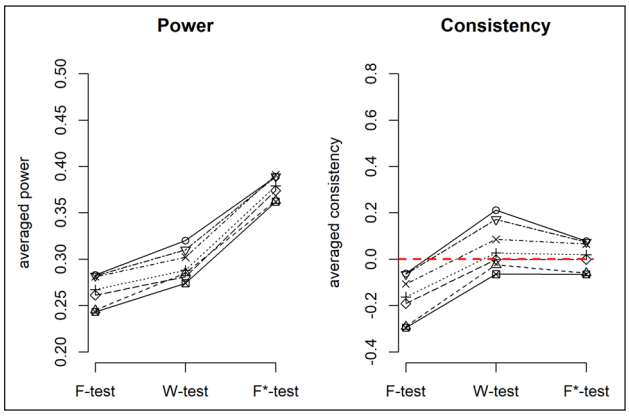

Figure 12: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation betwen sample sizes and SDs, and positive correlation between SDs and means (cell f in Table 1).

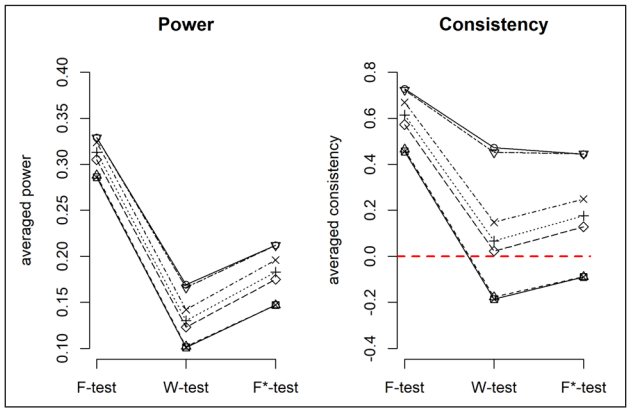

Figure 13: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation betwen sample sizes and SDs, and negative correlation between SDs and means (cell h in Table 1).

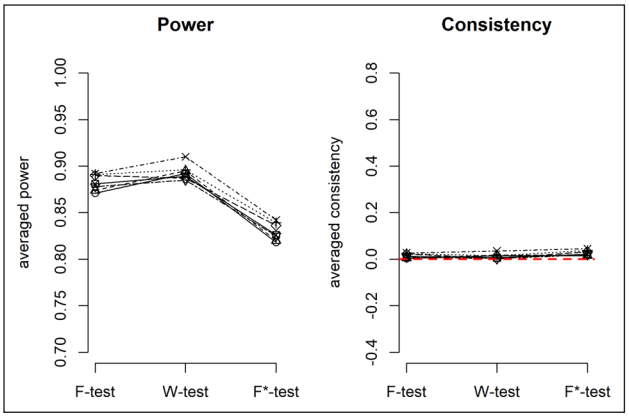

Figure 14: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation betwen sample sizes and SDs, and positive correlation between SDs and means (cell e in Table 1).

correlation between *SDs* and means. Note that for all tests, the effect of heteroscedasticity is approximately the same regardless of the shape of the distribution. Moreover,

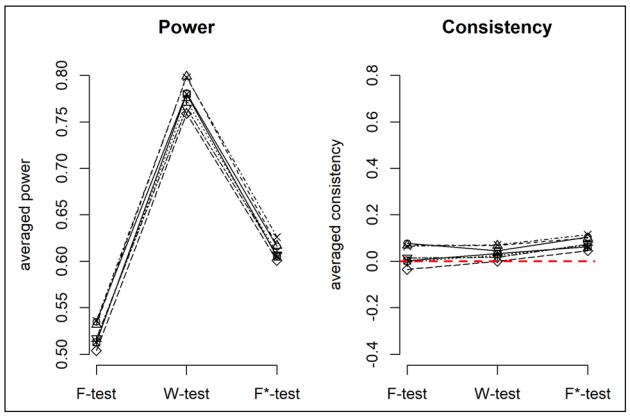

Figure 15: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation betwen sample sizes and SDs, and negative correlation between SDs and means (cell i in Table 1).

there is one constant observation in our simulations: whatever the configuration of the *n*-ratio, the consistency of the three tests is closer to zero when there is a negative correlation between the *SD* and the mean (meaning that the group with the highest mean has the lower variance).

We can draw the following conclusions about the statistical power of the three tests:

1) When all assumptions are met, the *W*-test falls slightly behind the *F*-test and the *F**-test, both in terms of power and consistency.
2) When variances are equal between groups and distributions are not normal, the *W*-test is slightly more powerful than both the *F*-test and the *F**-test, even with heavy-tailed distributions.
3) When the assumption of equal variances is violated, the *F*-test is either too liberal or too conservative, depending on the correlation between sample sizes and *SDs*. On the other side, the *W*-test is not influenced by the sample sizes and *SDs* pairing. However, it is influenced by the *SD* and means pairing.
4) The last conclusion generally remains true when both assumptions of equal variances and normality are not met.

Recommendations

Taking both the effects of the assumption violations on the alpha risk and on the power, we recommend using the *W*-test instead of the *F*-test to compare groups means. The *F*-test and *F**-test should be avoided, because a) the equal variances assumption is often unrealistic, b) tests of the equal variances assumption will often fail to detect differences when these are present, c) the loss of power when using the *W*-test is very small (and often even negligible), and d) the gain in Type I error control is considerable under a wide range of realistic conditions. Also, we recommend the use of balanced designs (i.e. same sample sizes 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 samples sizes between groups. Our simulations show that the Type