Power study of anova versus Kruskal-Wallis test

T. Van Hecke
University College Ghent,
Faculty of Applied Engineering Sciences,
Ghent (Belgium)
Tanja.VanHecke@hogent.be
November 16, 2010

Abstract
This paper describes the comparison of the anova and the Kruskal-Wallis test by means of the power when violating the assumption about normally distributed populations. The permutation method is used as a simulation method to determine the power of the test. It appears that in the case of asymmetric populations the non-parametric Kruskal-Wallis test performs better than the parametric equivalent anova method.

Keywords: anova; kruskal-wallis; permutation test; power

1 Introduction
Wilcox (1995) and Glass et al. (1972) have done research into the effect of the violation of the normality assumption in the case of anova (Analysis of Variance). Wilcox concluded that non-normality has some effect on the Type I error, but the effect is minimal when the variances are equal. However, the effect is much more noticeable in regard to the Type II error. Glass et al. focused more on the skewness of the population and concluded the same as Wilcox when the variances were equal. Blair et al. (1980) made an analysis of the power of the $t$-test and the Wilcoxon’s rank-sum test. While it is generally agreed that parametric procedures are a little more powerful than nonparametric procedures when the assumptions of the parametric procedures are met, what about the case of data for which those assumptions are not met, for example, the typical Likert scale data? Such data violate the normality assumption and often the homogeneity of variance assumption made when we conduct a traditional parametric analysis. Nanna and Sawilowsky [7] demonstrated that with Likert scale data, the Wilcoxon rank sum test has a considerable power advantage over the parametric $t$-test. The Wilcoxon procedure has a power advantage with both small and large samples, with the advantage actually increasing with sample size. It is our intention to investigate whether similar conclusions like
those of Nanna and Sawilowsky can be drawn about the power for the anova and the Kruskal-Wallis method [6].

2 Anova versus Kruskal-Wallis test

2.1 Assumptions

The anova $F$-test is used to test the equality of $k$ ($k > 2$) population means, so the null hypothesis is

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k. \quad (1)$$

However, the $F$-test has some assumptions that are frequently ignored and often violated when used in real world applications. These assumptions include that the data is normally distributed and that the population variances are equal.

A one-way anova may yield inaccurate estimates of the $p$-value when the data are not normally distributed at all. The Kruskal-Wallis test is the non-parametric analogue of a one-way anova, which does not make assumptions about normality. Like most non-parametric tests, it is performed on the ranks of the measurement observations. It does, however, assume that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes (for example one is skewed to the right and another is skewed to the left), the Kruskal-Wallis test may give inaccurate results, see Fagerland and Sandvik (2009).

Instead of null hypothesis (1), the null hypothesis of the Kruskal-Wallis test states that the samples are from identical populations.

2.2 How the Kruskal-Wallis test works

The Kruskal-Wallis test starts by substituting the rank in the overall data set for each measurement value. The smallest value gets a rank of 1, the second-smallest gets a rank of 2, etc. Tied observations get average ranks; thus if there were four identical values occupying the fifth, sixth, seventh and eighth smallest places, all would get a rank of 6.5.

The sum of the ranks $R_i$ is calculated for each group $i$ ($i = 1, 2, \ldots, k$) of size $n_i$, then the test statistic $H$ from (2) is calculated, which basically represents the variance of the ranks among groups, with an adjustment for the number of ties. $H$ is approximately $\chi^2$ distributed, with the degrees of freedom equal to the number of groups $k$ minus 1.

$$H = \frac{12}{N (N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3 (N + 1), \quad N = \sum_{i=1}^{k} n_i. \quad (2)$$
3 Power by means of the permutation test

3.1 Permutation Test

The permutation test (see Good (1994) and Efron et al. (1993) and also known as randomization test) was developed first by R. A. Fisher, the founder of classical statistical testing. However, in his later years Fisher lost interest in the permutation method because there were no computers in his days to automate such a laborious method. The randomization test is a test procedure in which data are randomly re-assigned so that an exact \( p \)-value is calculated based on the permutated data. That is often quite impossible, as too laborious: suppose that we have three groups with 20 observations per group. There are \( \frac{60!}{(20!)^3} \) possible different permutations of those observations into three groups. In practice the Monte Carlo sampling is used, i.e. we draw random samples out of the permutations to estimate the result of drawing all possible samples.

In the case of a one-way anova design with \( k \) samples of different treatments each of size \( n \), the basic idea is that if the null hypothesis is true, then any permutation of the observations between the \( k \) treatments has the same chance to occur as any other permutation. An observation in group 1 could just as easily have fallen in group 3, and vice versa. So we simply permute the observations across groups, calculate the \( F \), repeat this multiple times, and find the percentage out of \( n_{\text{tot}} \) repetitions in which the calculated values of \( F \) exceeded the \( F_{\text{obs}} \) obtained from the original data. This is the \( p \)-value under the null hypothesis. In resampling, instead of consulting a theoretical distribution, the researcher artificially simulates chance.

The steps for a multiple-treatment permutation test:

- Compute the \( F \)-value of the given samples, called \( F_{\text{obs}} \).
- Re-arrange the \( kn \) observations in \( k \) samples of size \( n \).
- For each permutation of the data, compare the \( F \)-value with the \( F_{\text{obs}} \).

For the upper tailed test, compute the \( p \)-value as

\[
p = \frac{\#(F > F_{\text{obs}})}{n_{\text{tot}}}.
\]

If the \( p \)-value is less than or equal to the predetermined level of significance \( \alpha \), then we reject \( H_0 \).

3.2 Comparing the power of anova and the Kruskal-Wallis test

The power of a test can be estimated by performing Monte Carlo simulations in which artificial data is generated. This analysis will be conducted using a program that will generate random data from a specified distribution with given parameters. Anova and the Kruskal-Wallis test will then be conducted at
a significance level of $\alpha$. This simulation will be repeated $n_{\text{tot}}$ times for each set of parameters, and then the proportion of times the null hypothesis is rejected is recorded. If the parameters are set for unequal means, the empirical power of the test is resulted. From these results, the power can be analyzed for anova and the Kruskal-Wallis test to see which method performs better under violations of assumptions.

4 Testexample

We consider the testexample where three different samples should help us to decide whether the populations they are taken from, are equal. To compare the power of the Kruskal-Wallis and the anova method, we consider the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$  \hspace{.5cm} (4)

and the alternative hypothesis

$$H_1 : \mu_1 + d = \mu_2 \text{ and } \mu_1 + 2d = \mu_3.$$  \hspace{.5cm} (5)

We perform the permutation test (2500 permutations) for $\alpha = 0.05$ in case of 2500 samples from the chosen distributions for the anova and the Kruskal-Wallis test.

![Figure 1: The power of anova and the Kruskal-Wallis test as a function of the differences between the means for various sample sizes in the case of a normal distribution.](image)

From Figure 1 it is clear that in the case of symmetrical distributions (such as the normal distribution, with $\sigma = 1$) the power of the Kruskal-Wallis test is comparable to that of the parametric anova test.
Figure 2: The power of anova and the Kruskal-Wallis test as a function of the differences between the means for various sample sizes in the case of a lognormal distribution.

But in the case of asymmetrical distributions (such as the lognormal distribution, with $\sigma = 1$ or the $\chi^2$ distribution with 3 d.f.), the performance of the non-parametric Kruskal-Wallis test results in a higher power (see Figure 2 and Figure 3 respectively), which is more evident as $n$ increases. In all figures the power is presented as a function of $d$.
5 Conclusions

For non-symmetrical distributions the non-parametrical Kruskal-Wallis test results in a higher power compared to the classical one-way anova. The results of the simulations show that an analysis of the data is needed before a test on differences in central tendencies is conducted. Although the literature and textbooks state that the $F$-test is robust under the violations of assumptions, these results show that the power suffers a significant decrease.

References


