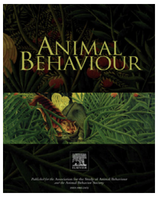
Animal Behaviour 136 (2018) 119–126

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav



Commentary

Beware the *F* test (or, how to compare variances)

D. J. Hosken*, D. L. Buss, D. J. Hodgson*

Centre for Ecology & Conservation, University of Exeter, Cornwall, Penryn U.K.

ARTICLE INFO

Article history:
Received 15 May 2017
Initial acceptance 1 June 2017
Final acceptance 23 November 2017

MS number: 17-00407R

Keywords:
Box–Anderson test
F test
jackknife
Levene's test
normality
permutation
power
variance

Biologists commonly compare variances among samples, to test whether underlying populations have equal spread. However, despite warnings from statisticians, incorrect testing is rife. Here we show that one of the most commonly employed of these tests, the *F* test, is extremely sensitive to deviations from normality. The *F* test suffers greatly elevated false positive errors when the underlying distributions are heavy tailed, a distribution feature that is very hard to detect using standard normality tests. We highlight and assess a selection of parametric, jackknife and permutation tests, consider their performance in terms of false positives, and power to detect signal when it exists, then show correct methods to compare measures of variation among samples. Based on these assessments, we recommend using Levene's test, Box–Anderson test, jackknifing or permutation tests to compare variances when normality is in doubt. Levene's and Box–Anderson tests are the most powerful at small sample sizes, but the Box–Anderson test may not control type I error for extremely heavy-tailed distributions. As noted previously, do not use *F* tests to compare variances.

> *Never use an F-test to test equality of variances* (Van Valen, 2005, **page 30**)
>
> *The effects of nonnormality on the distribution theories for the test statistics … are catastrophic* (Miller, 1998, **page 264**)

Evolutionary biologists and behavioral ecologists study variation alongside averages, and commonly wish to partition observed variation among various causes. This is of course the basis of analysis of variance (ANOVA) and its associated family of tests, where variation is partitioned among and within experimental treatments (predictors), to determine their influence on the response variable(s).

Sometimes, however, we are also interested in comparing the size of the variances themselves, among samples or treatments, to ask is there more variation in A than in B? Classic examples include comparing variation in behavioural plasticity, sex-specific variation in fitness, variance in sex ratios, variance in dietary breadth or preference, variation in preferred group size, and even how intra-individual variation in trait size can affect mating success (e.g. Brown & Robinson, 2016; Craft, 2016; Hosken, 2001; MacLeod & Clutton Brock, 2013; Shafir, Menda, & Smith, 2005; Sutherland, 1985; reviewed in Krebs & Davies, 1978, 1997; Westneat & Fox, 2010).

Another common reason to compare sample variances is as a diagnostic check for homogeneity of variance, prior to using ANOVA. Given the importance of the question ('Do the variances differ?'), we seek a statistical test that tells us the probability of detecting the observed signal were the null hypothesis to be true. This *P* value is commonly considered 'significant' if it lies below the conventional threshold of 0.05. So a test of variances must, if it is to be accurate and effective, satisfy two statistical conditions. First, it should have a low probability of concluding different variances when in fact the samples are drawn from the same underlying population. This is the type I (or false positive) error rate, and conventionally it should be 0.05. Second, the test should have a high probability of detecting a significant difference when samples are drawn from populations with genuinely different variances. This is called statistical 'power'. Inevitably power decreases with decreasing difference in variance between the underlying populations, such that small differences in population variances can be hard to detect.

A standard statistical approach, among biologists at least, is to use the *F* test to ask whether variance ratios differ significantly from unity. However, as Van Valen (1978, 2005), Miller (1998) and many

* Correspondence: D. J. Hosken and D. J. Hodgson, Centre for Ecology & Conservation, University of Exeter, Cornwall, Penryn TR10 9EZ, U.K.

E-mail addresses: d.j.hosken@exeter.ac.uk (D. J. Hosken), d.j.hodgson@exeter.ac.uk (D. J. Hodgson).

https://doi.org/10.1016/j.anbehav.2017.12.014
0003-3472/© 2018 The Association for the Study of Animal Behaviour. Published by Elsevier Ltd. All rights reserved.

other statisticians (e.g. Box, 1953) have noted, this is inappropriate. Unfortunately, biologists have not heeded warnings from statisticians (as we have noted when serving as both editors and referees), and incorrect testing keeps occurring. As part of the continuing battle against inappropriate and anticonservative (failure to control type I error) statistical analyses, we reiterate points raised by Van Valen (2005) and Miller (1998) by bringing this issue to the attention of a larger audience. We provide a comparison of statistical tests designed to compare sample variances, and use numerical simulations to demonstrate risks of false positive and false negative conclusions with increasingly severe deviations from normality. We focus on absolute variation in continuous variables, but point readers to Van Valen (1974) for suggestions on discrete variables.

Denouncement of the *F* test might seem rather heretical, given its deep roots in the statistical training of all biologists. The bad news is that *F* tests of the equality of variances are highly sensitive to deviations from normality of the underlying data distributions (Fig. 1). Van Valen (2005) linked this sensitivity to violations of the central limit theorem, but Miller (1998) attributed the problem more properly to a direct mathematical dependence of the variance of the sample variance on the kurtosis of the underlying probability distribution, damped by the sample size. The *F* test is very insensitive to the data's third moment, skew, but highly sensitive to its fourth, kurtosis (Miller, 1998; Fig. 1). Kurtosis measures the clustering of data around the mode, relative to variance: leptokurtic distributions have most data clustered tightly around the mode, coupled with very extreme values, and are therefore 'heavy tailed'. Platykurtic distributions are less clustered around the mode, coupled with a paucity of extreme values, and are therefore 'light tailed'. Heavy-tailed distributions risk very high rates of falsely positive *F* tests (i.e. type I error >0.05), while light-tailed distributions can yield painfully conservative tests (i.e. type I error <0.05). The good news is that *F* tests used in standard ANOVA are very robust to minor deviations from normality, for two reasons. First, the numerator of ANOVA tests represents variance among means;

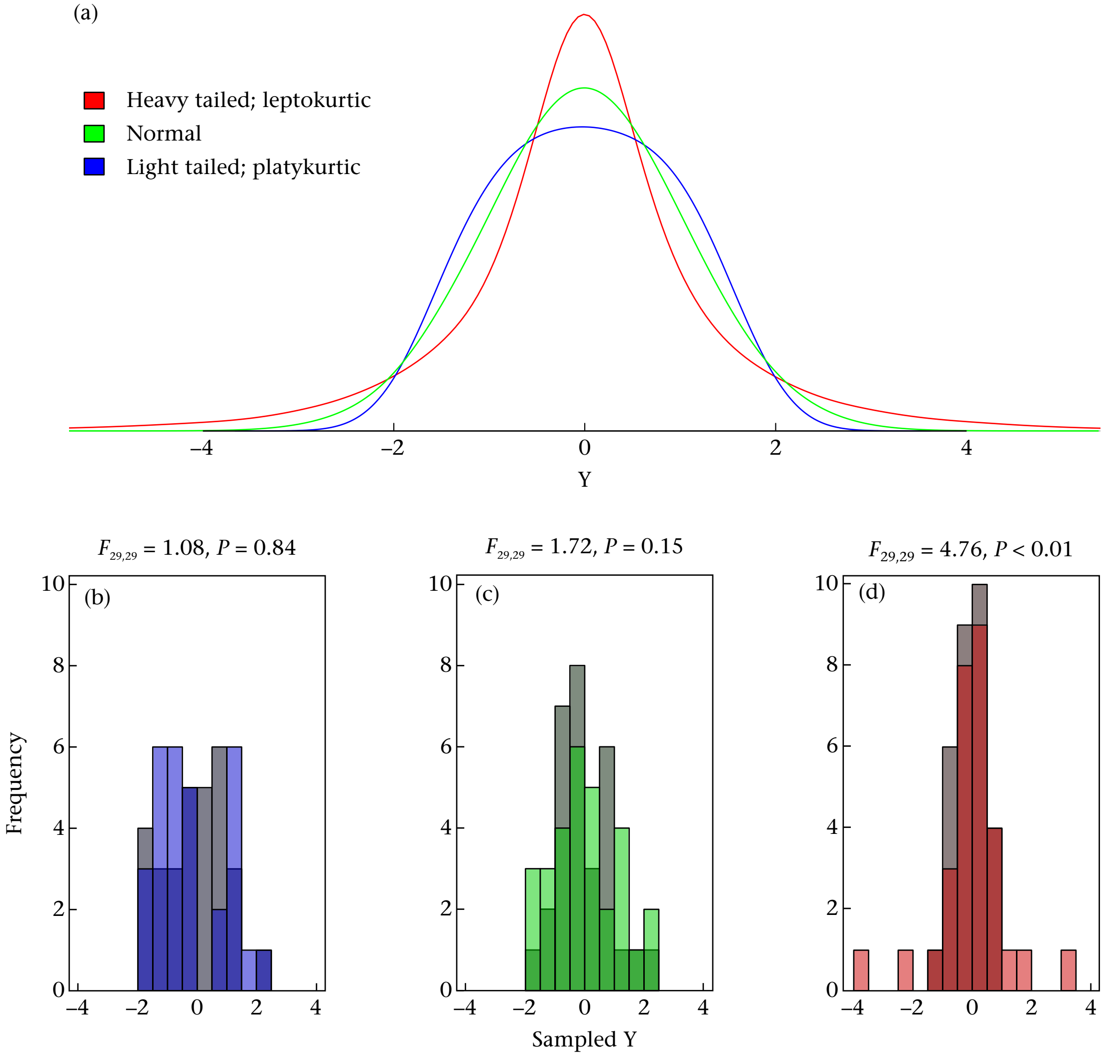


Figure 1. The influence of kurtosis on *F* test comparisons of sample variances. (a) Probability distribution functions of a population's phenotypic measurement 'Y': normal/Gaussian distribution (green); a heavy-tailed distribution (red; kurtosis parameter $\delta = 0.5$) and a light-tailed distribution (blue; $\delta = 100$). Each distribution has a mean of 0 and a standard deviation of 1. From each population we draw two samples of $N = 30$, mimicking the null hypothesis of no difference in variance. (b–d) Histograms of the samples from each population, and the results of *F* tests. In each case, darker bars show where the samples overlap. (b) Two samples drawn from a light-tailed distribution overlap considerably, have similar variance (the spread of the grey and light blue bars is similar) and yield an *F* ratio close to 1. (c) Two samples from a normal distribution overlap, but the light green sample has greater variance (although the *P* value correctly concludes not significantly so). (d) Two samples from a heavy-tailed population have overlapping means but the light red sample has a much greater variance (and the *P* value yields a type I error). These scenarios have been chosen to mirror simulations of type I error rates.