**Taking Parametric Assumptions Seriously:** **Arguments for the Use of Welch’s *F*-test instead of the Classical *F*-test in One-way ANOVA**

Marie Delacre1\*, Christophe Leys1, Youri L. Mora1, and Daniël Lakens2

1Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

2Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, Netherlands

\*Correspondence concerning this article should be addressed to Marie Delacre, Service of analysis of the data, Université Libre de Bruxelles, Bruxelles. Address: Avenue Franklin D. Roosevelt 50; Email: [marie.delacre@ulb.ac.be](mailto:marie.delacre@ulb.ac.be); Telephone number: 02/6503243.

Word Count: 4972

**Funding**

This work was supported by the Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013.

**Author Note**

All materials required to reproduce the analyses reported in this article are available at https://osf.io/ru9tz/

# Abstract

Parametric tests rely on two main assumptions: normality of the distribution and equality of variances. We argue that these assumptions are often unrealistic in the field of psychology. We underline the current lack of attention to parametric assumptions through an analysis of researchers’ practices. Through Monte Carlo simulations we illustrate the consequences of performing the classic parametric *F-*test for ANOVA when the test assumptions are not met on the Type I error rate and statistical power. Under realistic deviations from the assumption of equal variances the classic *F*-test can yield severely biased results and lead to invalid statistical inferences. We examine two common alternatives to the *F*-test, namely the Welch’s ANOVA (*W*-test) and the Brown-Forsythe test (*F\*-*test). Our simulations show that the *W*-test is a better alternative and we therefore recommend using the *W*-test by default when comparing means. We provide a detailed example explaining how to perform the *W*-test in SPSS and R. We summarize our conclusions in five practical recommendations that researchers can use to improve their statistical practices.

When comparing independent groups researchers often analyze the means by performing a classical Analysis of Variance (ANOVA) *F-*test (Erceg-Hurn & Mirosevich, 2008). The *F*-test relies on the assumptions that the data are sampled from a normal distribution, and that the data are sampled from distributions that have equal variances (or homoscedasticity, see Lix, Keselman, & Keselman, 1996). While a deviation from the normality assumption still yields quite robust conclusions when using *F*-test (Glass, Peckham, & Sanders, 1972), unequal variances rapidly become problematic (Grissom, 2000). Yet, researchers rarely provide information about these assumptions when they report an *F*-test. When we examined statistical tests reported in 116 articles in the *Journal of Personality and Social Psychology* published in the year 2016, fourteen percent of these articles reported a One-Way *F*-test, but only one article indicated to take the homogeneity of variances assumption into account by reporting corrected degrees of freedom for unequal variances, that could signal the use of the *W*-test instead of the classical *F*-test. A similar investigation (Hoekstra, Kiers, & Johnson, 2012) yielded similar conclusions about the lack of attention to both the homoscedasticity and the normality assumptions. Despite the fact that the *F-*test is currently used by default, alternatives exist that are often a better choice, such as the Welch’s W ANOVA (*W*-test), the Alexander-Govern test, James second order test and the Brown-Forsythe ANOVA (*F\**-test). In this paper, we will discuss the pertinence of the assumptions of parametric tests, and provide simulations comparing *F-*test with the most adequate alternatives. As we argue in this article, the *W*-test has nearly the same statistical power than the *F*-test when variances are unequal, but provides better Type 1 error control. Since the *W*-test is available in practically all statistical software packages, researchers can improve their statistical inferences by replacing the *F*-test by the *W*-test.

# Why Should you Think about the Assumptions Underlying Parametric Tests?

When the assumptions of parametric tests are not met, a bias in Type I (α) or Type II (β) error may occur (Leys & Schumann, 2010, Lix et al., 1996). Type I error is the probability to falsely reject the null hypothesis. Type II error is the probability to falsely accept the null hypothesis. The statistical power (1-β) of a test is the probability of correctly rejecting the null hypothesis.

To check the assumptions underlying the *F*-test a two-step procedure is recommended in many textbooks (Field, 2013; Howell, 2012). In the first step, researchers are recommended to statistically and/or visually examine the normality and equality of variances assumptions. In the second step, researchers should choose the best statistical test (see Delacre et al., 2017). However, this two-step procedure is known to be problematic (Rasch, Kubinger, & Moder, 2011; Ruxton, 2006; Zimmerman, 2004; Wilcox, Granger, & Clark, 2013).

To statistically examine deviations from normality, the Kolmogorov-Smirnov and the Shapiro-Wilk tests are usually suggested. The first is the most widely used test but suffers from a drastic lack of power, and therefore is not recommended (see Supplemental Material 1, and Ghasemi & Zahediasl, 2012; Thode, 2002; Wilcox, 2005). The Shapiro-Wilk test is a more powerful alternative (Ghasemi & Zahediasl, 2012; Supplemental Material 1). Even so, there are still two complementary limitations of using a two-step procedure when examining the normality assumption. When sample sizes are small, all tests have low power, which is problematic since departures from normality have especially severe consequences for small sample sizes (Supplemental Material 2 and 3). When samples are large, all tests are powerful but departures from normality do not strongly affect the robustness of the *F*-test (and thus the assumption check would lead one to reject parametric tests when they perform adequately). One recommendation to cope with these limitations is to combine the Shapiro-Wilk test with graphical methods (for more information, see Ghasemi & Zahediasl, 2012; Öztuna, Elhan, & Tüccar, 2006).

The two-step procedure is equally problematic when statistically testing the equality of variances assumption. Tests assessing the equality of variances (such as Levene’s test) lack power when samples sizes are small. This is problematic given that even small differences can inflate error rates in the *F*-test and Student’s *t*-tests (Delacre et al., 2017). To conclude, assumption tests are at best a very limited approach to deciding whether or not a statistical test that relies on the normality and equal variances assumptions should be performed. This is especially problematic given that, as we will argue in the next section, the normality and equal variances assumptions are often not realistic.

## **Is the Normality Assumption Realistic?**

It has been argued that there are many fields in psychology where the assumption of normality does not hold (Cain, Zhang, & Yuan, 2016). Data can depart from normality either on one or both indicators of skewness (a measure of asymmetry of the shape of the distribution) and kurtosis (a distribution with positive kurtosis will be more peaked and have heavier tails than the normal distribution. while a distribution with negative kurtosis will be flatter and have lighter tails than the normal distribution). According to Wilcox (2005), in social and behavioral sciences, distributions can be very similar to the normal curve but with thicker tails (such as a mixed-normal distribution). These conclusions are consistent with Micceri (1989) and Yuan, Bentler and Chan (2004) showing departures from normality in various data sets.

For example, in the field of psychology, when assessing a wellness score for the general population, data may be sampled from a left-skewed distribution, because most people are probably not depressed (see Heun, Burkart, Maier, & Bech, 1999). Some examples can also be found in the field of neurosciences. When studying reaction times, data are often sampled from right-skewed distributions because it is relatively uncommon to observe slow response times (Cain et al., 2016; Palmer, Horowitz, Torralba, & Wolfe, 2011; Van Zandt, 2000). In sum, there are many common situations in which normally distributed data is an unlikely assumption.

## **Is the Homogeneity of Variances Assumption Realistic?**

Equality of variances (or homoscedasticity) is a mathematical demand that is ecologically very unlikely (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). In a previous paper (Delacre et al., 2017), we identified three different causes of heteroscedasticity: the variability inherent to the use of measured variables, the variability induced by quasi-experimental treatments on measured variables, and the variability induced by different experimental treatments on randomly assigned subjects. One additional source of variability is the presence of unidentified moderators (Cohen, Cohen, West, & Aiken, 2003).

First, psychologists, as many scholars from various fields in human sciences, often use measured variables (e.g. age, gender, educational level, ethnic origin, depression level, etc.) instead of random assignment to conditions. Prior to any treatment, parameters of pre-existing groups can vary largely from one population to another, as suggested by Henrich, Heine and Norenzayan (2010). For example, Green, Deschamps and Paez (2005)have shown that the scores of competitiveness, self*-*reliance and interdependence are more variable in some ethnic groups than in others. This stands true for many pre-existing groups such as gender, cultures, or religions and for various outcomes (see for example Adams, Van de Vijver, De Bruin, & Bueno Torres, 2014; Beilmann, Mayer, Kasearu, & Realo, 2014; Church et al., 2012; Cohen & Hill, 2007; Haar, Russo, Suñe, & Ollier-Malaterre, 2014; Montoya & Briggs, 2013). Moreover, groups are sometimes defined with the intention to have different variabilities. For example, as soon as a selective school admits its students based on the results of aptitude tests, the variability will be smaller compared to a school that accepts all students. In this example, the goal is not to alter the variability but is an inherent statistical implication of such theoretical positions.

Second, a quasi-experimental treatment can have different impacts on variances between pre-existing groups, that can even be of theoretical interest. For example, in the field of linguistics and social psychology, Wasserman and Weseley (2009) investigated the impact of language gender structure on sexist attitudes of women and men. They tested differences between sexist attitude scores of subjects who read a text in English (i.e. a language without grammatical gender) or in Spanish (i.e. a language with grammatical gender). The results showed that (for a reason not explained by the authors), the women’s score on the sexism dimension was more variable when the text was read in Spanish than in English (*SD*spanish=.80 > *SD*english=.50). For men, the reverse was true (*SD*spanish=.97 < *SD*english=1.33)[[1]](#endnote-1).

Third, even when the variances of groups are the same before treatment (due to a complete randomization in the group assignment), unequal variances can emerge later, as a consequence of an experimental treatment (Bryk & Raudenbush, 1988; Cumming, 2013; Erceg-Hurn & Mirosevich, 2008; Keppel & Wickens, 2004). For example, Koeser & Sczesny (2014) have compared arguments advocating either masculine generic or gender-fair language with control messages in order to test the impact of these conditions on the use of gender-fair wording (measured as a frequency). They report that the standard deviations increase after treatment in all experimental conditions.

Fourth, more often than not, psychological processes are captured in situations where many variables are unidentified and/or left uncontrolled (Cohen et al., 2013). Since some of these variables can act as moderators, they can generate heteroscedasticity. To conclude, there are many common situations in which the homogeneity of variances assumption is unlikely to be true. Therefore, it is very useful to have statistical tools that do not rely on the equality of variances assumption when aiming to compare groups based on their means.

# The Mathematical Differences Between the *F*-test, *W*-test, and *F*\*-test

In this section, we will explain the mathematical differences in how the *F*-test, *W*-test and *F\**-test are computed, with a focus on the way standard deviations are pooled across groups to stress the implications on heteroscedasticity (see appendix for a numerical example).

As shown in formula 1, The *F* statistic is calculated by dividing the inter-group variance by a pooled error term, where and *n*j are respectively the variance estimates and the sample sizes from each independent group, and where *k* is the number of independent groups:

|  |  |  |
| --- | --- | --- |
|  | *F =* | (1) |

The degrees of freedom in the numerator (formula 2) and in the denominator (formula 3) of the *F*-test are computed as follows:

|  |  |  |
| --- | --- | --- |
|  | *dfn = k-*1 | (2) |
|  | *dfd = N-k, where N =* | (3) |

As a generalization of the Student’s *t*-test, the *F*-test is calculated based on a pooled error term, which implies that all samples are estimates of a common population variance. The *F*-test suffers from the same limitations as the Student’s *t*-test when sample sizes are unequal between groups, in that the Type I error rate is no longer controlled at the desired level when variances are unequal between groups. When the larger variance is associated with the larger sample size, there is a decrease in the Type I error rate (Nimon, 2012; Overall, Atlas, & Gibson, 1995), because the error term increases, and therefore, the *F-*value decreases, leading to fewer significant findings than expected with a specific Type I error level. When the larger variance is associated with the smaller sample size, the Type I error rate is inflated (Nimon, 2012; Overall et al., 1995). This inflation is caused by the under evaluation of the error term, which increases the *F-*value, and thus leads to more significant results than expected, based on the nominal Type I error level. Moreover, when the number of groups increases, the *F-*test becomes increasingly liberal as soon as the variances of the distributions in each group are not similar, even when sample sizes are equal between groups.

To address the problems with error control in the *F*-test when variances are unequal, several authors have proposed alternative approaches to statistical tests on more than two means, which do not rely on the homogeneity of variances assumption (e.g., Welch, 1951). Tomarken and Serlin (1986) have shown that from the available alternatives, the *F\**-test and *W*-test are the best choices. Both tests are available in SPSS, which is a widely used software in psychological sciences (Hoekstra et al., 2012). The *F\** statistic proposed by Brown and Forsythe (1974) is computed as follows:

|  |  |  |
| --- | --- | --- |
|  | *F\* =* | (4) |

Where xj and are respectively the group mean and the group variance, and is the overall mean.

As can be seen in formula 4 the numerator of the *F\** statistic is equal to the sum of squares between groups (which is equal to the numerator of the *F* statistic when one compares two groups). In the denominator of the statistic, the variance of each group is weighted by 1 minus the relative frequency of each group, so that the variance associated with the group with the smallest sample size is given more weight. As a result, when the larger variance is associated with the larger sample size, *F\** is larger than *F*, because the denominator decreases, leading to more significant findings compared with the *F*-test. On the other hand, when the larger variance is associated with the smaller sample size, *F\** is smaller than *F*, because the denominator increases, leading to fewer significant findings than expected with the *F*-test. The degrees of freedom in the numerator and in the denominator of *F\**-test are computed as follows:

|  |  |  |
| --- | --- | --- |
|  | *dfn = k-*1 | (5) |
|  | *dfd=* | (6) |

Formula 7 provides the *W*-test computation. The squared deviation between group means and the general mean are weighted by instead of *n*j in the numerator of the *W*-test (Brown & Forsythe, 1974).

|  |  |  |  |
| --- | --- | --- | --- |
|  | *W =* , | | (7) |
| where | *,* |

The degrees of freedom of the *W*-test are approximated as follows:

|  |  |  |
| --- | --- | --- |
|  | *dfn =* | (8) |
|  | *dfd =* | (9) |

When there are only two groups to compare, the *F\**-test and *W*-test are identical (i.e., they have exactly the same statistical value, degrees of freedom and significance). However, when there are more than two groups to compare, the tests differ. To better understand how to compute all statistics, a set of fictional raw data simulating the example of a three-group design is available in the appendix. The following section will present Monte Carlo simulations assessing these three tests on both Types I and II error rates.

# Monte Carlo simulations: *F*-test vs. *W*-test vs. *F\**-test

## We chose to examine the Type I error rate and the statistical power of the *F-*test, *W*-test and *F\*-*test. The James second-order test and the Alexander-Govern’s test were not included, because they yield very similar results to the *W*-test but are less readily available in statistical software packages. For a more extended description of these two alternatives, see Schneider and Penfield (1997).

We conducted simulations for 2560 scenarios that we deemed most relevant. For different distributions underlying the data, k-1 samples (where k varied from 2 to 5) were generated from a population where and sample sizes ( were 20, 30, 40, 50 or 100. The *SD* and the sample size of the last group was a function of the sample sizes ratio (*n*-ratio = ; ranging from 0.5 to 2 in steps of 0.5) and the *SD*-ratio (ranging from .5 to 4). This setup resulted in a wide range of conditions in which the normality assumption was met or not, and where the homoscedasticity assumption was met or not, as summarized in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1.  *Number of conditions in the simulation where the normality and homoscedasticity assumptions were met or not* | | | |
|  | Distributions | | |
|  | *Normal* |  | *Other* |
| Homoscedasticity | 80 |  | 560 |
| Heteroscedasticity | 240 |  | 1680 |

## **Type I Error Rate of the *F*-test vs. *W*-test vs. *F\**-test**

**Simulating error rates when the normality assumption is met.** We simulated 1,000,000 data sets under the null hypothesis (all means are equal across groups) and computed the *p*-value distribution across the three tests under 320 scenarios where the normality assumption is met (see Table 1). When the null-hypothesis is true, the *p*-value distribution should be uniform, and the Type I error rate is supposed to be equal to the alpha level (here 5%). As explained above, when comparing two groups, the *W*-test and *F\**-test are mathematically identical and should yield identical error rates.

We classified the 320 scenarios where the normality assumption is met into five categories. Previous findings highlighted differences in terms of the Type I error rate of the *F*-test as a function of the correlation between sample sizes and *SD*s (see Nimon, 2012; Overall et al., 1995). Therefore, scenarios where variances are unequal between groups were divided into three categories: 1) sample sizes are unequal and there is a positive correlation between sample sizes and *SD*s (i.e. the group with the biggest sample size has the biggest *SD*), 2) sample sizes are unequal and there is a negative correlation between sample sizes and *SD*s or 3) sample sizes are equal across all groups. In the last two categories of simulations, 4) variances are equal between groups and sample sizes are unequal or 5) variances are equal between groups and sample sizes are equal. For each category, the sample size of the first group was varied from 20 to 100 and the number of groups in the ANOVA varied from 2 to 5. Table 2 reports the main conclusions based on a sample size of 30 in the first group (which seemed ecological and didactic) and provides Type 1 error rates for ANOVAs with two and three groups (for results from additional scenarios, please see the Supplemental Material). Because the Type I error rates were not normally distributed in each category, we report the median and the Median Absolute Deviation (MAD), a more robust measure of central tendency and dispersion than the mean and *SD* (Leys, Ley, Klein, Bernard, & Licata, 2013). There is no measure of dispersion for the fifth category because there was only one condition where both equal variances, equal sample sizes, and normality were met, when there are 30 subjects in the first group.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | Table 2.  *Type I error rate of the F-test, W-test and F\*-test, for five categories of conditions when we compare two or three groups extracted from normal distributions, when there are 30 subjects in the first group* |  | | |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | | Two groups | | | |  | | Three groups | | | | | **Category** | *F* | | *F\** | *W* |  | | *F* | | *F\** | *W* | | 1 | **0.018***(0.003)* | | **0.050***(<0.001)* | **0.050***(<0.001)* |  | | **0.034***(0.001)* | | **0.058***(0.002)* | **0.050***(<0.001)* | | *2* | **0.112***(0.015)* | | **0.051***(<0.001)* | **0.051***(<0.001)* |  | | **0.112***(0.022)* | | **0.056***(0.004)* | **0.050***(<0.001)* | | *3* | **0.052***(<0.001)* | | **0.050***(<0.001)* | **0.050***(<0.001)* |  | | **0.061***(0.006)* | | **0.059***(0.006)* | **0.050***(<0.001)* | | *4* | **0.050***(<0.001)* | | **0.050***(<0.001)* | **0.050***(<0.001)* |  | | **0.050***(<0.001)* | | **0.050***(<0.001)* | **0.050***(<0.001)* | | *5* | **0.050** | | **0.050** | **0.050** |  | | **0.050** | | **0.050** | **0.050** | | | | |  | *Note*. The median Type I error rates for the *F*-test, *W*-test and *F*\*-test, and associated MAD (in brackets) are compared. In the first three categories, variances are unequal (*SD*-ratio ≠ 1) and sample sizes are respectively unequal between groups with a positive correlation between *ns* and *SD*s (Category 1), unequal between groups with a negative correlation between *ns* and *SD*s (Category 2), or equal (category 3). In the last two categories, variances are equal (*SD*-ratio = 1) and sample sizes are either unequal (category 4) or equal (category 5). |  | |

Table 2 shows that, concerning the Type I error, in all cases, the *W*-test performs as well as or better than the two alternatives. More specifically, as soon as variances are unequal between groups (categories 1 to 3), the *W*-test stays unaltered whereas the two other tests become unreliable.

These results can be generalized: when the number of groups increases, the *F*-test becomes increasingly unreliable. The Type I error rate is too low when there is a positive correlation between sample sizes and population standard deviations, but too high when there is either a negative correlation between sample sizes and population standard deviations or unequal variances with equal sample sizes[[2]](#endnote-2). The *F\**-test is robust against unequal population variances when there are two groups to compare (Table 2). When there are three groups to compare, the test is less affected by violations of the assumption of equal variances than the *F*-test, but the Type I error rate still increases when there are unequal population variances between groups. Additional simulations, presented in the Supplemental Material, show that the test gets more liberal as the sample sizes are smaller, and as the *SD*-ratio and the number of groups to compare increase. Finally, the *W*-test yields a more stable Type I error rate, regardless of the number of groups that are compared, and regardless of the *SD*-ratio.

**Simulating Type I error rates when the normality assumption is not met.** We tested the impact of non-normal distributions without heterogeneity of variances on the Type I error rate, based on560 scenarios (See Table 1). As an illustration, Table 3 contains information about the median Type I error rate and associated MAD of categories where we compare two or three groups of different sample sizes, extracted from 1) mixed normal distributions (symmetric with heavy tails distributions), 2) normal right-skewed distributions or 3) chi-squared distributions, when there are 30 subjects in the first group (for more scenarios, see the Supplemental Material). Table 3 shows that across the three categories, the *F*-test provides a better control of the Type I error rate than both *W*-test and *F\**-test. When there are two groups to compare, the *F\**-test and *W*-test have the same Type I error rate since, as stated above, both tests are identical. However, when there are three groups to compare, the *W*-test is more conservative than the *F*-test and *F\**-test with heavy tailed distributions, and more liberal than the *F*-test and *F\**-test with skewed distributions. Lastly, we underline that, in these scenarios, departures from normality have mild consequences.

|  |  |  |
| --- | --- | --- |
|  | Table 3.  *Type I error rate of the F-test, W-test and F\*-test, for three categories of conditions when we compare two or three groups extracted from non-normal distributions with equal variances and unequal sample sizes, when there are 30 subjects in the first group* |  |
| |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | | Two groups | | | |  | | Three groups | | | | | **Category** | *F* | | *F\** | *W* |  | | *F* | | *F\** | *W* | | 1 | **0.048***(<0.001)* | | **0.046***(<0.001)* | **0.046***(<0.001)* |  | | **0.046***(<0.001)* | | **0.045***(<0.001)* | **0.042***(<0.001)* | | *2* | **0.050***(<0.001)* | | **0.051***(0.001)* | **0.051***(0.001)* |  | | **0.049***(0.000)* | | **0.049***(<0.001)* | **0.053***(0.001)* | | *3* | **0.054***(0.002)* | | **0.058***(0.001)* | **0.058***(0.001)* |  | | **0.052***(0.001)* | | **0.052***(<0.001)* | **0.064***(<0.001)* | | | |
|  | *Note*. The median Type I error rates for the *F*-test, *W*-test and *F*\*-test, and associated MAD (in brackets) are compared into three categories of conditions where variances are equal (*SD*-ratio = 1) and sample sizes are unequal between groups. In the first two categories, groups are respectively extracted from mixed normal distributions (Category 1) or normal right-skewed distributions (Category 2). In the third category, k-1 samples are extracted from chi square distributions and the last sample from a normal left-skewed distribution (category 3). |  |

### In general, while the *W*-test is more robust than both the *F*-test and *F\**-test when distributions are normal and groups have unequal variances, it is less robust than the two other tests when the normality assumption is not met but groups have equal variances (Supplemental Material 2). The *W*-test is more affected by heavy-tailed and skewed distributions than the *F*-test, becoming slightly more conservative with heavy-tailed distributions, and more liberal with skewed distributions. Furthermore, the *W*-test becomes more liberal when highly skewed distributions are combined with unequal variances, particularly when sample sizes are unequal between groups, but it is still less liberal than the *F*-test (see Table 4).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4.  *Comparison of Type I error rate of the F-test, W-test and F\*-test when we compare two or three groups extracted from highly skewed distributions, when there are 30 subjects in the first group*   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | | Two groups | | | |  | Three groups | | | | | **Category** | *F* | | *F\** | *W* |  | *F* | | *F\** | *W* | | 1 | **0,120***(0,014)* | | **0,064***(0,003)* | **0,064***(0,003)* |  | **0,116***(0,023)* | | **0,062***(0,007)* | **0,070***(0,008)* | | *2* | **0.060***(0.003)* | | **0.056***(0.001)* | **0.056***(0.001)* |  | **0.066***(0.008)* | | **0.063***(0.007)* | **0.058***(0.002)* | | *Note*. The median Type I error rates for the *F*-test, *W*-test and *F*\*-test, and associated MAD, are compared between two categories where k-1 samples are extracted from chi-square distributions and the last sample from a normal left-skewed distribution. In both categories, variances are unequal and sample sizes are unequal (with negative correlation between *ns* and *SDs*; category 1) or equal (category 2). | | | | | | | | | | |

### In sum, taking all our simulations[[3]](#endnote-3) into account, we can draw the following conclusions for the Type I error rate:

### 1) With departure from normality but homoscedasticity, the *F*-test remains acceptable with at least 20 subjects per group, following Bradley (1978)’s acceptable interval of an error rate between 025 to .075 when the alpha is set to 0.05.

### 2) As soon as there is heteroscedasticity, regardless of normality, the Type I error rate of the *F*-test and *F\**-test will commonly fall outside of Bradley’s interval, even with big sample sizes. The Type I error rate of the *W*-test will generally fall inside of the Bradley’s interval as long as there are at least 20 subjects per groups, except when distributions are highly skewed, as discussed in the next point.

### 3) For highly skewed distributions, which are easily detectable based on a Shapiro-Wilk test, the *W*-test is the best choice, but one needs to collect at least 50 subjects per group to guarantee a Type 1 error rate that falls within Bradley’s interval. With less than 50 subjects per group, other options such as non-parametric approaches (see, e.g., Leys & Schumann, 2010) are advised.

## **Statistical power for the *F*-test, *W*-test, and *F\**-test**

In addition to the Type 1 error rate, the statistical power of a test is important to consider. The power of a test is a function of the effect size, the sample size, and the nominal alpha level (Cohen, 1988). We performed simulations in which we introduced a true effect (the mean = 1 in the last group, mean = 0 in all other groups). We again manipulated the distributions, variances and *SD*-ratios of the populations from which groups are extracted as well as sample sizes and sample ratios, as was done when examining the Type 1 error rate. Based on these results, we computed the relative power difference of the *W*-test and *F\**-test in comparison with the *F*-test within categories, as follows:

|  |  |
| --- | --- |
| Relative difference of power of the *W*-test = | (9) |
| Relative difference of power of the *F\**-test = | (10) |

Formulas (9) and (10) will return a positive value when the *W*-test (or *F\**-test) is more powerful than the *F*-test, and they will return a negative value when the *W*-test (or *F\**-test) is less powerful than the *F*-test.

First, it is often argued that the *W*-test and *F\**-test are less powerful than the *F*-test when the assumptions of normality and equal variances are met. However, our simulations show that under these assumptions, the loss of power is always smaller than 3% of the power of the *F*-test. Moreover, the relative differences in power between the *F*-test and both *W*-test and *F\**-test tend towards zero when the number of subjects per group increases (Table 5).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5.  *Relative difference of power of the W-test and F\*-test, in comparison with the F-test, when we compare two or three groups extracted from normal distributions with equal variances and unequal sample sizes, as a function of the first sample size* | | | | | | | |
|  | | Two groups | | | Three groups | | |
| **n1** | *F\** | | *W* |  | | *F\** | *W* |
| 20 | **-1.6%***(0.90%)* | | **-1.6%***(0.90%)* |  | | **-1.2%***(0.50%)* | **-2.6%***(0.30%)* |
| *30* | **-1.0%***(0.60%)* | | **-1.0%***(0.60%)* |  | | *-***0.6%***(0.30%)* | **-1.3%***(<0.10%)* |
| *40* | **-0.6%***(0.30%)* | | **-0.6%***(0.30%)* |  | | *-***0.4%***(0.20%)* | **-0.8%***(0.10%)* |
| *50* | **-0.4%***(0.20%)* | | **-0.4%***(0.20%)* |  | | *-***0.1%***(<0.10%)* | **-0.5%***(0.10%)* |
| *100* | **-0.0%***(<0.10%)* | | **-0.0%***(<0.10%)* |  | | **-0.0%***(<0.10%)* | **-0.0%***(<0.10%)* |
| *Note*. The median difference of power (in %) of *W*-test and *F*\*-test (in comparison with the *F*-test), and associated MAD (in brackets) are compared when both assumptions of normality and equal variances are met. | | | | | | | |

Second, when the assumption of normality is met but variances are unequal, since the *F*-test becomes either too liberal or too conservative, depending on the correlation between *ns* and *SD*s, the *W*-test is sometimes more powerful, sometimes less powerful than the *F*-test (Table 6). However, the *W*-test is preferable since (a) it is more stable; and (b) its power is closer than the expected power (computed using the Welch’s power curve, Minitab assistance, *n.d*).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 6.  *Relative difference of power of the W-test and F\*-test, in comparison with the F-test, for three categories of conditions when we compare two or three groups extracted from normal distributions, when there are 30 subjects in the first group* | | | | | | | | |
|  | Two groups | | | | | Three groups | | |
| **Category** | | *F\** | *W* | |  | | *F\** | *W* |
| 1 | | **70.9%***(28.9%)* | | **70.9%***(28.9%)* |  | | **57.7%***(20.2%)* | **18.8%***(23.4%)* |
| *2* | | **-29.3%***(16.7%)* | | **-29.3%***(16.7%)* |  | | **-24.8%***(15.4%)* | **-29.3%***(29.4%)* |
| *3* | | **-1.8%***(0.9%)* | | **-1.8%***(0.9%)* |  | | **-2.2%***(1.5%)* | **-30%***(10.6%)* |
| *Note*. The median difference of power (in %) of *W*-test and *F*\*-test (in comparison with the *F*-test), and associated MAD (in brackets) are compared when population distributions are normal and variances are unequal (*SD*-ratio ≠ 1) and either sample sizes are unequal between groups with a positive correlation between *ns* and *SD*s (Category 1), or unequal between groups with a negative correlation between *ns* and *SD*s (Category 2), or equal (category 3). | | | | | | | | |

Third, when the assumption of normality is not met but there is homoscedasticity, the *W*-test tends to be a little more powerful than the *F*-test, especially when distributions are highly skewed (Table 7). This is due to the fact that the *W*-test is more liberal. Therefore, whenever researchers have good reasons to think that variances are equal between groups and given that the departure from normality does not impact the *F*-test too much, the classical *F*-test would be a better choice in these circumstances. However, since we argue that (a) heteroscedasticity is often present but can be difficult to detect and (b) the *W*-test and *F*-test have very similar power when homoscedasticity is met, we still advice to use the *W*-test.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 7.  *Relative difference of power of the W-test and F\*-test, in comparison with the F-test, for three categories of conditions when we compare two or three groups extracted from normal distributions that have equal variances and unequal sample sizes, when there are 30 subjects in the first group* | | | | | | | |
|  | Two groups | | | | Three groups | | |
| **Category** | | *F\** | *W* |  | | *F\** | *W* |
| 1 | | **-1%***(0.6%)* | **-1%***(0.6%)* |  | | **-0.6%***(0.3%)* | **-1.3%***(<0.1%)* |
| *2* | | **-0.4%***(0.1%)* | **-0.4%***(0.1%)* |  | | **-0.3%***(0.1%)* | **+2.3%***(0.5%)* |
| *3* | | **-0.5%***(0.7%)* | **-0.5%***(0.7%)* |  | | **-0.6%***(0.4%)* | **+6.3%***(0.8%)* |
| *Note*. The median difference of power (in %) of *W*-test and *F*\*-test (in comparison with the *F*-test), and associated MAD (in brackets) are compared when population variances are equal (*SD*-ratio = 1), sample sizes are equal between groups and sample are extracted from normal distribution (Category 1), k-1 samples are extracted from normal right-skewed distributions and one sample is extracted from a normal left-skewed distribution (Category 2), or k-1 samples are extracted from chi-squared distributions and one sample is extracted from a normal left-skewed distribution (category 3). | | | | | | | |

Note that globally, when the assumption of homoscedasticity is met but distributions are not normal, and more specifically with heavy tailed distributions, we observed results that are in contrast with previous work (Wilcox, 1998) for all tests comparing means. Wilcox (1998) concluded that there is a loss of power when comparing means from heavy-tailed distributions (e.g. double exponential or a mixed normal distribution) when compared to normal distributions. This finding is based on the argument that heavy-tailed distributions are associated with bigger standard deviations than normal distributions, and that the effect size for such distributions is therefore smaller (Wilcox, 2011). However, it is important to avoid the confusion between kurtosis and the standard deviation. DeCarlo (1997) explains that kurtosis and *SD* are completely independent, meaning that one can find distributions that have similar *SD* but different kurtosis. We ran simulations manipulating kurtosis while keeping the *SD* unaltered (by comparing a normal distribution with a mixed and a double-exponential distribution that have different kurtosis, but the same *SD*). Results show that when heavy-tailed distributions have equal standard deviations and *SD*-ratios as normal distributions, there are no substantial differences in power as a function of the kurtosis of the underlying distribution (see Supplemental Material 3).

Finally, when both the normality and equal variances assumptions are not met, the *F*-test is never reliable. However, the *W*-test is not very reliable either (see Supplemental Material 3). Indeed, the *W*-test will be more powerful than the *F*-test, but again, just because it is too liberal (see the conclusions in the Type I error rate section). Therefore, under these precise circumstances, we recommend researchers switch to non-parametric alternatives (Leys & Schumann, 2010).

***Figures 1 and 2***

# Recommendations

In sum, we provide five recommendations:

1) Use the *W*-test instead of the *F*-test to compare groups means. The *F*-test and *F\**-test should be avoided, because the equal variances assumption is often unrealistic, tests of the equal variances assumption will often fail to detect differences when these are present, the loss of power when using the *W*-test is very small (and often even negligible), and the gain in Type I error control is considerable under a wide range of realistic conditions.

2) Do not neglect the descriptive analysis of the data. A complete description of the shape and characteristics of the data (e.g. histograms and boxplots) is important. When at least one statistical parameter relating to the shape of the distribution (e.g. variance, skewness, kurtosis) seems to vary between groups, comparing results of the *W*-test with results of a nonparametric procedure is useful in order to better understand the data.

3) Use the Shapiro-Wilk test to detect departures from normality (combined with graphical methods). Contrary to the Kolmogorov-Smirnov test, the Shapiro-Wilk test will almost always detect distributions with high skewness, even with very small sample sizes. With small sample sizes, the *W*-test will not control Type I error rate when skewness is present and detecting departures for normality is therefore especially important in small samples. When comparing at most four groups, the *W*-test should be avoided if the Shapiro-Wilk test reject the normality assumption, with less than 50 observations per group. When comparing more than four groups, the *W*-test should be avoided if the Shapiro-Wilk test rejects the normality assumption, with less than 100 subjects per group When normality cannot be assumed because of high kurtosis or high skewness, we recommend the use of alternative tests that are not based on means comparison, such as the trimmed means test[[4]](#endnote-4) or nonparametric tests. For more information, see Erceg-Hurn and Mirosevich (2008).

4) Perform a-priori power-analyses. Fifty subjects per group are generally enough to control the Type I error rate, but power analyses are still important in order to determine the required sample sizes to achieve sufficient power to detect a statistically significant difference (see Albers & Lakens, 2018).

5) Use balanced designs (i.e. same sample size in each group) whenever possible. When using the *W*-test, the Type I error rate is a function of criteria such as the skewness of the distributions, and whether skewness is combined with unequal variances and unequal sample sizes between groups. Our simulations show that the Type I error rate control is in general slightly better with balanced designs.

# References

Cohen, J and Cohen, P and West, S G and Aiken, L S (2013). *Applied multiple regression/correlation analysis for the behavioural sciences*. Mahwah, NJ: Erlbaum.

Cumming, G (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY : Routledge.

Field, A (2013). *Discovering statistics using IBM SPSS statistics*. London, UK: Sage.

Ghasemi, A and Zahediasl, S (2012). Normality tests for statistical analysis: A guide for non-Statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486‑489. [doi:10.5812/ijem.3505](https://doi.org/10.5812/ijem.3505)

Gibbons, J D and Chakraborti, S (2011). *Nonparametric statistical inference*. Boca Raton, FL: Chapman and Hall/CRC.

Haar, J M and Russo, M and Suñe, A and Ollier-Malaterre, A (2014). Outcomes of work–life balance on job satisfaction, life satisfaction and mental health: A study across seven cultures. *Journal of Vocational Behavior*, *85*(3), 361–373. [doi:10.1016/j.jvb.2014.08.010](http://dx.doi.org/10.1016/j.jvb.2014.08.010)

Keppel, G and Wickens, T D (2004). *Design and analysis: A researcher’s handbook*. [Upper Saddle River](https://fr.wikipedia.org/wiki/Upper_Saddle_River) , [New Jersey](https://fr.wikipedia.org/wiki/New_Jersey): Prentice Hall.

Liu, H (2015). Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance. Virginia Commonwealth University.

Minitab assistance (n.d). One-Way ANOVA. Retrieved from http://support.minitab.com/en-us/minitab/17/Assistant\_One\_Way\_ANOVA.pdf.

Nachar, N (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13–20. [doi:10.20982/tqmp.04.1.p013](http://dx.doi.org/10.20982/tqmp.04.1.p013)

Palmer, E M and Horowitz, T S and Torralba, A and Wolfe, J M (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58-71. [doi:10.1037/a0020747](http://dx.doi.org/10.1037/a0020747.supp)

Rasch, D and Kubinger, K D and Moder, K (2011). The two-sample t test: Pre-testing its assumptions does not pay off. *Statistical papers*, *52*(1), 219–231. [doi:10.1007/s00362-009-0224-x](http://dx.doi.org/10.1007/s00362-009-0224-x)

Ruxton, G D (2006). The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test. *Behavioral Ecology*, *17*(4), 688–690. [doi:10.1093/beheco/ark016](http://dx.doi.org/10.1093/beheco/ark016)

Schneider, P and Penfield, D (1997). Alexander and Govern’s approximations: Providing an alternative to ANOVA under variance heterogeneity. *The Journal of Experimental Education, 65*(3). 271-286.

Simmons, J P and Nelson, L D and Simonsohn, U (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366, doi:10.1177/0956797611417632

Thode, H C (2002). *Testing for normality*. New York, NY: Marcel Dekker.

Tomarken, A J and Serlin, R C (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90–99. doi:10.1037//0033-2909.99.1.90

Van Zandt, T (2000). How to fit a response time distribution. *Psychonomic bulletin & review*, *7*(3), 424–465. [doi:10.3758/bf03214357](http://dx.doi.org/10.3758/bf03214357)

Wasserman, B D and Weseley, A J (2009). ¿Qué? Quoi? Do languages with grammatical gender promote sexist attitudes? *Sex Roles*, *61*(9‑10), 634-643. [doi:10.1007/s11199-009-9696-3](http://dx.doi.org/10.1007/s11199-009-9696-3)

Welch, B L (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3/4), 330–336. [doi:10.2307/2332579](http://dx.doi.org/10.2307/2332579)

Wilcox, R R (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300–314. [doi:10.1037/0003-066x.53.3.300](http://dx.doi.org/10.1037/0003-066x.53.3.300)

Wilcox, R R (2005). *Introduction to Robust Estimation and Hypothesis Testing*. [Cambridge, Massachusetts](https://en.wikipedia.org/wiki/Cambridge,_Massachusetts), US: Elsevier Academic Press.

Wilcox, R R and Granger, D A and Clark, F (2013). Modern robust statistical methods: Basics with illustrations using psychobiological data. *Universal Journal of Psychology*, *1*(2), 21–31.

Yuan, K H and Bentler, P M and Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. [doi:10.1007/bf02295644](http://dx.doi.org/10.1007/bf02295644)

Zimmerman, D W (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *The Journal of general psychology*, *127*(4), 354–364. [doi:10.1080/00221300009598589](http://dx.doi.org/10.1080/00221300009598589)

Zimmerman, D W (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173–181. [doi:10.1348/000711004849222](http://dx.doi.org/10.1348/000711004849222)

**Author’s note:** First author performed simulations. First, second and fourth authors contributed to the design. All authors contributed to the writing and the review of the literature. The Supplemental Material, including the full R code for the simulations and plots can be obtained from <https://github.com/mdelacre/Welch-ANOVA>. The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

# Appendix

## **The Mathematical Development of the *F*-test, *W*-test, and *F\**-test: Numerical Example**

A summary is presented in Table A1. The complete example is available on Github. The DV is a score that can vary from 0 to 40. The IV is a three-level factor A (levels = A1, A2 and A3).

|  |  |  |  |
| --- | --- | --- | --- |
|  | A1 | A2 | A3 |
| ni | 41.00 | 21.00 | 31.00 |
|  | 24 | 23 | 27 |
|  | 81.75 | 10.075 | 38.40 |

Table A1. *Summary of the data of the fictive case*

The global mean (i.e. the mean of the global dataset) is a weighted mean of the group means:

The *F*-test statistic and degrees of freedom are computed by applying formulas 1, 2 and 3:

*F =*  2.377

*df*n = 3-1 = 2

*df*d = 93-3 = 90

The *F\**-test and his degrees of freedom are computed by applying formulas 4, 5 and 6.

*F\* =* 3.088

|  |  |
| --- | --- |
| *df*n = 3-1 = 2 | |
| *dfd =*  81,149 | |
| where | 79,11 |

Finally, the *W*-test and his degrees of freedom are computed in applying formulas 7, 8 and 9:

|  |  |
| --- | --- |
| *W =*  4.606 | |
| Where | *w =* 3,39  24,10 |
| *dfn =* 3-1 = 2 | | |
| *dfd = =* 59,32 | | |

One should notice that in this example, the biggest sample size has the biggest variance. As previously mentioned, it means that the *F*-test will be too conservative, because the *F* value decreases. The *F*\* ANOVA will also be a little too conservative, even if the test is less affected than the *F*-test. As a consequence: *W* > *F\** > *F*.

1. Note that this example is for didactic reasons, the differences have not been tested and might not differ significantly [↑](#endnote-ref-1)
2. To yield a robust test, the Type I error rate has to be sufficiently close to the nominal 5% level. In order to assess the robustness of the three tests in our simulations, we follow Bradley (1978) and consider the Type I error rate as ‘close enough’ to the nominal 5% if it falls in the interval [0.025; 0.075]. [↑](#endnote-ref-2)
3. All data are available on the following link: https://osf.io/ru9tz/ (See Type I error rate.xlsx). [↑](#endnote-ref-3)
4. The null hypothesis of the trimmed means test assumes that trimmed means are the same between groups. A trimmed mean is a mean computed on data after removing the lowest and highest values of the distribution (Erceg-Hurn & Mirosevich, 2008). Trimmed means and means are equal when data are symmetric. On the other hand, when data are asymmetric, trimmed means and means differ. [↑](#endnote-ref-4)