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
Available online 30 January 2018
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/

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;

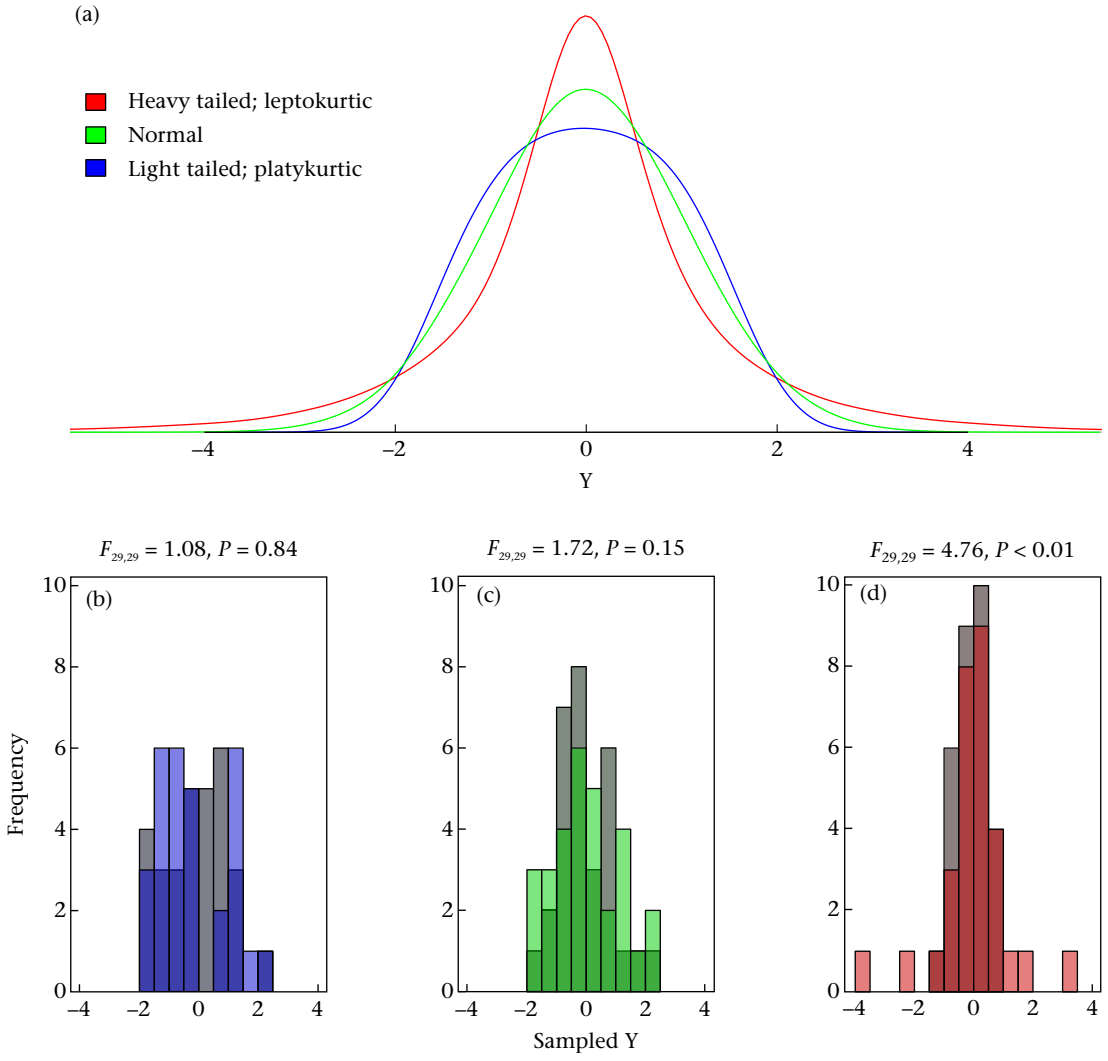

Figure 1. The influence of kurtosis on *F* test comparisons of sample variances. (a) Probability distribution functions of a population's phenotypic measurement 'Y': normal/Gaussian distribution (green); a heavy-tailed distribution (red; kurtosis parameter $\delta = 0.5$) and a light-tailed distribution (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.

hence kurtoses of the underlying distributions have been 'averaged away'. Second, the denominator of ANOVA tests will (usually) have large degrees of freedom that dampen the influence of kurtosis. Perversely though, the use of *F* tests (and their multisample extension, Bartlett's test) to check ANOVA's assumption of homogeneous variance across treatments, remains highly sensitive to departures from normality. To quote Zar (1999, page 204), 'Because of the poor performance of tests for variance homogeneity …. it is not recommended that [they] be performed as tests of the underlying assumptions of [ANOVA]'.

Defenders of the *F* test might cite the availability of statistical tests for the normality of data distributions. However, tests of normality have low power (they incorrectly fail to reject H_o except at very large sample sizes), and it is particularly hard to detect the heavy distribution tails that can have so much influence on both the magnitude of variance and the outcome of any *F* test. Affirmative results of normality tests (e.g. nonsignificant goodness-of-fit tests) should not be used to justify using the *F* test to compare equality of variances (Van Valen, 2005). Basically *F* tests should be avoided, and since Bartlett's test is a generalization of the *F* test to *k* samples, it should also be avoided or at least used with extreme caution.

A COMPARISON OF VARIANCE COMPARISONS

So, what tests are appropriate to use in tests of equality of variance? For univariate tests of absolute variation, Van Valen (2005) recommended three relatively simple and appropriate tests: jackknifing, Smith's test and Levene's test. Miller (1998) did not scrutinize Smith's test, but dissected a selection of robust parametric (including Levene's test and the Box–Anderson test) and nonparametric options.

Here we compare parametric tests (Levene's, Box–Anderson, Smith's) and resampling tests (jackknifing), and to the latter group we append a discussion of bootstrapping and permutation testing. We do not cover nonparametric tests based on ranked data and ranked variances because they either require assumptions of equal medians, throw away data, are not robust or are inefficient (Miller, 1998). Each test we consider has strengths and weaknesses, and they vary in their robustness to the problems that plague *F* testing of variance equality. We hope this comparison helps to guide the choice of tests for biologists wishing to compare sample variances but are suffering from, or simply worried about, non-normality.

PARAMETRIC TESTS

Levene's Test

The most commonly used and simplest of the univariate equality of variance tests is Levene's test. For each sample first find the median (or, if that is not possible, the mean), and then calculate the absolute deviation of each datum from the median ($y_i = |x_i - \text{median}(x)|$). This generates a new variable (y_i = deviance), which increases with increasing variation in the sample. Then calculate the mean and variance of the deviances among samples, and these can be tested for equality by *t* test or an *F* test. This is very straightforward and has been implemented as the Levene Test function in the 'car' package in R (Fox & Weisberg, 2011).

Formally, Levene's test is a test of all the even moments of a distribution rather than just a test of variances, but the test is dominated by the effect of the variance and is robust in that sense. It has been recommended that for very long-tailed symmetrical distributions, the 10% of data in either tail can be removed before testing. However, Van Valen (2005) suggested that removal of biologically important data is hardly ever justified for the small increase in the precision of estimates that this procedure generates. The test is conservative, but only just so for all but the heaviest-tailed distributions (type I errors lie below, but not far below, the critical threshold of 0.05; Fig. 2) and is robust even to extreme changes in skew and (pertinently, as the next even moment) kurtosis. Levene's test ranks among the most powerful of the tests compared here, at all sample sizes (Figs 3–5).

Box–Anderson Test

Box and Anderson (1955) developed an approximately robust test, based on permutation theory, which is discussed in Miller's (1998) review of variance comparisons. The test scales the numerator and denominator degrees of freedom of the standard *F* test, to better match the theoretical variances under the normal distribution and those under the permutation distribution. The significance of the *F* ratio should be judged based on degrees of freedom $df1 = \widehat{d}(N_1 - 1)$ and $df2 = \widehat{d}(N_2 - 1)$ where $\widehat{d} = \left(1 + \frac{\widehat{b}_2 - 3}{2}\right)^{-1}$ and $\widehat{b}_2 = \frac{(\sum_{i=1}^{2} N_i)(\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^4)}{(\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2)^2}$.

In R, this significance can be queried using pf(statistic, df1, df2). This test satisfies type I error rates of 0.05 for all but the most extreme heavy-tailed distributions, for which it is anticonservative (Fig. 2). It ranks among the most powerful tests of equality of variance (Figs 3–5).

Smith's Test

Smith's test is general, but rarely used even though it is robust and normality is not required (Van Valen, 2005; apparently published only in Grüneberg et al., 1966). It is also the only univariate test that can be used to compare published summaries of variation.

With a sample size of *N*, the variance of the sample variance is given as the square of the standard error of the variance:

$$s^2_{s_j^2} = \frac{\sum_{i=1}^{N}(x_{i,j} - \bar{x})^4 - s_j^4\left(\frac{N_j - 3}{N_j}\right)}{(N_j - 2)(N_j - 3)}.$$

For *k* samples, the following statistic is approximately χ^2-distributed with *k*-1 degrees of freedom:

$$\chi^2_{k-1} = \sum_{j=1}^{k}\frac{s_j^4}{s^2_{s_j^2}} - \frac{\left(\sum_{j=1}^{k}\frac{s_j^2}{s^2_{s_j^2}}\right)^2}{\sum_{j=1}^{k}\frac{1}{s^2_{s_j^2}}},$$

and the significance of this statistic can be assessed using tables of significance or by querying the cumulative distribution function (e.g. using pchisq(statistic, df) in software R; R Core Team, 2016). Our simulations show that Smith's test is hardly affected by even the most extreme skews and kurtoses, but is extremely conservative, delivering type I error (rejection of a true null: a false positive) rates consistently and dramatically less than 5% (i.e. type I errors lie well below the critical threshold of 0.05; Fig. 2). It is not commonly used in any of the empirical sciences, and this super-conservatism also yields low power to detect real differences (Figs 3–5; spectacularly low power with sample size $N = 10$), which will probably not improve its popularity.

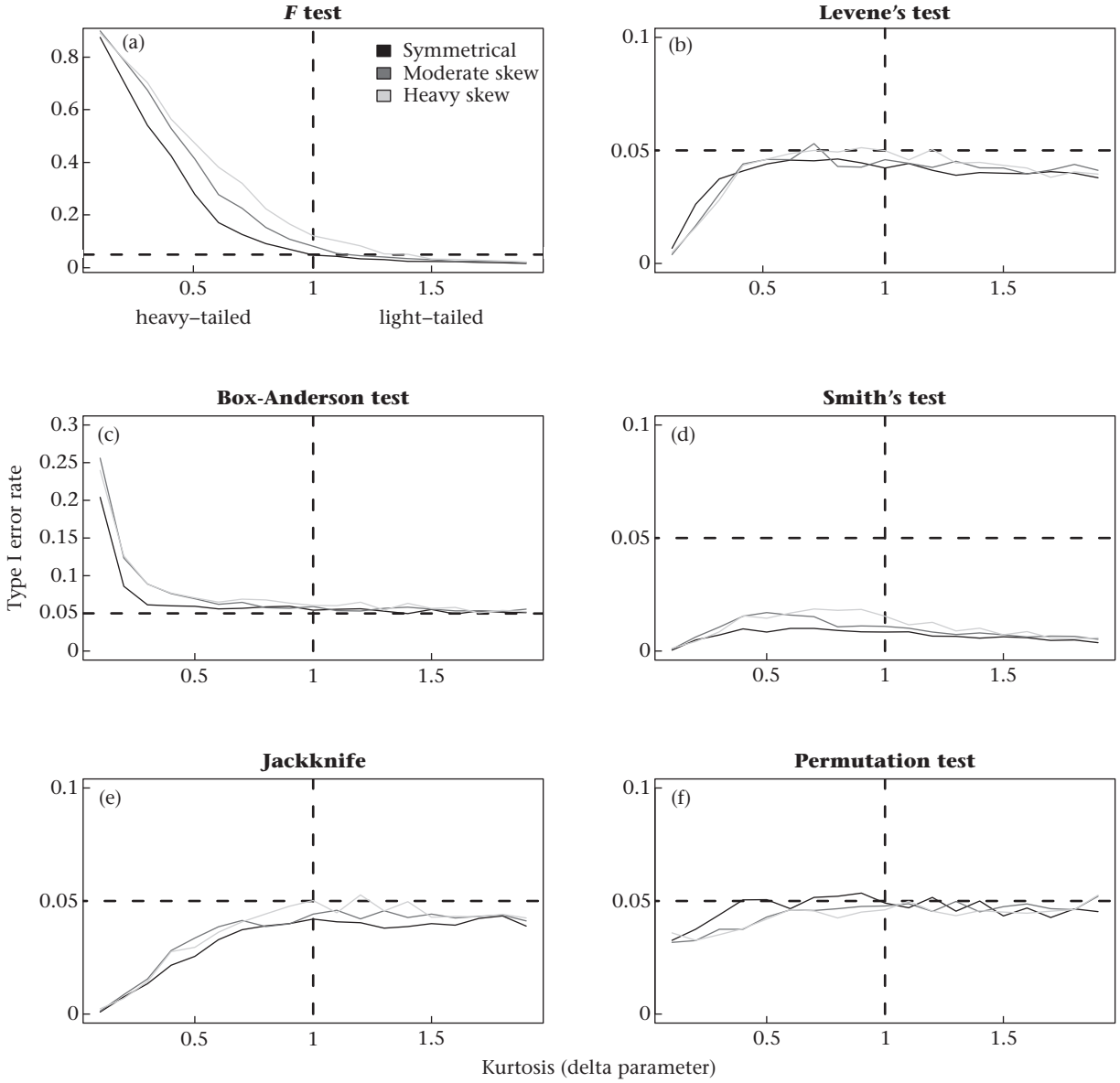

Figure 2. Rates of false positive conclusions from tests of the equality of variance of samples with $N = 30$, drawn from two populations. (a) F test, (b) Levene's test, (c) Box–Anderson test, (d) Smith's test, (e) jackknife, (f) permutation test. Type I error rates are simulated from identical background populations of the sinh-arcsinh family with mean 0, standard deviation 1 and kurtosis (on the x-axis) defined by the delta parameter (small values = heavy tailed; 1 = normal; large values = light tailed). Line shadings represent different skews, described by the epsilon parameter: black = symmetrical (epsilon = 1); mid-grey = moderate skew (epsilon = 0.5); light grey = heavy skew (epsilon = 1.5). Dashed vertical lines mark a symmetrical (normal) distribution (kurtosis = 1); dashed horizontal lines mark an error rate of 0.05 which is our convention for accepted likelihood of falsely rejecting null ($P = 0.05$). Well-behaved tests converge on a type I error rate of 0.05.

RESAMPLING TESTS

The Bootstrap

One method often used in testing equality of variances is the bootstrap (random sampling with replacement). This is one of a family of randomization techniques that has become commonplace with the advent of the desktop computer. However, some bootstrap methods are poor, nonrobust performers (Hall & Wilson, 1991) and generally, for very heavy-tailed distributions, the technique is prone to providing incorrect but increasingly well-supported results as sample size increases (Wu, 1988).

The Jackknife

Jackknifing is another randomization technique and is now fairly standard. It requires reasonable sample sizes (>20) and involves dropping one datum at a time and calculating a variance for each group to be tested and for the total variance, until each datum has been dropped in turn. The variance of the variances can then be calculated and since these are distributed as t with N-1 degrees of freedom, they can be compared with t or F tests. The jackknife is robust to skew and to all but the most extreme kurtoses (Fig. 2), is conservative, but more so than Levene's test (i.e. the type 1 error surface is below 0.05). It is relatively powerful at reasonable sample sizes (Figs 3 and 5) but, being based on subsamples of the data, suffers low power at small sample sizes (Fig. 4). However, it is the only test that can provide confidence intervals on variance estimates (also see Bissell & Ferguson, 1975).

Permutation Tests

The final test we consider here, data permutation, is completely data driven, relying entirely on the sample data to consider the

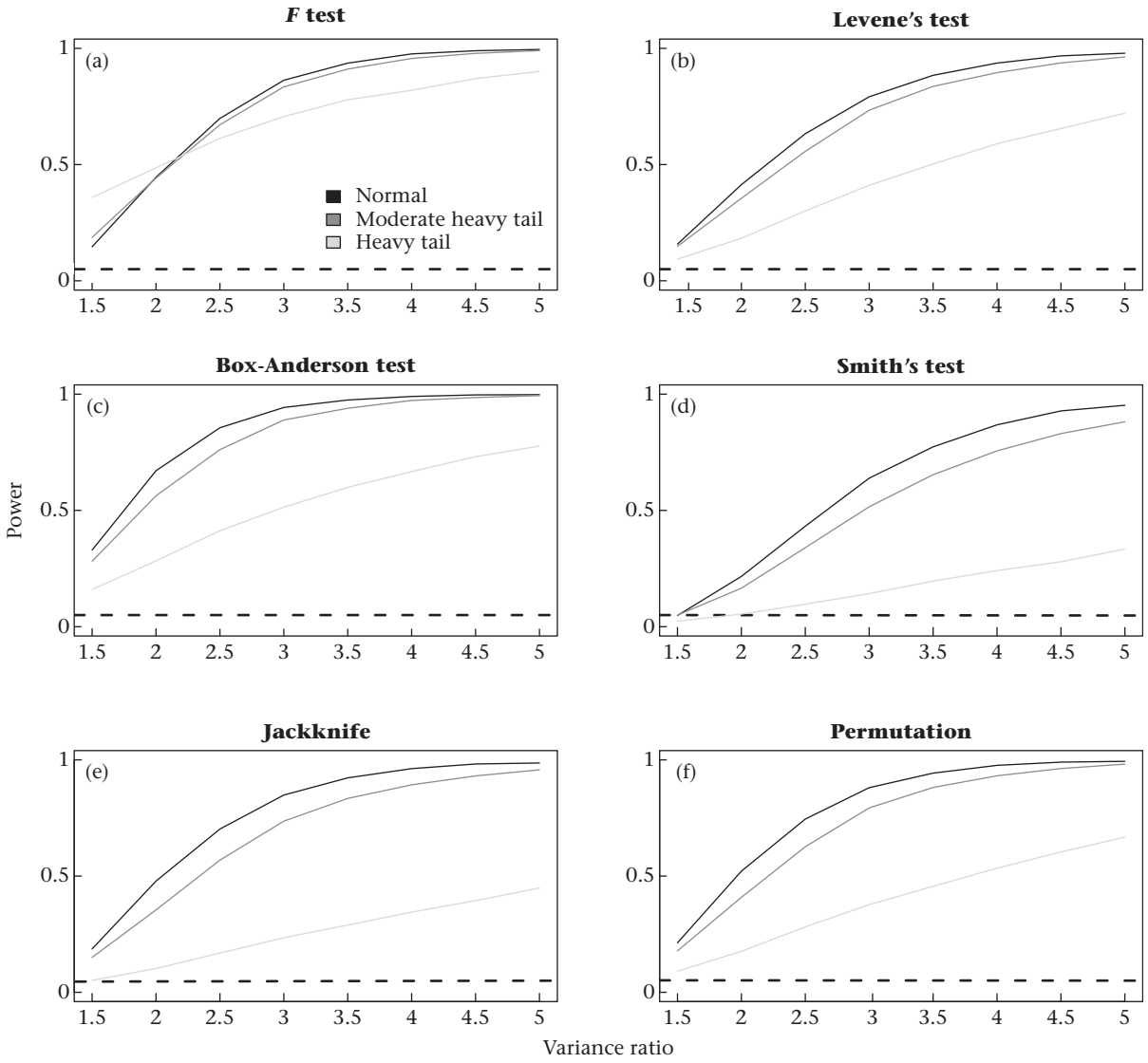

Figure 3. Simulations to determine the power (ability to detect real signal at significance threshold = 0.05) of tests that compare sample variances. (a) F test, (b) Levene's test, (c) Box–Anderson test, (d) Smith's test, (e) jackknife, (f) permutation test. Samples drawn with $N = 30$ from underlying populations following sinh-arcsinh probability distributions, with mean 0, skew parameter 0 and sharing different values of kurtosis parameter delta. For each test, the x-axis changes the variance ratio of the two underlying populations, from 1 to 5. Dashed line shows the threshold type I error rate, which should ideally equal 0.05 for variance ratio = 1 and should be recreated by 'power' simulations at this variance ratio. Line shadings: black = normal (delta = 1); mid-grey = moderately heavy tailed (delta = 0.75); light grey = heavy tailed (delta = 0.5). The 'apparent' high power of the *F* test for variance ratios close to 1 is in fact due to type I error (see Fig. 2). Power trajectories converge to a maximum of 1 with increasing variance ratio.

evidence for or against differences in variance between the two underlying populations. In other words, it requires no distributional assumptions for the test statistic and therefore loses power dramatically at small sample sizes. Data from the two samples are shuffled (sampled without replacement) between two fake samples, and the variance ratio is calculated. This is repeated many times (here, 10K) to create an empirical distribution of variance ratios under the null hypothesis of no difference. The observed variance ratio of the real samples is compared to this null distribution, and significant differences are inferred when this observation lies in the lower or upper 2.5% of the distribution of outcomes. This test therefore uses the variance ratio, which might be called *F*, but it is not an *F* test. Permutation tests are computationally expensive, but for most real-world examples the power of the modern personal computer is more than sufficient. See Rodríguez-Muñoz, Bretman, Slate, Walling, and Tregenza (2010) for an application to sex differences in reproductive variance in a wild insect. The permutation test is robust to skew and kurtosis and, perhaps self-evidently, provides type I error rates of 0.05 or below (Fig. 2). It is powerful at reasonable sample sizes (Figs 3 and 5) but, being based on data shuffles, suffers low power at small sample sizes (Fig. 4). We note, however, that the permutation approach is more powerful than the jackknife at small sample sizes (Fig. 4).

COMPARISON OF FALSE POSITIVES AND POWER

Simulations of Type I Error (False Positive) Rates

For each test described here, including the *F* test of sample variances, we asked, 'how often would we mistakenly conclude different variances when in fact the samples are drawn from the same underlying population?' This is the risk of false positive outcome, or the type I error rate [i.e. Pr(reject $H_0|H_0$ True)]. We simulated populations of 10K measurements drawn from adapted

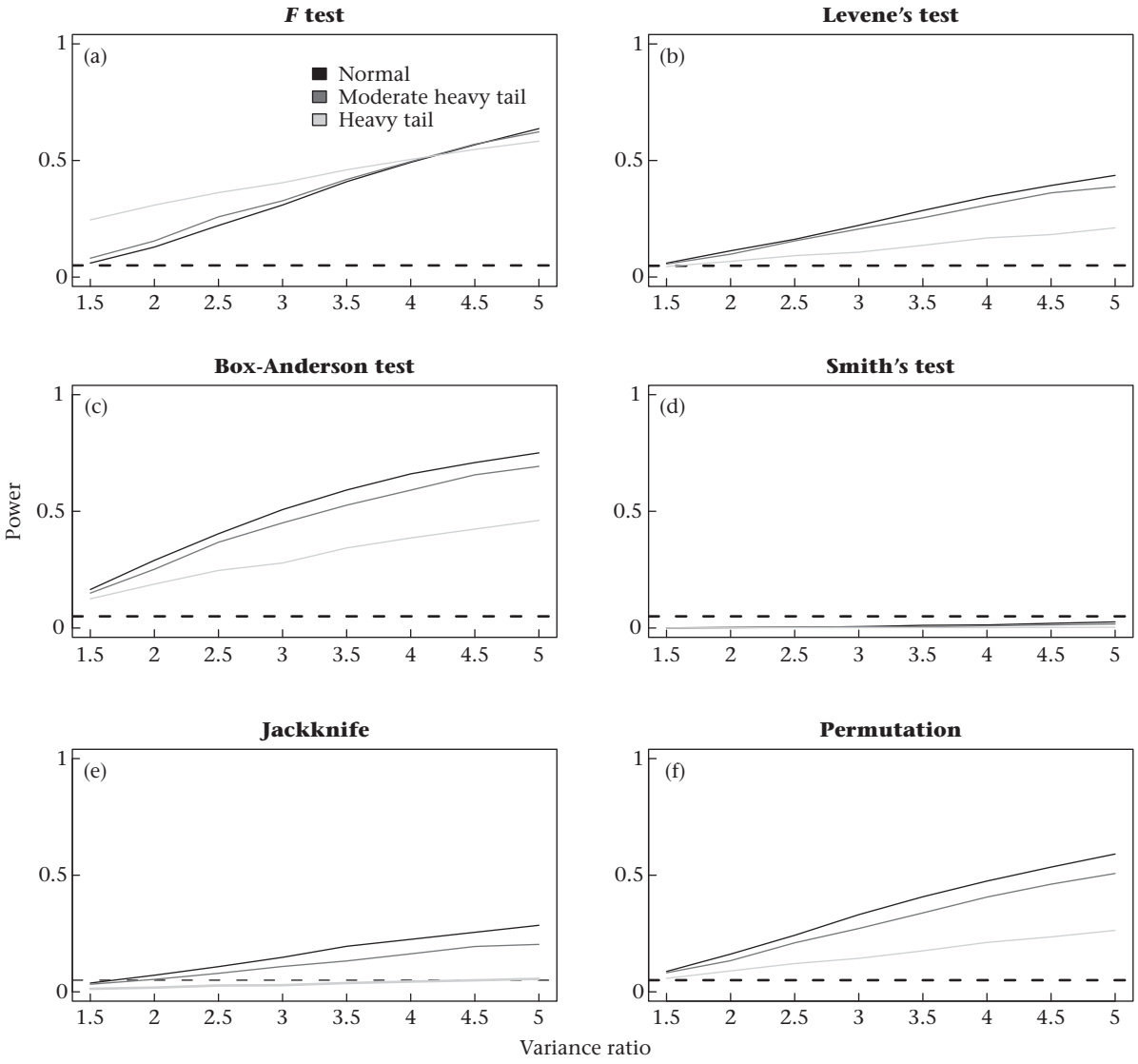

Figure 4. Simulations to determine the power (ability to detect real signal at significance threshold = 0.05) of tests that compare small-sample variances. (a) F test, (b) Levene's test, (c) Box–Anderson test, (d) Smith's test, (e) jackknife, (f) permutation test. Samples drawn as in Fig. 3 but with $N = 10$. Power trajectories fail to converge to 1, across the selected range of variance ratios, because of small sample size.

normal distributions. We used the sinh-arcsinh family of distributions (Jones & Pewsey, 2009) for which skew is manipulated using shape parameter ε (positive values yield long tails above the mode, while negative values yield long tails below the mode), and kurtosis using shape parameter δ (increasing values move from leptokurtic (data clustered around the mode, but heavy tailed) to platykurtic (data spread around the mode, but light tailed) distributions, recreating the normal distribution at $\delta = 1$). We simulated populations factorially across a range of skews and kurtoses, and scaled all populations to have zero mean and unit standard deviation.

$$y \sim N(0,1)$$

$$y^{*} = \sinh\left(\left(\frac{1}{\delta}\right)(\operatorname{arcsinh}(y) + \varepsilon)\right)$$

$$y^{**} = \frac{y^{*} - \mu_{y^{*}}}{\sigma_{y^{*}}}$$

here, y is a sample from the standard normal distribution, y^{*} is its sinh-arcsinh transformation and y^{**} scales the transformed distribution back to zero mean and unit variance.

For each assessment of type I errors, we drew two samples (each with $N = 30$) from the simulated population y^{**}, compared variances, stored the P value of the test, and repeated 10K times. For each simulated population and each test, the type 1 error rate is the proportion of tests deemed significant at a threshold $\alpha = 0.05$. The relative performance of the tests we assess can then be judged by the type I error rate for an underlying normal distribution (ideally = 0.05, and usefully conservative when <0.05), and by the sensitivity of this risk of false positives with changes in skew and kurtosis (Fig. 2). We checked our simulations by confirming that for each combination of δ and ε, the average ratio of the variances of the two samples was one.

Simulations of Power

The second valuable characteristic of a statistical test is its power, i.e. its ability to detect signal when that signal is real. We only analysed power of the tests in relation to changes in kurtosis because all were relatively robust to distributional skew (Fig. 2). For these simulations we drew two samples of $N = 30$ from distributions with mean zero, that shared kurtoses of $\delta = 0.5$ (heavy tailed),

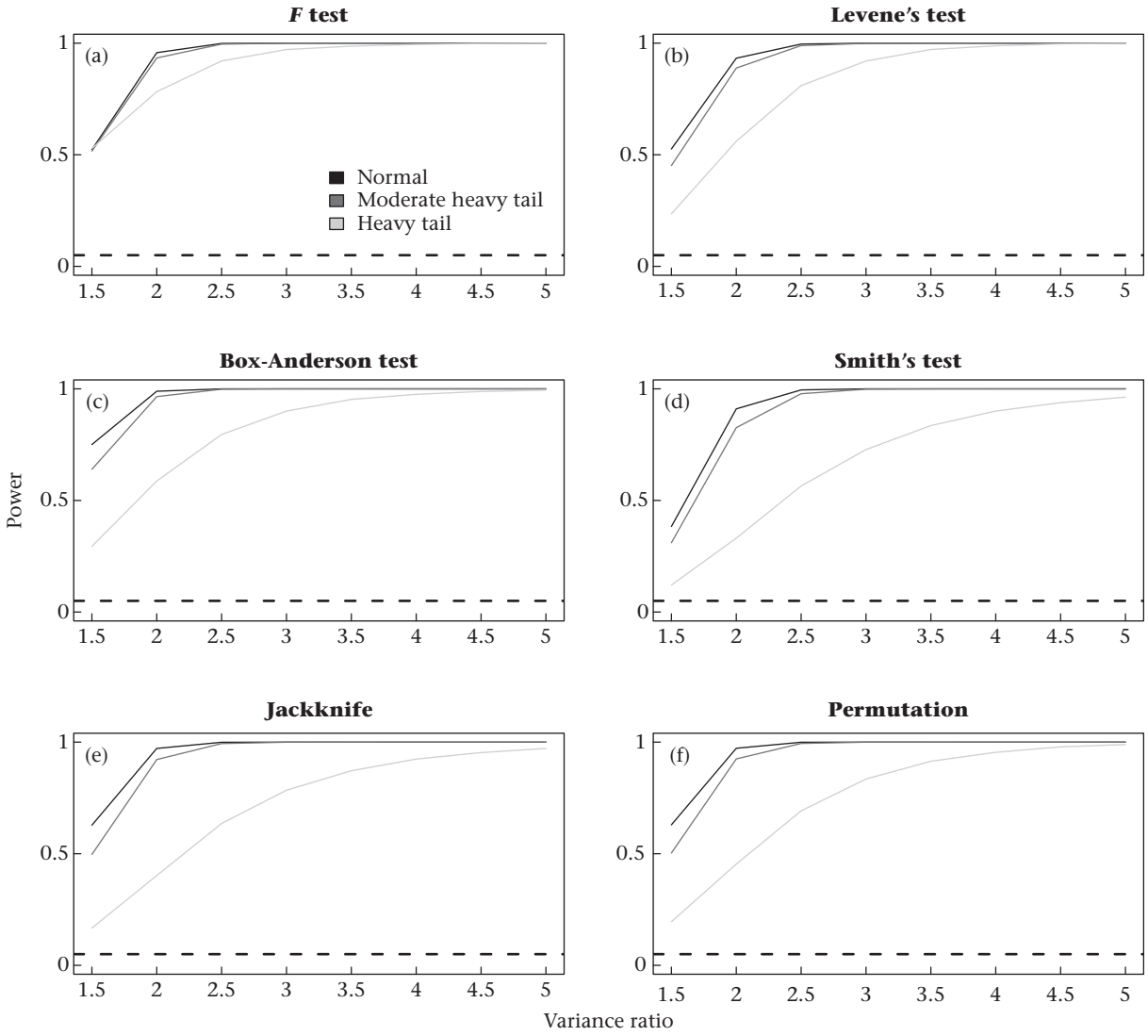

Figure 5. Simulations to determine the power (ability to detect real signal at significance threshold = 0.05) of tests that compare large-sample variances. (a) F test, (b) Levene's test, (c) Box–Anderson test, (d) Smith's test, (e) jackknife, (f) permutation test. Samples drawn as in Fig. 3 but with $N = 100$. Power trajectories converge rapidly to 1 due to large sample sizes.

0.75 (moderately heavy tailed) or 1 (normal), but whose variances increased in ratio from 1 to 5. Using 10K simulations of each parameter combination, we measured power as the probability of detection of these real variance ratios. This is the complement of the type II error rate (power = 1 − Pr(false negative)). Somewhat confusingly, tests can provide what appears to be high power when signal is weak: this is in fact a consequence of high type I error rates (see the apparent power of the *F* test in Fig. 3, related to its high type I error rate in Fig. 2). We therefore require a test that has a type I error rate of 0.05 at a variance ratio of 1, but whose ability to detect genuine signal increases rapidly as the variance ratio moves away from 1. We repeated these power analyses for small sample sizes ($N = 10$, Fig. 4) and large sample sizes ($N = 100$, Fig. 5).

Comparison of False Positives and Power

Our analyses, summarized in Figs 2–5, bring together a set of considerations of test specificity and sensitivity from the statistical literature of several decades ago (e.g. Miller, 1968; Shorack, 1969; reviewed in Van Valen, 1978, 2005; Miller, 1998). Our main point is that the *F* test, although apparently powerful to detect real differences in variance, is indeed highly anticonservative (i.e. type I error (falsely rejecting H_o) is high) with even small deviations in kurtosis from the normal distribution, and while less sensitive to skew, deviations in this moment also reduce the test's usefulness (Fig. 2, F test). To reiterate and emphasize our starting position, if the experimenter or analyst is ever in any doubt about the assumption of normality, the *F* test should be avoided for the testing of equality of variances.

The remaining tests have strengths and weaknesses. We suggest Smith's test is not a viable alternative to the *F* test because of its extreme conservatism (i.e. type I error rates are much lower than 0.05). The permutation test is immune to kurtosis and skews when considering type I errors, but like the jackknife, has low power (fails to reject H_o when H_o is false). This lack of power is further exaggerated at small sample sizes, because the tests are driven by the data themselves and rely on resampling, but the permutation test trumps the jackknife for power when $N = 10$ (Fig. 4).

This leaves two rivals for the crown of 'best test of equality of variances': Levene's test and the Box–Anderson test. Levene's test is favoured by its conservatism at all values of skew and kurtosis. The Box–Anderson test is the most powerful at all sample sizes, but only just so, and this power comes at a cost of anticonservatism for extremely heavy-tailed distributions.

A final point worthy of note is that power declines with increasingly heavy-tailed distributions, whatever test is chosen. Differences in dispersion of heavy-tailed distributions are simply very hard to detect.

WHO CARES?

We have chosen not to name or shame those who have used the *F* test for equality of variances. Many examples of its misuse are caught in time by referees during peer review. However, errors do slip through the peer review net, and some of these are recent and include papers in *Animal Behaviour*. Examples of misuse fall into two camps: (1) studies whose hypotheses relate directly to the comparison of two or more variances; and (2) studies that use *F* tests or Bartlett's test to test homogeneity of variance as an assumption of ANOVA. '*F* test equality of variance' is difficult to search for using bibliographic search engines, because of the vast number of hits for studies using ANOVA or hierarchical variance partitioning. However, a quick search of Google Scholar using the keywords 'variance-ratio Animal Behaviour' revealed 15 examples from the first camp within the first few pages, including six from *Animal Behaviour*. Most of these examples cite Zar (1999), or alternative editions of this classic textbook, to justify their choice of test, despite his repeated warnings about the sensitivity of *F* tests and Bartlett's test to non-normality.

Diagnostic tests of homogeneity of variance are even more prevalent, and raise an interesting slant on our argument. *F* tests risk type I errors for heavy-tailed distributions. A significant *F* test could therefore reveal either that the variances are not homogeneous or that the underlying population distribution is heavy tailed. On the other hand, a nonsignificant diagnostic *F* test could reveal either that the underlying populations have similar variance and are not heavy tailed or that there is low power to detect either effect due to small sample size. We recommend much more stringent approaches to the verification of ANOVA's assumptions.

CONCLUSION

Variation is not just one of the fundamental requirements for organic evolution, it is a concept that occupies and unifies many fields of biological investigation. Whether one is interested in viral gene transcription, behavioral repertoires, reproductive skew or elephant parasites, comparing variation can be revealing and important (e.g. Dukas & Real, 1993; Hosken & Blanckenhorn, 1999; Sutherland, 1985). Unfortunately, biologists often compare homogeneity of variances incorrectly. Rather than name and shame here, we thought it would be more helpful to point out this problem, reiterating Van Valen's (1978, 2005) previous discourse, alert biologists to the pitfall and provide simple solutions. Our simulations of type I error rates associated with various tests confirm the sensitivity of *F* test comparisons of variances to deviations from normality, particularly those associated with heavy-tailed data distributions. Overall, Levene's test tends to be the best means of comparing variances. It is robust to deviations from normality, is conservative but not painfully so and is powerful enough to detect signal when signal exists. For sufficiently large sample sizes, permutation tests also seem to be robust and relatively powerful. But whatever you do, when comparing variances, do not use the *F* test.

Author contributions

DHos conceived the idea; DHos and DHod designed the study; DHod performed the simulations; DBuss did bibliographic searches; DHos and DHod wrote the paper. All authors contributed critically to the drafts, declare no conflict of interest and give final approval for publication.

Acknowledgments

We thank Van Valen and Miller for their inspirational previous work on this topic and the referees who helped us clarify the submission significantly. DHod is supported by NERC standard grant NE/L007770/1 and by NERC International Opportunities Fund NE/N006798/1 and DHos by the Leverhulme Trust (RF-2015-001).

References

Bissell, A. F., & Ferguson, R. A. (1975). The jackknife – toy, tool or two edged weapon? *Statistician, 24*, 79–100. https://doi.org/10.2307/2987663.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40*, 318–335. https://doi.org/10.1093/biomet/40.3-4.318.

Box, G. E. P., & Anderson, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B, 17*, 1–26.

Brown, A. L., & Robinson, B. W. (2016). Variation in behavioral plasticity regulates consistent individual differences in *Enallagama* damselfly larvae. *Animal Behaviour, 112*, 63–73. https://doi.org/10.1016/j.anbehav.2015.11.018.

Craft, B. B. (2016). Risk sensitive foraging: Changes in choice due to reward quality and delay. *Animal Behaviour, 111*, 41–47. https://doi.org/10.1016/j.anbehav.2015.09.030.

Dukas, R., & Real, L. A. (1993). Effects of nectar variance on learning by bumble bees. *Animal Behaviour, 45*, 37–41. https://doi.org/10.1006/anbe.1993.1004.

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Grüneberg, H., Bains, G. S., Berry, R. J., Riles, L., Smith, C. A. B., & Weiss, R. A. (1966). *A search for genetic effects of high natural radioactivity in South India*. London, U.K: Her Majesty's Stationery Office.

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics, 47*, 451–454. https://doi.org/10.2307/2532163.

Hosken, D. J. (2001). Size and fluctuating asymmetry in sexually selected traits. *Animal Behaviour, 62*, 603–605. https://doi.org/10.1006/anbe.2001.1809.

Hosken, D. J., & Blanckenhorn, W. U. (1999). Female multiple mating, inbreeding avoidance and fitness: It is not only the magnitude of the costs and benefits that counts. *Behavioral Ecology, 10*, 462–464. https://doi.org/10.1093/beheco/10.4.462.

Jones, M. C., & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika, 96*, 761–780. https://doi.org/10.1093/biomet/asp053.

Krebs, J. R., & Davies, N. B. (1978). *Behavioral ecology: An evolutionary approach*. Oxford, U.K: Blackwells.

Krebs, J. R., & Davies, N. B. (1997). *Behavioral ecology: An evolutionary approach* (4th ed.). Oxford, U.K: Blackwells.

MacLeod, K. J., & Clutton Brock, T. H. (2013). No evidence for adaptive sex ratio variation in the cooperatively breeding meerkat *Suricata suricatta*. *Animal Behaviour, 85*, 645–653. https://doi.org/10.1016/j.anbehav.2012.12.028.

Miller, R. G., Jr. (1968). Jackknifing variances. *Annals of Mathematical Statistics, 39*, 567–582. https://doi.org/10.1214/aoms/1177698418.

Miller, R. G., Jr. (1998). *Beyond ANOVA: Basics of applied statistics*. Boca Raton, FL: Chapman & Hall.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rodríguez-Muñoz, R., Bretman, A., Slate, J., Walling, C. A., & Tregenza, T. (2010). Natural and sexual selection in a wild insect population. *Science, 328*, 1269–1272. https://doi.org/10.1126/science.1188102.

Shafir, S., Menda, G., & Smith, B. H. (2005). Caste-specific differences in risk sensitivity in honeybees, Apis mellifera. *Animal Behaviour, 69*, 859–868. https://doi.org/10.1016/j.anbehav.2004.07.011.

Shorack, G. R. (1969). Testing and estimating ratios of scale parameters. *Journal of the American Statistical Association, 64*, 999–1013. https://doi.org/10.1080/01621459.1969.10501032.

Sutherland, W. J. (1985). Chance can produce a sex difference in variance in mating success and explain Bateman's data. *Animal Behaviour, 33*, 1349–1352. https://doi.org/10.1016/S0003-3472(85)80197-4.

Van Valen, L. (1974). Multivariate structural statistics in natural history. *Journal of Theoretical Biology, 45*, 235–247. https://doi.org/10.1016/0022-5193(74)90053-8.

Van Valen, L. (1978). The statistics of variation. *Evolutionary Theory, 4*, 33–43.

Van Valen, L. (2005). The statistics of variation. In B. Hallgrimsson, & B. K. Hall (Eds.), *Variation: A central concept in biology* (pp. 29–48). Burlington, MA: Elsevier Academic Press. https://doi.org/10.1016/B978-012088777-4/50005-3.

Westneat, D. F., & Fox, C. W. (2010). *Evolutionary behavioral ecology*. Oxford, U.K: Oxford University Press.

Wu, C. F. J. (1988). Discussion of the papers by Hinkley and DiCiccio and Romano. *Journal of the Royal Statistics Society B, 50*, 364–365.

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.