# Additional file to

**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.

**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/

## **Supplemental Material 1: Detecting Abnormality with Kolmogorov-Smirnov Test, Shapiro-Wilk Test and Jarque-Bera Test**

The Kolmogorov Smirnov test is available in many packages such as R or SPSS. This test will compute the Kolmogorov distance, which is the maximum distance between observed distribution and theoretical distribution (i.e. normal distribution; Wilcox, 2005): . In SPSS, the observed distribution is automatically compared to a normal distribution whose mean and standard deviation are estimated by the sample. Moreover, the lillefors correction is applied by default (i.e. a correction on the critical values; Steinskog, Tjøstheim, & Kvamstø, 2007). In R, the observed distribution can be compared to normal distribution of any mean and standard deviation, that must be specified as an argument of the function. Moreover, the lillefors correction is not applied by default, implement it implies to use another function.

The Jarque-Bera test is based on skewness and kurtosis and is specifically designed to detect departures from the normal curve (Öztuna, Elhan, & Tüccar, 2006; Steinskog et al., 2007). The Shapiro-Wilk test, also specifically designed to detect departures from the normal curve (DeCarlo, 1997), is based on the correlation between the observed quantiles and normal quantiles (Ghasemi & Zahediasl, 2012; Öztuna et al., 2006). Both tests are known to be more powerful than the Kolmogorov-Smirnov test (regardless of the Lillefors correction).

### **Simulations.** Firstly, in order to estimate the Type I error rate of the four tests, we simulated 1,000,000 samples following a standard normal distribution (Figure A1.1), thanks to the function “rnorm” (from the package “stats”; “R: The Normal Distribution,” 2016), under five conditions, as a function of the sample size (10, 20, 30, 50 and 100[[1]](#footnote-1)), and we performed the four tests on simulated samples. Regardless of the correction, we chose to perform the Kolmogorov-Smirnov test comparing the observed distribution to a normal distribution where mean and standard deviation are estimated by the sample, as it is done in SPSS. Two steps were repeated for each condition: in a first step, the *p*-values of the four tests were extracted for each dataset, and in a second step, the percent of *p*-values under the nominal alpha risk (5%) was computed for each test and for each condition. Results are in Table S1.1.

|  |
| --- |
|  |
| *Figure A1.1*. normal curve centered around a mean of 0, and with a standard deviation of 2. |

Secondly, in order to estimate the power, we simulated 1,000,000 samples, under 25 conditions, as a function of the sample size (10, 20, 30, 50 and 100), and the distributions underlying the data, and we performed the four tests on simulated samples. As previously, regardless of the correction, we chose to perform the Kolmogorov-Smirnov test comparing the observed distribution to a normal distribution where mean and standard deviation are estimated by the sample. Two steps were repeated for each condition: in a first step, the *p*-values of the four tests were extracted for each dataset, and in a second step, the percent of *p*-values under the nominal alpha risk (5%) was computed for each test and for each condition.

In order to generate data from different distributions, we used different R commands:

### **Mixed normal distribution (mean=0, SD=1; Figure A1.2).** In order to assess the power of all tests when data are extracted from a mixed normal distribution, where P(X~N(0,2.53))= .9 and P(X~N(0,.6325))= .1, conducting to a mixed normal distribution where mean = 0 and *SD* = 1, with a kurtosis of 12.80, data were generated by means of the function “rmixnorm” (from the package “bda”; Wang & Wang, 2015). Results are in Table S1.2.

|  |
| --- |
|  |
| *Figure A1.2*. distributions where mean = 0 and *SD* = 1, as a function of the distribution underlying the data (Mixed normal distribution vs normal) |

### **Uniform distribution (mean=0, SD=1; Figure A1.3).** In order to assess the power of all tests when data are extracted from a uniform distribution, data were generated by means of the function “runif” (from the package “stats”; “R: The Uniform Distribution,” 2016). Results are in Table S1.3.

|  |
| --- |
|  |
| *Figure A1.3*. Distributions where mean = 0 and *SD* = 1, as a function of the distribution underlying the data (uniform vs normal) |

### **Normal skewed distribution with positive skewness of +0.99 (mean=0, SD=1; Figure A1.4).** In order to assess the power of all tests when data are moderately skewed, data were generated by means of the function “rsnorm” (from the package “fGarch”; “R: Skew Normal Distribution,” 2017). Results are in Table S1.4.

|  |
| --- |
|  |
| *Figure A1.4*. Distributions where mean = 0 and *SD* = 1, as a function of the distribution underlying the data (normal skewed vs normal) |

**Chi-squared distribution (mean=2, SD=1.41; Figure A1.5).** In order to assess the power of all tests when data are highly skewed, data were generated by means of the functions “rchisq” (“R: The (non-central) Chi-squared Distribution,” 2016). Results are in Table S1.5.

|  |
| --- |
|  |
| *Figure A1.5*. Distributions where mean = 1 and *SD* = 1.41, as a function of the distribution underlying the data (chi-squared vs normal) |

### **Type I error rate, as a function of the sample size.** When the lillefors correction is not applied, the Type I error rate of the Kolmogorov-Smirnov test is very close to 0, meaning that the Kolmogorov-Smirnov test will reject the null hypothesis less than 1%, regardless of the sample size, meaning that the test is too conservative. On the other side, when the lillefors correction is applied, the test has a good Type I error rate control, in the same way that the Shapiro-Wilk test. The Jarque-Bera test is also a little too conservative.

### **Power, as a Function of the Sample Size and Distribution Underlying the Data.**

**Mixed normal distribution.** When the data follow a mixed normal distribution, the power of the Kolmogorov-Smirnov test without lillefors correction barely achieves 30% when there are 100 subjects in the sample, meaning that the test will often fail to detect any departures from the normal distribution. All other tests are more powerful. With very small sample sizes (i.e. ), the most powerful test is the Shapiro-Wilk test. With at least 30 subjects per groups, the Jarque-Bera test is a little more powerful than the Shapiro-Wilk test (Table S1.3). However, for both test, 100 subjects are required to detect normality violations with a power of 95%.

**Uniform distribution.** When the data are uniformly distributed, the power of the Kolmogorov-Smirnov test without lillefors correction barely achieves 1% when there are 100 subjects in the sample, meaning that the test will practically never detect any departures from the normal distribution. The Jarque-Bera test will only detect that distributions are not normal with big samples sizes (i.e. ni=100; Table S1.3). While the Kolmogorov-Smirnov test with lillefors correction will more often detect departures from the normal assumption than both pre-mentioned tests, the Shapiro-Wilk test appear to be the most powerful alternative (Table S1.3). However, even the Shapiro-Wilk test requires 100 subjects per group in order to detect the normality violations with a power of 95%.

**Normal skewed distribution.** When the data follow a normal skewed distribution (i.e. have a moderate skewness; skewness = 1), the Kolmogorov-Smirnov test without lillefors has a power of 1% when there are 30 subjects in the sample. It increases to 3% when there are 50 subjects, and to 16% when there are 100 subjects. All other solutions are more powerful, but the Shapiro-Wilk test appears to be the best solution, regardless of the sample size (Table S1.4). However, even the Shapiro-Wilk test requires 50 subjects per group in order to detect the normality violations with a power of 95%.

**Chi-squared distribution.** With moderate to big sample sizes (i.e. ), when the data follow a chi-squared distribution (i.e. when data are highly skewed), all tests will almost always detect the departure from the normal assumption. It is already true with 30 subjects in the sample, except with the Kolmogorov-Smirnov test without lillefors correction. With small sample sizes (), the Shapiro-Wilk is always more powerful than the Kolmogorov-Smirnov test with correction and Jarque-Bera test (Table S1.5). Even with very small sample sizes, the test is very powerful.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table S1.1  *Real alpha risk, when nominal alpha risk = 5%, and distributions are normal* | | | | |
|  | **Test** | | | |
| n | **Kolmogorov-Smirnov**  **(no correction)** | **Kolmogorov-Smirnov**  **(lillefors correction)** | **Shapiro-Wilk** | **Jarque-Bera** |
| 10 | < 0.01 | 0.05 | 0.05 | 0.01 |
| 20 | < 0.01 | 0.05 | 0.05 | 0.02 |
| 30 | < 0.01 | 0.05 | 0.05 | 0.03 |
| 50 | < 0.01 | 0.05 | 0.05 | 0.04 |
| 100 | < 0.01 | 0.05 | 0.05 | 0.04 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table S1.2  Power when nominal alpha risk = 5%, and distributions are mixed normal | | | | | |
|  | **Test** | | | |
| n | **Kolmogorov-Smirnov**  **(no correction)** | **Kolmogorov-Smirnov**  **(lillefors correction)** | **Shapiro-Wilk** | **Jarque-Bera** |
| 10 | < 0.01 | 0.19 | 0.23 | 0.14 |
| 20 | 0.03 | 0.32 | 0.44 | 0.42 |
| 30 | 0.06 | 0.41 | 0.59 | 0.60 |
| 50 | 0.13 | 0.56 | 0.77 | 0.80 |
| 100 | 0.30 | 0.80 | 0.95 | 0.96 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table S1.3  Real alpha risk, when nominal alpha risk = 5%, and distributions are uniform | | | | | |
|  | **Test** | | | |
| n | **Kolmogorov-Smirnov**  **(no correction)** | **Kolmogorov-Smirnov**  **(lillefors correction)** | **Shapiro-Wilk** | **Jarque-Bera** |
| 10 | < 0.01 | 0.06 | 0.08 | < 0.01 |
| 20 | < 0.01 | 0.10 | 0.20 | < 0.01 |
| 30 | < 0.01 | 0.14 | 0.38 | < 0.01 |
| 50 | < 0.01 | 0.26 | 0.75 | < 0.01 |
| 100 | 0.01 | 0.59 | 1 | 0.56 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table S1.4  *Power when nominal alpha risk = 5%, and distributions are normal skewed* | | | | |
|  | **Test** | | | |
| n | **Kolmogorov-Smirnov**  **(no correction)** | **Kolmogorov-Smirnov**  **(lillefors correction)** | **Shapiro-Wilk** | **Jarque-Bera** |
| 10 | < 0.01 | 0.13 | 0.19 | 0.04 |
| 20 | < 0.01 | 0.23 | 0.44 | 0.16 |
| 30 | 0.01 | 0.35 | 0.68 | 0.28 |
| 50 | 0.03 | 0.56 | 0.93 | 0.54 |
| 100 | 0.16 | 0.91 | 1 | 0.96 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table S1.5  *Power when nominal alpha risk = 5%, and distributions are normal skewed* | | | | |
|  | **Test** | | | |
| n | **Kolmogorov-Smirnov**  **(no correction)** | **Kolmogorov-Smirnov**  **(lillefors correction)** | **Shapiro-Wilk** | **Jarque-Bera** |
| 10 | 0.11 | 0.80 | 0.93 | 0.44 |
| 20 | 0.61 | 0.99 | 1.00 | 0.90 |
| 30 | 0.94 | 1.00 | 1.00 | 0.99 |
| 50 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 1.00 | 1.00 | 1.00 | 1.00 |

## **Supplemental Material 2: Type I error rate of the *F*-test, *W*-test and *F\**-test**

Assuming a Type I error rate of 5% under the null, a test can yield either a significant result (*p*-value < 5%; or a “false positive” -FP) or a non-significant result (*p*-value > 5%; or a “true negative”-TN).

The Type I error rate is the relative frequency of effects detected as significant, under the null:

=

In order to estimate the Type I error rate for the *F*-test and 2 well known alternatives when population variances are unequal (*W*-test and *F\**-test, both available on SPSS), we simulated 1,000,000 datasets of k samples (where k is ranging from 2 to 5) under different conditions. There were 8 blocks of simulations each of them generated under 80 conditions (yielding 4\*8\*80,000,000 simulations in total).

Two steps were repeated for each condition: in a first step, the *p*-values of the three tests were extracted for each dataset, and in a second step, the percent of *p*-values under the nominal alpha risk (5%) was computed for each condition and for each test.

In the first 7 blocks, in each condition, *k*-1 samples were generated from a population where and sample sizes ( were 20,30,40,50 or 100. The standard deviation 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 (0.5,1,2 or 4). The set of simulations was repeated seven times, varying the distributions underlying the data. We used R commands to generate data from different distributions:

* ***k* normal distributions** (Figure A2.1): In order to assess the Type I error rate of the different tests under the assumption of normality, data were generated by means of the function “rnorm” (from the package “stats”; “R: The Normal Distribution,” 2016) .
* ***k* double exponential distributions** (Figure A2.2): In order to assess the impact of high kurtosis on the Type I error rate of all tests, data were generated by means of the function “rdoublex” (from the package “smoothmest”; "R: The double exponential (Laplace) distribution," 2012).
* ***k* mixed normal distributions** (Figure A2.3): In order to assess the impact of extremely high kurtosis on the Type I error rate of all tests, regardless of variance, data were generated by means of the function “rmixnorm” (from the package “bda”; Wang & Wang, 2015).
* ***k* normal right skewed distributions** (Figure A2.4): In order to assess the impact of moderate skewness on the Type I error rate, data were generated by means of the function “rsnorm” (from the package “fGarch”; “R: Skew Normal Distribution,” 2017). The normal skewed distribution was chosen because it is the only skewed distribution where the standard deviation ratio can vary without having an impact on skewness.
* ***k*-1 normal left skewed distributions** (Figure A2.4) **and 1 normal right skewed distribution** (Figure A2.5): In order to assess the impact of unequal shapes, in terms of skewness, on the Type I error rate, when data have moderate skewness, data were generated by means of the functions “rsnorm” (from the package “fGarch”; “R: Skew Normal Distribution,” 2017).
* ***k*-1 chi-squared distributions with two degrees of freedom** (See Figure A2.6)**, and one normal right skewed distribution** (Figure A2.4): In order to assess the impact of high asymetry on the Type I error rate, *k*-1 distributions were generated by means of the functions “rchisq” (“R: The (non-central) Chi-squared Distribution,” 2016; see Figures A2.6). Because the chi-squared is non-negative, it is not possible to generate chi-squared where = 1, 4 or 8 and µi is the same than the chi-squared with two degrees of freedom. However, we wanted to assess the impact of different *SD*-ratio on Type I error rate. For these reasons, the last distribution was generated by means of “rsnorm” in order to follow a normal skewed distribution with positive skewness of +0.99 and mean = 2 (from the package “fGarch”; “R: Skew Normal Distribution,” 2017).
* ***k*-1 chi-squared distributions with two degrees of freedom** (See Figure A2.6)**, and one normal left skewed distribution** (Figure A2.5): In order to assess the impact of unequal shapes, in terms of skewness, on Type I error rate when distributions have extreme skewness, *k*-1 distributions were generated by means of the functions “rchisq” (“R: The (non-central) Chi-squared Distribution,” 2016). The last distribution was generated by means of “rsnorm” in order to follow a normal right skewed distribution with a mean of 2 (from the package “fGarch”; “R: Skew Normal Distribution,” 2017).

|  |
| --- |
|  |
| **Figure A2.1**: centered normal probability density function, as a function of . |

|  |
| --- |
|  |
| **Figure A2.2**: centered double exponential probability density function, as a function of . |

|  |
| --- |
|  |
| **Figure A2.3**: centered mixed normal probability density function, as a function of . |

|  |
| --- |
|  |
| **Figure A2.4**: centered normal right skewed probability density function, as a function of .   |  | | --- | |  | | **Figure A2.5**: centered normal left skewed probability density function, as a function of . | |  | | **Figure A2.6**: chi-squared with 2 degrees of freedom probability density function. | |

There is commonly a confusion between kurtosis and variance (DeCarlo, 1997). One reason is that the standard scale parameter of the double exponential distribution is which is not equal to the standard deviation: β = . A last set of simulations was created and in each condition, *k*-1 samples were generated from a double exponential distribution where β were 2 (i.e. j2.82) and sample sizes ( were 20,30,40,50 or 100. The scale parameter β 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 (0.5,1,2 or 4).

In order to facilitate the assessment of the quality of each test, we classified conditions into five categories for each block of simulation. Previous findings highlighted differences in terms of the Type I error rate of the *F*-test as a function of the correlation between *ns* 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 *ns* 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 *ns* 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.

Finally, we computed 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 categories are split as a function of the number of subjects in the first group. Considering the large amount of data, we decided to report data on the osf platform, at the following link: https://osf.io/ru9tz/ (See Type I error rate.xlsx).

**Results of the *F*-test.** When both assumptions of homoscedasticity and normality are met, the Type I error control is very close to the nominal 5%, whatever the sample sizes, the number of groups to compare, and whether designs are balanced (i.e. the sample sizes are equal between groups) or not, which is conform to our expectations.

However, consistently with research conducted by Minitab statisticians (Minitab assistance, *n.d*), simulations show that the Type I error rate of the *F*-test can differ noticeably from the nominal Type I error rate (i.e., 5%) when the groups have different variances. In light of the definition of Bradley (1978), one can consider the alpha risk “sufficiently close” to the nominal alpha risk if its value falls in the interval [0.025; 0.075] (Hayes & Cai, 2007). Considering this norm, one can observe that when there is a positive correlation between *ns* and *SD*s, the test is often too conservative and when there is a negative correlation between *ns* and *SD*s, the test is often too liberal. Interestingly, there is no improvement when one increases sample sizes. Consistently with what we observed in a previous paper (see Delacre, Lakens, & Leys, 2017), when there are only two groups to compare, the test is robust against unequal variances when sample sizes are equal between groups (and the bigger is the sample size, the truest it is). However, when there are more than two groups to compare, unequal variances are problematic even when sample sizes are equal between groups. In this case, the test becomes too liberal.

When data are not normally distributed but there are equal variances between groups, the Type I error rate is close to the nominal 5%. One can conclude that the impact of normality violations on Type I error rate is smaller than the impact of unequal variances.

However, when both normality and equal variances are not met, while the impact of heavier tailed distributions on Type I error rate remains marginal unless kurtosis is very big , the test becomes more liberal when distributions are skewed (even with moderate skewness or when the skewness has the same valence between groups). It is consistent with other studies showing that tests based on means are more affected by skewness than by kurtosis (DeCarlo, 1997).

Finally, when the number of groups increases, the test becomes more liberal. It improves the test when there is a positive correlation between *ns* and *SD*s (because the Type I error rate increases and it is closer to the nominal 5%), however, it is problematic when there is either negative correlation between *ns* and *SD*s, or heteroscedasticity with balanced designs.

**Results of the *F\**-test.** As long as the assumptions of homoscedasticity and normality are met, the *F\**-test has a good Type I error control, regardless of the sample sizes, the number of groups and of whether sample sizes are equal or not.

Under the normality assumption, the test is robust against heteroscedasticity when there are only two groups to compare, which is not surprising since the *F\**-test is identical to the *W*-test when there are only two groups. However, when the number of groups increases, the test is affected by violations of the assumption of equal variances between groups (although it is less affected by these violations than the *F*-test, by construction). Most of the time, the test becomes more and more liberal when k increases.

When data are not normally distributed but there are equal variances between groups, and sample sizes are small, the test becomes more liberal when there are big differences between groups, in terms of skewness.

Finally, when neither the assumption of homoscedasticity nor the assumption of normality is met, the test becomes a little more liberal when distributions are skewed, even with balanced designs, contrary to the statements of Lix, Keselman, & Keselman (1996). In conclusion, we do not recommend the use of the *F\**-test as a default strategy to compare means.

**Results of the *W*-test[[2]](#footnote-2).** As well as the *F*-test and *F\**-test, when the assumption of homoscedasticity and normality are met, the *W*-test has a good Type I error control, regardless of the number of groups and of whether sample sizes are equal or not.

Contrary to the *F*-test and *F\**-test, the *W*-test is totally immune to the violations of the assumption of equal variances, as long as the normality assumption is met. However, the test is more sensitive to the normality violations than both the *F*-test and *F\**-test. Even under the homoscedasticity assumption, one can observe an effect of skewness: when skewness moves away from 0, the test becomes more liberal (especially when distributions are largely skewed combined with skewness’s of opposite sign), even with balanced designs. When the number of groups increases, the test is more and more liberal.

Moreover, it seems to have an interaction effect between violations of both normality and heteroscedasticity: when there is heteroscedasticity, the test becomes even more liberal with skewed distributions, particularly when a high *SD*-ratio is associated with a negative correlation between *ns* and *SD*s, and when distributions are largely skewed combined with skewness’s of opposite sign. The interaction effect between skewness and heteroscedasticity is consistent with the statement of Lix, Keselman, & Keselman (1996) about the Alexander-Govern and the James second order tests.

When variances are unequal, it makes no sense to compare directly the *F*-test with the *W*-test and *F\**-test, because the *F*-test is not valid, as previously explained. However, when there are equal variances between groups, one can observe that:

1. the *W*-test is a little more affected by heavy tailed distributions, in terms of Type I error rate, than the *F*-test (meaning that it becomes more conservative with heavy tailed distributions than the *F*-test, but the difference is marginal unless kurtosis is really high, like when data are extracted from mixed distributions).
2. the *W*-test becomes too liberal with highly skewed distributions (such as with chi-squared distributions with two degrees of freedom), while the *F*-test has a good Type I error control in these situations.

These two observations are more and more true when the number of groups increases (i.e. when the number of groups increases, the test becomes even more conservative with heavy tailed distributions and even more liberal with moderately/highly skewed distributions; for example, the test is robust to moderate asymmetry when there are 3 groups to compare, however it is not when the number of groups increases), even if sample sizes are big (more than 50 subjects per group).

In conclusion, the *W*-test is more robust than the *F*-test and *F\**-test, in terms of type I error, almost all the time, even when nonnormality are combined with largely unequal variances (i.e. when the *W*-test is the less efficient). However, the three tests should be avoided when small samples extracted from highly skewed distributions are combined with unequal variances. In all other case, we recommend considering the *W*-test as the default to compare means.

## **Supplemental Material 3: power of the *F*-test, *W*-test and *F\**-test**

Assuming the null hypothesis is false and a Type I error rate of 5%, a test can yield either a significant result (*p*-value < 5%; or a “true positive” -TP) or a non-significant result (*p*-value > 5%; or a “false negative”-FN). The power is the relative frequency of effects detected as significant, when the null is false (i.e. when there are real differences between groups):

Power= =

In order to compute the power of the *F*-test and 2 famous alternatives when population variances are unequal (*W*-test and *F\**-test of comparison of means, both available on SPSS), we performed 1,000,000 simulations of k samples (2 or 3)[[3]](#footnote-3) generated under 560 conditions (yielding 2\*560,000,000 simulations in total).

In each condition, *k*-1 samples were generated from a population where and sample size was 20,30,40,50 or 100. The standard deviation and the sample size of the last sample is a function of the sample sizes ratio (*n*-ratio = ; ranging from 0.5 to 2 in steps of 0.5) and the *SD*-ratio (0.5,1,2 or 4). In all conditions, the mean of *k*-1 groups was 0 and the mean of the last group was 1. Note that because standard deviations vary from one condition to another but the means difference is constant, the effect size is not systematically the same in all conditions.

The set of simulations was repeated seven times varying the distributions underlying the data.

* ***k* normal distributions**: In order to assess the power of all tests when the normality assumption is met, data were generated by means of the function “rnorm” (from the package “stats”; “R: The Normal Distribution,” 2016) .
* ***k* double exponential distributions**: In order to assess the impact of high kurtosis on the power of all tests, data were generated by means of the function “rdoublex” (from the package “smoothmest”; "R: The double exponential (Laplace) distribution," 2012).
* ***k* mixed normal distributions**: In order to assess the impact of extremely high kurtosis on the power of all tests, regardless of variance, data were generated by means of the function “rmixnorm” (from the package “bda”; Wang & Wang, 2015).
* ***k* normal skewed distributions with positive skewness of +0.99**: In order to assess the impact of moderate skewness on the power of all tests, data were generated by means of the function “rsnorm” (from the package “fGarch”; “R: Skew Normal Distribution,” 2017). The normal skewed distribution was chosen because it is the only skewed distribution where the standard deviation ratio can vary without having an impact on skewness.
* ***k*-1 normal skewed distributions with positive skewness of +0.99** **and 1 normal skewed distribution with negative skewness of -0.99**: In order to assess the impact of unequal shapes, in terms of skewness, on the power of all tests, when data have moderate skewness, data were generated by means of the functions “rsnorm” (from the package “fGarch”; “R: Skew Normal Distribution,” 2017).
* ***k*-1 chi-squared distributions with two degrees of freedom, and one normal skewed distribution with positive skewness of +0.99**: In order to assess the impact of high asymetry on the power of all tests, *k*-1 distributions were generated by means of the functions “rchisq” (“R: The (non-central) Chi-squared Distribution,” 2016). Because the chi-squared is non-negative, it is not possible to generate chi-squared where = 1, 4 or 8 and µi is the same than the chi-squared with two degrees of freedom. However, we wanted to assess the impact of different *SD*-ratio on Type I error rate. For these reasons, the *k*th distribution was generated by means of “rsnorm” in order to follow a normal skewed distribution with positive skewness of +0.99 and mean = 2 (from the package “fGarch”; “R: Skew Normal Distribution,” 2017).
* ***k*-1 chi-squared distributions with two degrees of freedom, and one normal skewed distribution with negative skewness of -0.99**: In order to assess the impact of unequal shapes, in terms of skewness, on power of all tests, when distributions have extreme skewness, *k*-1 distributions were generated by means of the functions “rchisq” (“R: The (non-central) Chi-squared Distribution,” 2016). The *k*th distribution was generated by means of “rsnorm” in order to follow a normal skewed distribution with positive skewness of +0.99 and mean = 2 (from the package “fGarch”; “R: Skew Normal Distribution,” 2017).

Finally, because of a common confusion between kurtosis and variance (DeCarlo, 1997, see Supplemental Material 2), a last set of simulations was created and in each condition, *k*-1 samples were generated from a double exponential distribution where β were 2 (i.e. j2.82) and sample sizes ( were 20,30,40,50 or 100. The scale parameter β and the sample size of the *k*th 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 (0.5,1,2 or 4).

The observed power was computed by repeating two steps for each condition: in a first step, the *p*-values of the *F*-test, *W*-test and *F*\*-test were extracted for each dataset, and in a second step, the percent of *p*-values under the nominal alpha risk (5%) was computed for each condition and for each test. We used R commands to generate data from different distributions.

In order to insure the reliability of our calculation method, the observed power, computed when data were extracted from normal distributions, was compared with theoretical power, i.e. the power computed using the power function of the *W*-test, *F*-test and *F\**-test. When assumptions underlying each test are met (i.e. normality for all tests, and equal variances for *F*-test), the computed power is very consistent with theoretical power, one can therefore conclude that the method is reliable. The theoretical power was then compared with observer power, computed when data were extracted from nonnormal distributions, in order to assess the impact of normality violations on power.

In order to facilitate the assessment of the quality of each test, we classified conditions into five categories for each block of simulation, as for the Type I error rate (see Supplemental Material 2). Considering the large amount of data, we decided to report data on the osf platform, at the following link: https://osf.io/ru9tz/ (See Power.xlsx).

**Results of the *F*-test.** When the normality assumption is met, but the homoscedasticity assumption is not met, the power of the *F*-test is not consistent with theoretical expectations. It is particularly true with unequal sample sizes between groups: when there is a positive correlation between *ns* and *SD*s, power is smaller than expectations, meaning that the power-curve will conduct to overestimate the real power (even when there are 100 subjects per groups). On the other side, when there is a negative correlation between *ns* and *SD*s, power is bigger than expectations, meaning that the power-curve will conduct to underestimate the real power (even when there are 100 subjects per groups). Finally, with equal sample sizes between groups and unequal variances, the power curve with either underestimate the real power (with small sample sizes; i.e ni=20) or overestimate the real power (with big sample sizes; i.e. ni=100).

When the assumption of equal variances is met, one obtains a gain in power especially when distributions have a big kurtosis, or when high skewnesses are combined with skewnesses of opposite signs. However, the bigger are sample sizes, the closest is the power from the power in normal cases. For examples, with 50 subjets per group, deviations between the observed power and the expected power decreases, whatever distributions the data are extracted from. Finally, when the assumption of equal variances is not met, the effect of non normal distributions can become bigger than where variances are equal between groups.

## **Results of the *W*-test**. When the normality assumption is met, but the homoscedasticity assumption is not, contrary to what was observed for *F*-test, the power of the *W*-test is very consistent with theoretical expectations, because the *W*-test is robust against homoscedasticity violations.

However, the *W*-test is in general more affected by abnormality violations than *F*-test, except when homoscedasticity is combined with equal sample sizes and only two groups to compare. In all other situations, there is a bigger gain in power with *W*-test than with *F*-test, particularly when distributions have a high kurtosis, or are highly skewed with unequal skewnesses between groups. Moreover, the gain in power is more important when sample sizes are unequal between groups. However, the bigger are sample sizes, the closest is the power from the power in normal cases.

**Results of the *F\**-test.** When there are only two groups to compare, the *F\**-test and *W*-test are identical. The power of *F\**-test is therefore very consistent with theoretical expectations, even when variances are unequal between groups. However, when there are more than two groups to compare and unequal variances between groups, power of the *F\**-test is more consistent with theoretical expectations than power of the *F*-test, but less consistent than power of the *W*-test. Whatever the correlation between *ns* and *SD*s (positive, negative or null), power is bigger than expectations, meaning that the power-curve will conduct to underestimate the real power. This is particularly true with small sample sizes. When sample sizes increase, the gain in power decreases (and one observes a power lower than expectations, even when there are 100 subjects per groups).

Moreover, with heavy tailed distributions, there is a gain in power in comparison with normal distributions. When distributions are skewed, there is either a gain or a loss in power, in comparison with normal distributions, depending on the highness of skewness, and if skewness are of same or opposite signs between groups.

Finally, when the assumption of equal variances is not met, the *F\**-test is more affected by normality violations than the *F*- test, but less affected by abnormality violations than the *W*-test. However, as both other tests, the effect of skewed distributions becomes bigger than where variances are equal between groups and depends on the situation (either a gain or a loss in power).

**Reference**

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological methods*, *2*(3), 292-307. [doi:10.1037//1082-989x.2.3.292](http://dx.doi.org/10.1037/1082-989x.2.3.292)

Ghasemi, A., & 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)

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior research methods*, *39*(4), 709–722. [doi:10.3758/bf03192961](http://dx.doi.org/10.3758/bf03192961)

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of educational research*, *66*(4), 579–619.

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

Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of four different normality tests in terms of Type I error rate and power under different distributions. *Turkish Journal of Medical Sciences*, *36*(3), 171–176.

R: Skew Normal Distribution. (n.d.). Retrieved January 22, 2017, from http://finzi.psych.upenn.edu/library/fGarch/html/dist-snorm.html

R: The double exponential (Laplace) distribution. (n.d.). Retrieved January 22, 2017, from https://artax.karlin.mff.cuni.cz/r-help/library/smoothmest/html/ddoublex.html

R: The (non-central) Chi-squared Distribution. (n.d.). Retrieved January 22, 2017, from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Chisquare.html

R: The Normal Distribution. (n.d.). Retrieved January 20, 2017, from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Normal.html

Steinskog, D. J., Tjøstheim, D. B., & Kvamstø, N. G. (2007). A cautionary note on the use of the Kolmogorov–Smirnov test for normality. *Monthly Weather Review*, *135*(3), 1151–1157.

Wang, B., & Wang, M. B. (2015). Package ‘bda’. Retrieved January 20, 2017, from http://ftp.cs.pu.edu.tw/network/CRAN/web/packages/bda/bda.pdf

Wilcox Rand, R., & Wilcox Rand, R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Academic Press.

1. These sample sizes were chosen because it will be shown later that depending on the distribution underlying the data, 50 or 100 subjects per group are needed in order that the type 1 error rate converge to the nominal 5%. [↑](#footnote-ref-1)
2. Note that the Type 1 error rate of the *W*-test is very close to Type 1 error rate of Jame’s test and Alexander-Govern test, meaning that they have very similar strengths and limitations. These two last tests are both available in R, but not in SPSS. [↑](#footnote-ref-2)
3. Note that it is not possible to compare results when k=2 and when k=3, because the sample size is not the same. [↑](#footnote-ref-3)