Monte Carlo simulations: *F*-test versus *W*-test versus *F**-test

We performed Monte Carlo simulations using R (version 3.5.0) to assess the Type I and Type II error rates for the three tests. One million datasets were generated for 3840 scenarios that address the arguments present in the literature. In 2560 scenarios, means were equal across all groups (i.e. the null hypothesis is true), in order to assess the Type I error rate of the tests. In 1280 scenarios, there were differences between means (i.e. the alternative hypothesis is true) in order to assess the power of the tests. In all scenarios, when using more than 2 samples, all samples but one was generated from the same population, and only one group had a different population mean.

Population parameter values were chosen in order to illustrate the consequences of factors known to play a key role on both the Type I error rate and the statistical power when performing an ANOVA. Based on the literature review presented above, we manipulated the number of groups, the sample sizes, the sample size ratio (n-ratio $= \frac{n_k}{n_j}$), the SD-ratio (SD-ratio $=\frac{\sigma_k}{\sigma_j}$), and the sample size and variance pairing. In our scenarios, the number of compared groups (k) varied from 2 to 5. Sample sizes of k-1 groups (n_j) were 20, 30, 40, 50, or 100. The sample size of the last group was a function of the *n*-ratio, ranging from 0.5 to 2, in steps of 0.5. The simulations for which the *n*-ratio equals 1 are known as a balanced design (i.e. sample sizes are equal across all groups). The *SD* of the population from which was extracted last group was a function of the *SD*-ratio, with values of 0.5, 1, 2 or 4. The simulations for which the *SD*-ratio equals 1 are the particular case of homoscedasticity (i.e. equal variances across groups).

All possible combinations of *n*-ratio and *SD*-ratio were performed in order to distinguish positive pairings (the group with the largest sample size is extracted from the population with the largest *SD*), negative pairings (the group with the smallest sample size is extracted from the population with the smallest *SD*), and no pairing (sample sizes and/or population *SD* are equal across all groups). All of those conditions were tested with normal and non-normal distributions. When two groups are compared, conclusions for the three ANOVA tests (*F*, *F**, *W*) should yield identical error rates when compared to their equivalent *t*-tests (the *F*-test is equivalent to Student's *t*-test, and the *F**-test and *W*-test are equivalent to Welch's *t*-test; Delacre et al., 2017). When there are more than three groups, 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 (Harwell et al., 1992; Quensel, 1947).

For didactic reasons, we will report only the results where we compared three groups (k = 3). Increasing the number of groups increases how liberal all tests are. For interested readers, all figures for cases where we compare more than three groups are available here: https://osf.io/h4ks8/. Overall, the larger the sample sizes, the less the distributions of the population underlying the samples impact the robustness of the tests (Srivastava, 1959). However, increasing the sample sizes does not improve the robustness of the test when there is heteroscedasticity. Interested reader can see all details in the following Excel spreadsheet, available on github: « Type I error rate.xlsx ».

In sum, the simulations grouped over different sample sizes yield 9 conditions based on the *n*-ratio, *SD*-ratio, and sample size and variance pairing, as summarized in **Table 1**.

In all Figures presented below, averaged results for each sub-condition are presented under seven different configurations of distributions, using the following legend.

Type I Error Rate of the *F*-test, *W*-test, and *F**-test

As previously mentioned, the Type I error rate (α) is the long-run frequency of observing significant results when the null-hypothesis is true. When means are equal across all groups the Type I error rate of all test should be equal to the nominal alpha level. We assessed the Type I error rate of the *F*-test, *W*-test and *F**-test under 2560 scenarios using a nominal alpha level of 5%.

When there is no difference between means, the nine cells of **Table 1** simplify into five sub-conditions:

- Equal *n* and *SD* across groups (a)
- Unequal *n* but equal *SD* across groups (b and c)
- Unequal *SD* but equal *n* across groups (d and g)
- Unequal *n* and *SD* across groups, with positive correlation between *n* and *SD* (e and i)
- Unequal *n* and *SD* across groups, with negative correlation between *n* and *SD* (f and h)

Table 1: 9 conditions based on the *n*-ratio, *SD*-ratio, and sample size and variance pairing.

		***n*-ratio**		
		1	**>1**	**<1**
SD-ratio	1	a	b	c
	>1	d	e	f
	<1	g	h	i

Note: The *n*-ratio is the sample size of the last group divided by the sample size of the first group. When all sample sizes are equal across groups, the *n*-ratio equals 1. When the sample size of the last group is higher than the sample size of the first group, *n*-ratio >1, and when the sample size of the last group is smaller than the sample size of the first group, *n*-ratio <1. *SD*-ratio is the population *SD* of the last group divided by the population *SD* of the first group. When all samples are extracted from populations with the same *SD*, the *SD*-ratio equals 1. When the last group is extracted from a population with a larger *SD* than all other groups, the *SD*-ratio >1. When the last group is extracted from a population with a smaller *SD* than all other groups, the *SD*-ratio <1.

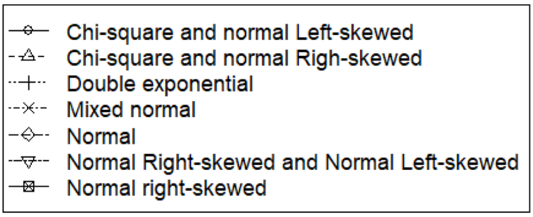
Chi-square and normal Left-skewed
Chi-square and normal Righ-skewed
Double exponential
Mixed normal
Normal
Normal Right-skewed and Normal Left-skewed
Normal right-skewed

Figure 1: Legend.