 | 0.469 | 0.424 | 0.380 | 0.311 |
| | 2 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | 3 | 0.500 | 0.531 | 0.576 | 0.620 | 0.689 |
| 7 | 1 | 0.500 | 0.475 | 0.438 | 0.401 | 0.343 |
| | 2 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | 3 | 0.500 | 0.525 | 0.562 | 0.599 | 0.657 |

For each combination of distribution (normal and binomial), sample sizes (25/25/25, 5/5/5, 5/5/25, and 5/25/25), test procedure (ANOVA, Brown-Forsythe, Welch, and Kruskal-Wallis), number of steps (two, three, four, five, and seven), and size of effect (no, small, medium, large, and very large), 50,000 hypothesis tests based on simulated random numbers were conducted, where the null hypothesis, corresponding to no difference between the locations of the populations, was challenged at an alpha level of 0.05 in all cases. Hence, 40,000,000 tests of simulated data were performed in the study. All simulations and analytical procedures were conducted using Microsoft Excel 2010.

## Results

Table 2 displays the number of significant tests where the discrete scale has two steps. For a better understanding of the reliability of the statistics presented in this section, it should be noted that the standard error of a sample proportion at a sample size of 50,000 is about 0.002 when the proportion is 0.5, and it decreases to about 0.001 when the proportion is 0.05 or 0.95. When one distribution is characterised by a significantly larger proportion of significant tests than the other for a given combination of effect size, sample sizes, and test method, this is indicated with an asterisk (*).

**Table 2**: Proportion of significant tests, mean value 0.5 (two steps)

| | | ANOVA | | Brown-Forsythe | | Welch | | Kruskal-Wallis | |
|---|---|---|---|---|---|---|---|---|---|
| **ES** | **n1/n2/n3** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** |
| 0 | 25,25,25 | 0.052 | 0.051 | 0.052 | 0.050 | 0.051 | 0.051 | 0.052 | 0.05 |
| | 5,5,5 | 0.058* | 0.052 | 0.057* | 0.041 | 0.000 | 0.039* | 0.059* | 0.045 |
| | 5,5,25 | 0.052 | 0.051 | 0.046 | 0.047 | 0.000 | 0.052* | 0.052* | 0.044 |
| | 5,25,25 | 0.052 | 0.051 | 0.070* | 0.053 | 0.011 | 0.056* | 0.044 | 0.047 |
| 0.1 | 25,25,25 | 0.112* | 0.107 | 0.112* | 0.106 | 0.110* | 0.104 | 0.112* | 0.102 |
| | 5,5,5 | 0.071* | 0.061 | 0.069* | 0.049 | 0.000 | 0.046* | 0.071* | 0.052 |
| | 5,5,25 | 0.074* | 0.070 | 0.061 | 0.062 | 0.001 | 0.067* | 0.074* | 0.060 |
| | 5,25,25 | 0.077* | 0.072 | 0.092* | 0.074 | 0.020 | 0.075* | 0.068 | 0.066 |
| 0.25 | 25,25,25 | 0.468 | 0.466 | 0.468 | 0.465 | 0.463 | 0.466 | 0.468* | 0.446 |
| | 5,5,5 | 0.131* | 0.111 | 0.128* | 0.091 | 0.000 | 0.088* | 0.132* | 0.098 |
| | 5,5,25 | 0.201 | 0.196 | 0.145 | 0.151* | 0.009 | 0.152* | 0.201* | 0.170 |
| | 5,25,25 | 0.235* | 0.223 | 0.215 | 0.215 | 0.079 | 0.199* | 0.218* | 0.207 |
| 0.4 | 25,25,25 | 0.858 | 0.879* | 0.858 | 0.878* | 0.855 | 0.889* | 0.858 | 0.870* |
| | 5,5,5 | 0.248* | 0.214 | 0.244* | 0.180 | 0.000 | 0.177* | 0.251* | 0.190 |
| | 5,5,25 | 0.447 | 0.455 | 0.308 | 0.321* | 0.041 | 0.327* | 0.445* | 0.400 |
| | 5,25,25 | 0.508 | 0.506 | 0.453 | 0.481* | 0.218 | 0.459* | 0.491 | 0.484 |
| 0.65 | 25,25,25 | 0.999 | 1.000 | 0.999 | 1.000 | 0.978 | 1.000* | 0.999 | 1.000 |
| | 5,5,5 | 0.525 | 0.519 | 0.521* | 0.456 | 0.000 | 0.500* | 0.535* | 0.488 |
| | 5,5,25 | 0.855 | 0.907* | 0.648 | 0.694* | 0.147 | 0.760* | 0.850 | 0.868* |
| | 5,25,25 | 0.894 | 0.921* | 0.845 | 0.910* | 0.422 | 0.910* | 0.889 | 0.919* |

As a scale with two steps has the greatest degree of continuity violation, one would expect the most differences between the discrete binomial and continuous normal cases. When all sample sizes are small, the ANOVA becomes more powerful as a result of scale discreteness (i.e. the probability of avoiding a Type II error is often higher when data are discrete than when they are continuous), but at the cost of an elevated probability of a Type I error. For some combinations of effect sizes and unequal sample sizes, the ANOVA becomes more powerful due to scale discreteness without an elevated probability of a Type I error. When all sample sizes are large, it becomes less powerful when the effect size is large, but more powerful when the effect size is small.

The Brown-Forsythe test becomes more powerful due to scale discreteness when all sample sizes are small and, for small and medium effect sizes, when exactly one sample size is small, but in both cases at the cost of an elevated probability of a Type I error. When exactly one sample size is large, it becomes less powerful for medium and larger effect sizes.

The Welch test algorithm does not work satisfactorily for coinciding dichotomous distributions when at least one sample size is small, which is the reason for the very low numbers for the discrete scale in those cases. Note, however, that for large sample sizes, it becomes more powerful when the effect size is small, but less powerful when the effect size is large.

The Kruskal-Wallis test becomes more powerful as a result of scale discreteness when at most one sample size is large, but in both cases at the cost of an elevated probability of a Type I error. For small and medium effect sizes, it becomes more powerful when exactly one sample size is small. For large sample sizes, however, it becomes more powerful at small and medium effect sizes but less powerful at large effect sizes.

Finally, note that there are no significant differences in performance between the four methods when they are used to analyse data on a discrete scale with two steps as long as the sample sizes are large; the only exception is that the Welch test performs less well when the effect size is very large.

Table 3 displays the number of significant tests where the discrete scale has three steps. Here, the ANOVA shows no significant difference in performance due to scale discreteness, with the exception that it becomes more powerful when all sample sizes are small and the effect size is large. The Brown-Forsythe test exhibits elevated power when at least one sample size is small, but again, at the cost of an elevated probability of a Type I error. The Welch test displays the opposite reaction: it becomes less powerful when at least one sample size is small, but with a reduced probability of a Type I error. The Kruskal-Wallis test behaves erratically for some sample size combinations, and becomes less powerful at some effect sizes but more powerful at others. However, there is no significant change in the probability of a Type I error for any combination of sample sizes. Finally, note that there are no significant differences in performance between the four methods when they are used to analyse data on a discrete scale with three steps as long as the sample sizes are large.

**Table 3**: Proportion of significant tests, mean value 1.0 (three steps)

| | | **ANOVA** | | **Brown-Forsythe** | | **Welch** | | **Kruskal-Wallis** | |
|---|---|---|---|---|---|---|---|---|---|
| **ES** | **n1/n2/n3** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** |
| 0 | 25,25,25 | 0.051 | 0.05 | 0.051 | 0.049 | 0.052* | 0.049 | 0.049 | 0.048 |
| | 5,5,5 | 0.054 | 0.052 | 0.049* | 0.041 | 0.033 | 0.041* | 0.044 | 0.046 |
| | 5,5,25 | 0.050 | 0.050 | 0.053* | 0.047 | 0.037 | 0.054* | 0.046 | 0.044 |
| | 5,25,25 | 0.050 | 0.051 | 0.056* | 0.053 | 0.044 | 0.057* | 0.046 | 0.046 |
| 0.1 | 25,25,25 | 0.108 | 0.108 | 0.108 | 0.107 | 0.108 | 0.105 | 0.104 | 0.103 |
| | 5,5,5 | 0.060 | 0.062 | 0.055* | 0.050 | 0.038 | 0.049* | 0.049 | 0.054* |
| | 5,5,25 | 0.069 | 0.071 | 0.067* | 0.062 | 0.049 | 0.067* | 0.062 | 0.062 |
| | 5,25,25 | 0.073 | 0.074 | 0.080* | 0.074 | 0.063 | 0.075* | 0.069 | 0.068 |
| 0.25 | 25,25,25 | 0.461 | 0.46 | 0.460 | 0.459 | 0.457 | 0.455 | 0.453* | 0.437 |
| | 5,5,5 | 0.113 | 0.110 | 0.106* | 0.091 | 0.076 | 0.088* | 0.094 | 0.096 |
| | 5,5,25 | 0.194 | 0.190 | 0.158* | 0.149 | 0.127 | 0.152* | 0.174* | 0.168 |
| | 5,25,25 | 0.226 | 0.224 | 0.221* | 0.212 | 0.189 | 0.197* | 0.214* | 0.207 |
| 0.4 | 25,25,25 | 0.865 | 0.875 | 0.864 | 0.874 | 0.859 | 0.877* | 0.859 | 0.861 |
| | 5,5,5 | 0.222* | 0.211 | 0.209* | 0.178 | 0.156 | 0.172* | 0.191 | 0.186 |
| | 5,5,25 | 0.438 | 0.435 | 0.333* | 0.324 | 0.294 | 0.319* | 0.402* | 0.385 |
| | 5,25,25 | 0.509 | 0.508 | 0.474 | 0.469 | 0.434 | 0.444* | 0.491* | 0.479 |
| 0.65 | 25,25,25 | 0.999 | 1.000 | 0.999 | 1.000 | 0.999 | 1.000 | 0.999 | 0.999 |
| | 5,5,5 | 0.496 | 0.501 | 0.474* | 0.446 | 0.336 | 0.442* | 0.444 | 0.458* |
| | 5,5,25 | 0.848 | 0.873* | 0.686 | 0.706* | 0.614 | 0.711* | 0.812 | 0.827* |
| | 5,25,25 | 0.905 | 0.918* | 0.852 | 0.886* | 0.791 | 0.880* | 0.894 | 0.905 |

Table 4 displays the number of significant tests where the discrete scale has four steps. In this case, the ANOVA shows no significant difference in performance due to scale discreteness, except that it becomes powerful when at most one sample size is large and the effect size is very large. The Brown-Forsythe test becomes more powerful when all sample sizes are small, but less powerful at unequal sample sizes when the effect size is very large. The Welch test performs somewhat erratically, as it exhibits reduced power when sample sizes are unequal, but increased power when all sample sizes are small and the effect size is medium or large. The Kruskal-Wallis test also behaves erratically: it becomes too conservative when all sample sizes are small, which reduces power. In contrast, it becomes more powerful at the medium effect size when all sample sizes are large and at unequal sample sizes when the effect size is medium or large. As in the previous cases, note that there are no significant differences in performance between the four methods when they are used to analyse data on a discrete scale with four steps as long as the sample sizes are large.

**Table 4**: Proportion of significant tests, mean value 1.5 (four steps)

| | | **ANOVA** | | **Brown-Forsythe** | | **Welch** | | **Kruskal-Wallis** | |
|---|---|---|---|---|---|---|---|---|---|
| **ES** | **n1/n2/n3** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** |
| 0 | 25,25,25 | 0.050 | 0.052 | 0.049 | 0.052 | 0.050 | 0.051 | 0.048 | 0.049 |
| | 5,5,5 | 0.049 | 0.052 | 0.044 | 0.042 | 0.040 | 0.040 | 0.040 | 0.045* |
| | 5,5,25 | 0.050 | 0.050 | 0.051* | 0.047 | 0.053 | 0.054 | 0.045 | 0.044 |
| | 5,25,25 | 0.050 | 0.052 | 0.055 | 0.054 | 0.054 | 0.056 | 0.047 | 0.048 |
| 0.1 | 25,25,25 | 0.111 | 0.110 | 0.110 | 0.109 | 0.110 | 0.107 | 0.107 | 0.105 |
| | 5,5,5 | 0.059 | 0.062 | 0.053* | 0.049 | 0.047 | 0.047 | 0.049 | 0.054* |
| | 5,5,25 | 0.072 | 0.071 | 0.065 | 0.062 | 0.066 | 0.068 | 0.063 | 0.061 |
| | 5,25,25 | 0.075 | 0.074 | 0.079 | 0.076 | 0.072 | 0.076* | 0.070 | 0.070 |
| 0.25 | 25,25,25 | 0.459 | 0.457 | 0.458 | 0.456 | 0.455 | 0.451 | 0.449* | 0.434 |
| | 5,5,5 | 0.109 | 0.111 | 0.100* | 0.091 | 0.092* | 0.087 | 0.093 | 0.098* |
| | 5,5,25 | 0.191 | 0.189 | 0.151 | 0.149 | 0.132 | 0.150* | 0.169 | 0.165 |
| | 5,25,25 | 0.227 | 0.224 | 0.215 | 0.21 | 0.181 | 0.194* | 0.213* | 0.204 |
| 0.4 | 25,25,25 | 0.933 | 0.939 | 0.933 | 0.938 | 0.930 | 0.940 | 0.928 | 0.930 |
| | 5,5,5 | 0.264 | 0.257 | 0.244* | 0.222 | 0.221* | 0.210 | 0.228 | 0.228 |
| | 5,5,25 | 0.530 | 0.525 | 0.403 | 0.40 | 0.352 | 0.389* | 0.479* | 0.470 |
| | 5,25,25 | 0.618 | 0.615 | 0.563 | 0.565 | 0.512 | 0.540* | 0.596* | 0.584 |
| 0.65 | 25,25,25 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 1.000 | 0.999 | 0.999 |
| | 5,5,5 | 0.492 | 0.504* | 0.465* | 0.452 | 0.425 | 0.435* | 0.441 | 0.458* |
| | 5,5,25 | 0.852 | 0.864* | 0.703 | 0.714* | 0.651 | 0.707* | 0.808 | 0.816 |
| | 5,25,25 | 0.913 | 0.918 | 0.862 | 0.879* | 0.835 | 0.872* | 0.900 | 0.903 |

Table 5 displays the number of significant tests where the discrete scale has five steps. The ANOVA displays a similar pattern as with four steps; there is no significant difference in performance due to scale discreteness, except that it becomes powerful when exactly one sample size is large and the effect size is very large. The Brown-Forsythe test also shows a similar pattern (as in the previous case), becoming more powerful when all sample sizes are small and the effect size is at least medium, but less powerful at unequal sample sizes when the effect size is very large. The performance of the Welch test, however, behaves somewhat differently when the number of steps is increased from four to five. It becomes more conservative when at least one sample size is small, which reduces its power when the effect size is small. The effect disappears when the effect size is medium or large, but returns when it is very large. The Kruskal-Wallis test continues to behave erratically along the same pattern as with four steps. Finally, under a medium effect size, the Kruskal-Wallis test has significantly less power than the other three methods even if all sample sizes are large.

**Table 5**: Proportion of significant tests, mean value 2.0 (five steps)

| | | ANOVA | | Brown-Forsythe | | Welch | | Kruskal-Wallis | |
|---|---|---|---|---|---|---|---|---|---|
| **ES** | **n1/n2/n3** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** |
| 0 | 25,25,25 | 0.050 | 0.050 | 0.050 | 0.049 | 0.049 | 0.049 | 0.047 | 0.047 |
| | 5,5,5 | 0.050 | 0.052 | 0.044 | 0.042 | 0.037 | 0.041* | 0.041 | 0.045* |
| | 5,5,25 | 0.050 | 0.050 | 0.049 | 0.047 | 0.047 | 0.053* | 0.044 | 0.044 |
| | 5,25,25 | 0.049 | 0.049 | 0.053 | 0.053 | 0.053 | 0.057* | 0.046 | 0.047 |
| 0.1 | 25,25,25 | 0.111 | 0.110 | 0.111 | 0.110 | 0.110 | 0.108 | 0.107 | 0.105 |
| | 5,5,5 | 0.059 | 0.062 | 0.052 | 0.050 | 0.045 | 0.048* | 0.049 | 0.054* |
| | 5,5,25 | 0.071 | 0.073 | 0.064 | 0.064 | 0.064 | 0.071* | 0.062 | 0.064 |
| | 5,25,25 | 0.076 | 0.078 | 0.079 | 0.080 | 0.074 | 0.080* | 0.071 | 0.072 |
| 0.25 | 25,25,25 | 0.463 | 0.462 | 0.463 | 0.461 | 0.458 | 0.456 | 0.450* | 0.441 |
| | 5,5,5 | 0.110 | 0.108 | 0.099* | 0.090 | 0.086 | 0.086 | 0.092 | 0.096* |
| | 5,5,25 | 0.192 | 0.188 | 0.153 | 0.148 | 0.151 | 0.148 | 0.170* | 0.165 |
| | 5,25,25 | 0.230 | 0.227 | 0.218 | 0.213 | 0.201 | 0.196 | 0.215* | 0.207 |
| 0.4 | 25,25,25 | 0.864 | 0.869 | 0.864 | 0.869 | 0.860 | 0.867 | 0.854 | 0.852 |
| | 5,5,5 | 0.216 | 0.216 | 0.197* | 0.183 | 0.173 | 0.174 | 0.185 | 0.189 |
| | 5,5,25 | 0.433 | 0.431 | 0.332 | 0.331 | 0.317 | 0.318 | 0.389 | 0.384 |
| | 5,25,25 | 0.515 | 0.511 | 0.466 | 0.468 | 0.437 | 0.436 | 0.491* | 0.479 |
| 0.65 | 25,25,25 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| | 5,5,5 | 0.493 | 0.500 | 0.464* | 0.448 | 0.415 | 0.428* | 0.443 | 0.452* |
| | 5,5,25 | 0.846 | 0.858* | 0.706 | 0.718* | 0.664 | 0.702* | 0.803 | 0.811 |
| | 5,25,25 | 0.912 | 0.918 | 0.861 | 0.875* | 0.843 | 0.869* | 0.899 | 0.902 |

Table 6 displays the number of significant tests where the discrete scale has seven steps. The ANOVA now becomes more conservative when at most one sample size is large, and it has reduced power at the medium effect size when all sample sizes are small. Both the Brown-Forsythe test and the Welch test lose power at very large effect sizes when sample sizes are unequal, but become more powerful at the large effect size when all sample sizes are small. The Kruskal-Wallis test becomes too conservative and loses power when all sample sizes are small. It is also characterised by significantly less power than the other three methods when the effect size is small or medium.

**Table 6**: Proportion of significant tests, mean value 3.0 (seven steps)

| | | ANOVA | | Brown-Forsythe | | Welch | | Kruskal-Wallis | |
|---|---|---|---|---|---|---|---|---|---|
| **ES** | **n1/n2/n3** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** | **Bin** | **Norm** |
| 0 | 25,25,25 | 0.051 | 0.051 | 0.051 | 0.051 | 0.051 | 0.051 | 0.049 | 0.049 |
| | 5,5,5 | 0.048 | 0.052* | 0.041 | 0.042 | 0.040 | 0.040 | 0.040 | 0.046* |
| | 5,5,25 | 0.049 | 0.052* | 0.047 | 0.048 | 0.051 | 0.054 | 0.043 | 0.045 |
| | 5,25,25 | 0.050 | 0.05 | 0.054 | 0.053 | 0.057 | 0.057 | 0.047 | 0.046 |
| 0.1 | 25,25,25 | 0.108 | 0.109 | 0.108 | 0.108 | 0.108 | 0.106 | 0.103 | 0.102 |
| | 5,5,5 | 0.059 | 0.058 | 0.050 | 0.047 | 0.048 | 0.046 | 0.048 | 0.051* |
| | 5,5,25 | 0.072 | 0.069 | 0.063* | 0.060 | 0.068 | 0.067 | 0.063 | 0.061 |
| | 5,25,25 | 0.076 | 0.076 | 0.077 | 0.075 | 0.076 | 0.075 | 0.071 | 0.068 |
| 0.25 | 25,25,25 | 0.457 | 0.459 | 0.456 | 0.458 | 0.451 | 0.452 | 0.440 | 0.436 |
| | 5,5,5 | 0.108 | 0.114* | 0.094 | 0.094 | 0.088 | 0.088 | 0.091 | 0.099* |
| | 5,5,25 | 0.188 | 0.188 | 0.152 | 0.149 | 0.152 | 0.148 | 0.166 | 0.166 |
| | 5,25,25 | 0.226 | 0.223 | 0.213 | 0.21 | 0.198 | 0.196 | 0.211 | 0.205 |
| 0.4 | 25,25,25 | 0.869 | 0.870 | 0.869 | 0.869 | 0.863 | 0.866 | 0.856 | 0.853 |
| | 5,5,5 | 0.215 | 0.213 | 0.191* | 0.181 | 0.177* | 0.170 | 0.184 | 0.187 |
| | 5,5,25 | 0.427 | 0.429 | 0.331 | 0.326 | 0.315 | 0.317 | 0.380 | 0.379 |
| | 5,25,25 | 0.519 | 0.514 | 0.472 | 0.465 | 0.442 | 0.437 | 0.49 | 0.482 |
| 0.65 | 25,25,25 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| | 5,5,5 | 0.491 | 0.499 | 0.453 | 0.447 | 0.420 | 0.416 | 0.437 | 0.449* |
| | 5,5,25 | 0.842 | 0.852 | 0.708 | 0.718 | 0.673 | 0.693* | 0.794 | 0.806* |
| | 5,25,25 | 0.913 | 0.921 | 0.860 | 0.873* | 0.848 | 0.867* | 0.896 | 0.902 |

Table 7 provides a qualitative summary of the simulation results. Note that cases where continuity violation has no or negligible impact on the probability of a Type I error ($\alpha$) or power ($1 - \beta$) are not explicitly discussed. For the ANOVA, scale discreteness is considered to have a marked impact on power because the number of differences between the normal distribution and the binomial distribution that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases. Sample sizes are also considered to have a marked impact on power, because power is reduced in several cases where sample sizes are unequal, but not when they are equal. However, continuity violation was not found to have a marked impact on the probability of a Type I error in any of the examined aspects, which is in line with previous research (Bevan et al., 1974).

**Table 7**: Impact of continuity violation – summary of simulation results

| | **ANOVA** | | **Brown-Forsythe** | | **Welch** | | **Kruskal-Wallis** | |
|---|---|---|---|---|---|---|---|---|
| Explanatory variable | *α* | 1–*β* | *α* | 1–*β* | *α* | 1–*β* | *α* | 1–*β* |
| Scale discreteness | Negligible impact | Marked impact | Marked impact | Marked impact | Marked impact | Marked impact | Marked impact | Marked impact |
| Effect size | n/a | Negligible impact | n/a | Negligible impact | n/a | Marked impact | n/a | Negligible impact |
| Sample sizes | Negligible impact | Marked impact | Negligible impact | Marked impact | Marked impact | Marked impact | Marked impact | Marked impact |

For the Brown-Forsythe test, scale discreteness is considered to have a marked impact on the probability of a Type I error because the significant differences between the normal and binomial distributions that can be seen when the discrete scale has only a few steps and at least one sample size is small seem to decrease when the number of steps increases. Furthermore, scale discreteness is considered to have a marked impact on power because the number of differences between the normal and binomial distributions that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases. Sample sizes are also considered to have a marked impact on power because power is increased in several cases where at least one sample size is small.

For the Welch test, scale discreteness is considered to have a marked impact on the probability of a Type I error because the significant differences that can be seen between the normal and binomial distributions when the discrete scale has only at a few steps seem to decrease when the number of steps increase. Sample sizes are also considered to have a marked impact on the probability of a Type I error because this probability is consistently different in several cases where at least one sample size is small, but not when all sample sizes are large. Furthermore, scale discreteness is considered to have a marked impact on power because the number of differences between the normal distribution and the binomial distribution that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases. Effect size is also considered to have a marked impact on power, particularly in combination with scale discreteness, because the number of observable differences between the normal and binomial distributions tends to decrease faster at the medium and large effect sizes than at the small and very large effect sizes. In addition, sample sizes are also considered to have a marked impact on power because the presence of unequal sample sizes seems to reduce power in general.

Finally, for the Kruskal-Wallis test, scale discreteness is considered to have a marked impact on the probability of a Type I error because the significant differences that can be seen between the normal and binomial distributions when the discrete scale has only a few steps, and all sample sizes are small, seem to be reversed when the number of steps increases. Sample sizes are also considered to have a marked impact on the probability of a Type I error because this probability is consistently different when all sample sizes are small, but not otherwise. Furthermore, scale discreteness is considered to have a marked impact on power because the number of differences between the normal and binomial distributions that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases, although the major difference occurs between two and three steps. Sample sizes are also considered to have a marked impact on power because power is changed in several cases where at least one sample size is small.

## Conclusion

Violation of continuity affects the performance of four statistical methods that are commonly used to compare locations across several groups. A dichotomous scale changes the probability of a Type I error for methods in all cases when all sample sizes are small and in many other cases when at least one sample size is small. However, the effect seems to decline as the number of scale points is increased, which is in line with theory (Krieg, 1999) and with similar published simulation results (e.g., Bevan et al., 1974). The probability of a Type II error also seems to decline as the number of scale points is increased, although the pattern is different for different methods and sample size combinations.

This should not be seen as an argument in favour of a larger number of steps when, for example, Likert-type and similar discrete scales are used. Even a small number of steps may be too many for the respondent if comprehensible instructions and labelling of response alternatives are not included to enable the respondent to conceptualize and respond in spatial terms (Cox, 1980). Often, and for a variety of reasons, scales with only a few steps must be used during data collection processes, and the results in this study can help determine a suitable statistical procedure to compare locations across groups in such situations.

In summary, ANOVA seems to be the most robust alternative of the four procedures when scales are discrete, as the violation of continuity has relatively little impact on its performance. The Brown-Forsythe test can become more powerful when scales are discrete and at least one sample size is small, but at the cost of an elevated probability of a Type I error. When all sample sizes are large and scales with at least three steps are used, neither the ANOVA nor the Brown-Forsythe test displays any significant sensitivity to continuity violation at any effect size level. Hence, these two tests can be used to make reliable analyses of discrete data in such situations. The Welch test can become less powerful when scales are discrete, in some cases even at large sample sizes. The Kruskal-Wallis test responds erratically to scale discreteness, particularly at unequal sample sizes, and has significantly less power than the other three methods when sample sizes are large.

Even though the impact of continuity violation on ANOVA and the three alternative methods examined here seem to be relatively small in most realistic situations (the most obvious exception is when the Welch test is used to analyse dichotomous data), applied researchers should consider the above results when using these statistical methods to analyse data collected with discrete scales. The main implications of this study can be summarised as follows:

- Collect data using continuous scales, if reasonable.
- Be aware that power can be reduced when discrete scales are used. The reduction in power becomes less pronounced when the number of scale points is increased, but in some situations, it remains significant for scales with up to seven points.
- Be aware that the actual probability of a Type I error may be affected when dichotomous scales are used if at least one sample size is small.
- Do not use the Welch test with dichotomous data.

Future research in this area should further explore the effects of data discreteness by combining continuity violation with, for example, heteroscedasticity. In general, the effects of concurrent violations can produce anomalous effects not observed in separate violations (see, e.g., Zimmerman, 1998). Other types of parametric methods should also be tested for their robustness against continuity violation.

## References

Aczel, A. D. & Sounderpandian, J. (2009). *Complete Business Statistics*, New York: McGraw-Hill Irwin.

Bevan, M. F., Denton, J. Q., & Myers, J. L. (1974). The robustness of the *F* test to violations of continuity and form of treatment population. *British Journal of Mathematical and Statistical Psychology*, *27*, 199-204. doi: 10.1111/j.2044-8317.1974.tb00540.x

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi: 10.1037/0033-2909.112.1.155

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *27*, 407-422. doi: 10.2307/3150495

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology*, *65*, 56-73. doi: 10.1111/j.2044-8317.2011.02014.x

Cribbie, R. A., Wilcox, R. R., Bewell, C., & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, *6*, 117-132.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, *42*, 237-288. doi: 10.3102/00346543042003237

Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin*, *101*, 159-165. doi: 10.1037/0033-2909.101.1.159

Ito, K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnainh, Ed., *Handbook of Statistics*, Vol. 1, Amsterdam, Holland.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977), An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, *30*: 213-221. doi: 10.1111/j.2044-8317.1977.tb00742.x

Khan, A., & Rayner, G. D. (2003). Robustness to non-normality of common tests for the many sample location problem. *Journal of Applied Mathematics and Decision Sciences*, *7*(4), 187-206. doi: 10.1155/S1173912603000178

Krieg, E. F. (1999). Biases induced by coarse measurement scales. *Educational and Psychological Measurement*, *59*, 749-766. doi: 10.1177/00131649921970125

Lantz, B. (2013). The impact of sample non-normality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology*, *66*, 224-244. doi: 10.1111/j.2044-8317.2012.02047.x

Lomax, R. G. & Hahs-Vaughn, D. L. (2007). *An Introduction to Statistical Concepts*. NJ: Lawrence Erlbaum Associates.

Rasmussen, J. L. (1988). Analysis of Likert-Scale Data: A Reinterpretation of Gregoire and Driver. *Psychological Bulletin*, *105*, 167-170. doi: 10.1037/0033-2909.105.1.167

Ryan, T. P. (2007). *Modern Experimental Design*. Hoboken, NJ: John Wiley & Sons.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456-461. doi: 10.1037/h0074469

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90-99.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, *67*(1), 55-68. doi: 10.1080/00220979809598344