INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32(1): 13, 1–12. DOI: https://doi.org/10.5334/irsp.198

RESEARCH ARTICLE

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

Marie Delacre*, Christophe Leys*, Youri L. Mora* and Daniël Lakens†

Student's *t*-test and classical *F*-test ANOVA rely on the assumptions that two or more samples are independent, and that independent and identically distributed residuals are normal and have equal variances between groups. We focus on the assumptions of normality and equality of variances, and argue that these assumptions are often unrealistic in the field of psychology. We underline the current lack of attention to these 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 under a range of realistic scenarios, 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 practical recommendations that researchers can use to improve their statistical practices.

Keywords: ANOVA; Welch test; parametric test; parametric assumptions; replicability crisis

When comparing independent groups researchers often analyze the means by performing a Student's *t*-test or classical Analysis of Variance (ANOVA) *F*-test (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998; Tomarken & Serlin, 1986). Both tests rely on the assumptions that independent and identically distributed residuals (1) are sampled from a normal distribution and (2) have equal variances between groups (or homoscedasticity; see Lix, Keselman, & Keselman, 1996). While a deviation from the normality assumption generally does not strongly affect either the Type I error rates (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Tiku, 1971) or the power of the *F*-test (David & Johnson, 1951; Harwell et al., 1992; Srivastava, 1959; Tiku, 1971), the *F*-test is not robust against unequal variances (Grissom, 2000). Unequal variances can alter both the Type I error rate (David & Johnson, 1951; Harwell et al., 1992) and statistical power (Nimon, 2012; Overall, Atlas, & Gibson, 1995) of the *F*-test.

Although it is important to make sure test assumptions are met before a statistical test is performed, researchers rarely provide information about test assumptions when they report an *F*-test. We examined statistical tests reported in 116 articles in the *Journal of Personality and Social Psychology* published in 2016. Fourteen percent of these articles reported a one-way *F*-test, but only one article indicated that the homogeneity of variances assumption was taken into account. They reported corrected degrees of freedom for unequal variances, which could signal the use of the *W*-test instead of the classical *F*-test. A similar investigation (Hoekstra, Kiers & Johnson, 2012) yielded 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, better alternatives exist, such as the Welch's *W* ANOVA (*W*-test), the Alexander-Govern test, James' second order test, and the Brown-Forsythe ANOVA (*F**-test). Although not the focus of the current article, additional tests exist that allow researchers to compare groups either based on other estimators of central tendency than the mean (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998), or based on other relevant parameters of distribution than the central tendency, such as standard deviations and the shape of the distribution (Grissom, 2000; Tomarken & Serlin, 1986). However, since most

* Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, BE

† Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, NL

Corresponding author: Marie Delacre (marie.delacre@ulb.ac.be)