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

Running title: Do not use *F*-tests to compare variances

In Press *Animal Behaviour*

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

Centre for Ecology & Conservation

University of Exeter, Cornwall

Penryn TR10 9EZ

*Joint Corresponding Authors. Email d.j.hodgson@exeter.ac.uk; d.j.hosken@exeter.ac.uk.

Abstract.

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 which 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.

Key words: Box-Anderson test, *F*-test, Jackknife, Levene's Test, Normality, permutation, power, variance.

Introduction

*"*Never use an *F*-test to test equality of variances*"* – Van Valen 2005

*"*The effects of nonnormality on the distribution theories for the test statistics…are catastrophic*"* – Miller 1998

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 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 reviewers), and incorrect testing keeps occurring. As part of the continuing battle against inappropriate and anti-conservative (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 (Figure 1). Van Valen (2005) links this sensitivity to violations of the Central Limit Theorem, but Miller (1998) attributes 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 quite insensitive to the data's third moment, skew, but highly sensitive to its fourth, kurtosis (Miller 1998; Figure 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, 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 multi-sample 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), “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. non-significant 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) recommends three relatively simple and appropriate tests: Jackknifing, Smith’s test and Levene’s test. Miller (1998) does not scrutinize Smith’s test, but dissects 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 straight forward and has been implemented as the leveneTest 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) suggests that removal of biological 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, Figure 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 (Figures 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

$df1 = \hat{d}(N_1 - 1)$ and $df2 = \hat{d}(N_2 - 1)$ where $\hat{d} = \left(1 + \frac{\hat{b}_2 - 3}{2}\right)^{-1}$ and

$$\hat{b}_2 = \frac{\left(\sum_{i=1}^{2} N_i\right)\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^4\right)}{\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2\right)^2}.$$

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 anti-conservative (Figure 2). It ranks among the most powerful tests of equality of variance (Figures 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) (Figure 2). It is not commonly used in any of the empirical sciences, and this super-conservatism also yields low power to detect real differences (Figures 3-5; spectacularly low power with sample size *N*=10), which will probably not improve its popularity.

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 common place with the advent of the desktop computer. However, some bootstrap methods are poor, non-robust 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 pretty 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 (Figure 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 (Figures 3 & 5) but, being based on subsamples of the data, suffers low power at small sample sizes (Figure 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 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 et al. (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 (Figure 2). It is powerful at reasonable sample sizes (Figures 3 & 5) but, being based on data shuffles, suffers low power at small sample sizes (Figure 4). We note, however, that the permutation approach is more powerful than the Jackknife at small sample sizes (Figure 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 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 δ=1). We simulated populations factorially across a range of skews and kurtoses, and scaled all populations to have zero mean and unit standard deviation.

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

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

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

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 (Figure 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 analyzed power of the tests in relation to changes in kurtosis because all were relatively robust to distributional skew (Figure 2). For these simulations we drew two samples of N = 30 from distributions with mean zero, that shared kurtoses of δ = 0.5 (heavy tailed), 0.75 (moderately heavy tailed) or 1 (Normal), but whose

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

295 *Comparison of False Positives and Power*

296 Our analyses, summarized in Figures 2-5, bring together a set of
297 considerations of test specificity and sensitivity from the statistical literature of
298 several decades ago (e.g. Miller, 1968; Shorack, 1969; reviewed in Van Valen,
299 1978, 2005; Miller, 1998). Our main point is that the *F*-test, although
300 apparently powerful to detect real differences in variance, is indeed highly
301 anti-conservative (i.e. Type I error (falsely rejecting H_o) is high) with even
302 small deviations in kurtosis from the Normal distribution, and while less
303 sensitive to skew, deviations in this moment also reduce the test's usefulness
304 (Figure 2 F-test). To reiterate and emphasise our starting position, if the
305 experimenter or analyst is ever in any doubt about the assumption of
306 Normality, the *F*-test should be avoided for the testing of equality of variances.

307 The remaining tests have strengths and weaknesses. We suggest Smith's
308 test is not a viable alternative to the *F*-test because of its extreme
309 conservatism (i.e. Type I error rates are much lower than 0.05). The
310 Permutation test is immune to kurtosis and skew when considering Type I
311 errors, but like the Jackknife, has low power (fails to reject H_o when H_o is
312 false). This lack of power is further exaggerated at small sample sizes,
313 because the tests are driven by the data themselves and rely on resampling,
314 but the Permutation test trumps the Jackknife for power when n=10 (Figure 4).

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

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

323 **Who cares?**

324 We have chosen not to name or shame those who have used the *F*-test for
325 equality of variances. Many examples of its misuse are caught in time by
326 referees during peer review. However, errors do slip through the peer review
327 net, and some of these are recent and include papers in *Animal Behaviour*.
328 Examples of misuse fall into two camps: (1) studies whose hypotheses relate
329 directly to the comparison of two or more variances; and (2) studies that use
330 *F*-tests or Bartlett's to test homogeneity of variance as an assumption of

331 ANOVA. "*F*-test equality of variance" is difficult to search for using
332 bibliographic search engines, because of the vast number of hits for studies
333 using ANOVA or hierarchical variance partitioning. However, a quick search of
334 Google Scholar using the keywords "variance-ratio Animal Behaviour"
335 revealed fifteen examples from the first camp within the first few pages,
336 including six from *Animal Behaviour*. Most of these examples cite Zar (1999),
337 or alternative editions of this classic textbook, to justify their choice of test,
338 despite his repeated warnings about the sensitivity of *F*-tests and Bartlett's
339 test to non-Normality.

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

349 **Conclusion**

350 Variation is not just one of the fundamental requirements for organic evolution,
351 it is a concept that occupies and unifies many field of biological investigation.
352 Whether one is interested in viral gene transcription, behavioral repertoires,
353 reproductive skew or elephant parasites, comparing variation can be revealing
354 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, don't 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: The authors 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? *The 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*, Second Edition. Thousand Oaks CA: Sage.

398 Francis, R. I. C. C., & Manly, B. F. (2001). Bootstrap calibrations to improve
399 the reliability of tests to compare sample means and variances.
400 *Envirometrics*, *12*, 713-729. https://doi.org/10.1002/env.494

401 Grüneberg, H., Bains, G. S., Berry, R. J., Riles, L., Smith, C. A. B., & Weiss, R.
402 A. (1966). *A Search for Genetic Effects of High Natural Radioactivity in*
403 *South India*. London: Her Majesty's Stationery Office.

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

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

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

413 Jones, M. C. & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, *96*,
414 761-780. https://doi.org/10.1093/biomet/asp053

415 Krebs, J. R. & Davies, N. B. (1978). Behavioral Ecology: an Evolutionary
416 Approach. Oxford: Blackwells.

417 Krebs, J. R. & Davies, N. B. (1997). Behavioral Ecology: an Evolutionary
418 Approach, 4th Edition. Oxford: Blackwells.

419 MacLeod, K. J. & Clutton Brock, T. H. (2013). No evidence for adaptive sex
420 ratio variation in the cooperatively breeding meerkat *Suricata suricatta.*

421 *Animal Behaviour, 85*, 645-653.
422 https://doi.org/10.1016/j.anbehav.2012.12.028

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

425 Miller, R. G. Jr. (1998). *Beyond ANOVA: Basics of Applied Statistics*.
426 Chapman & Hall, Boca Raton, Florida.

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

430 Rodríguez-Muñoz, R., Bretman, A., Slate, J., Walling, C. A. &Tregenza, T.
431 (2010). Natural and sexual selection in a wild insect population. *Science*,
432 *328*, 1269-1272. https://doi.org/10.1126/science.1188102

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

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

439 Sutherland, W. J. (1985). Chance can produce a sex difference in variance in
440 mating success and explain Bateman's data. *Animal Behaviour*, *33*,
441 1349-1352. https://doi.org/10.1016/S0003-3472(85)80197-4

442 Van Valen, L. (1974). Mutivariate structural statistics in natural history. *Journal*
443 *of Theoretical Biology*, *45*, 235-247. https://doi.org/10.1016/0022-
444 5193(74)90053-8

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

446 Van Valen, L. (2005). The statistics of variation. In: *Variation: A Central*
447 *Concept in Biology* (Eds, B Hallgrimsson & BK Hall), pp 29-48. Elsevier
448 Academic Press, Burlington, MA. https://doi.org/10.1016/B978-
449 012088777-4/50005-3

450 Westneat, D. F. & Fox, C. W. (2010). Evolutionary Behavioral Ecology.
451 Oxford: Oxford University Press.

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

454 Zar, J. H. (1999) *Biostatistical Analysis, Fourth Edition*. Upper Saddle River,
455 New Jersey: Prentice Hall.

Figure Captions

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 δ = 0.5) and a light-tailed distribution (blue; δ = 100). Each distribution has mean zero and standard deviation one. 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 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.

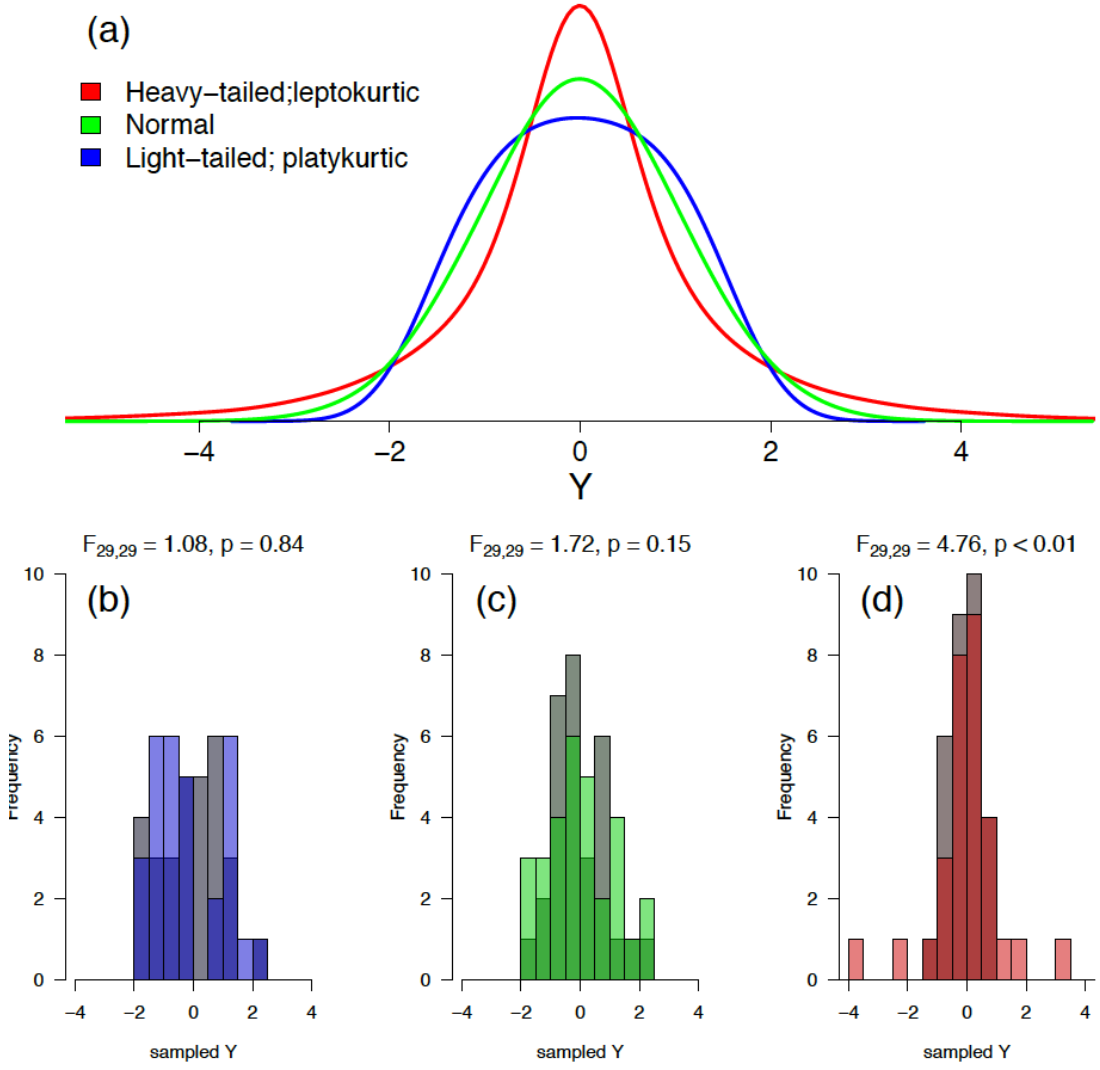
Figure 2. Rates of false positive conclusions from tests of the equality of variance of samples with $N = 30$, drawn from two populations. Type I error rates are simulated from identical background populations of the sinh-arcsinh family with mean 0, standard deviation 1, and kurtosis (on the x-axis) 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 = unskewed (epsilon = 1); mid-grey = moderate

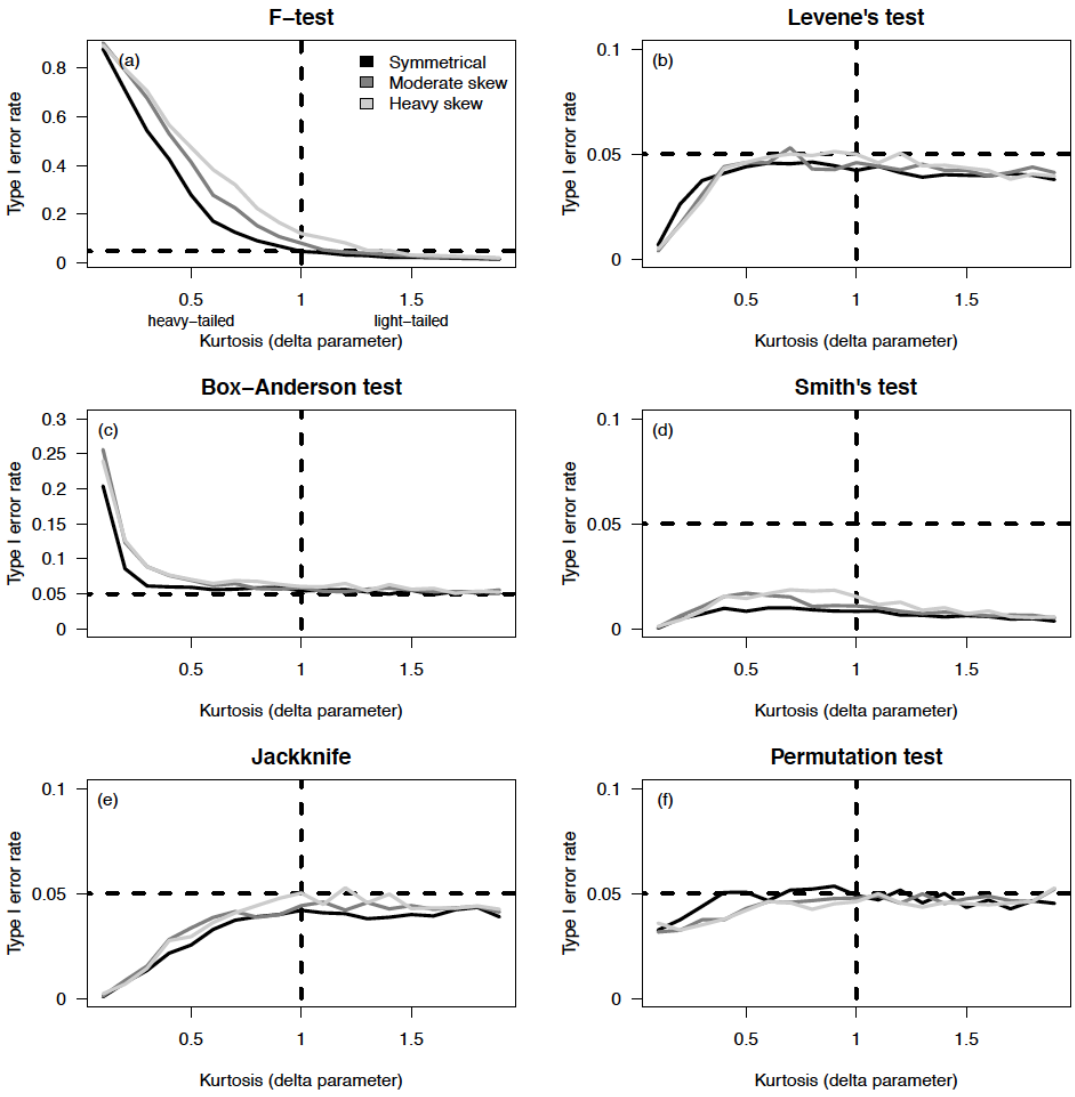
skew (epsilon = 0.5); light-grey = heavy skew (epsilon = 1.5). Well-behaved tests converge on a Type I error rate of 0.05.

Figure 3. Simulations to determine the power (ability to detect real signal at significance threshold = 0.05) of tests that compare sample variances. Samples drawn with *N* = 30 from underlying populations following sinh-arcsinh probability distributions, with mean zero, skew parameter zero, 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 represent: black = Normal (delta = 1); mid-gray = 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 Figure 2). Power trajectories converge to a maximum of 1 with increasing variance ratio.

Figure 4. Simulations to determine the power (ability to detect real signal at significance threshold = 0.05) of tests that compare small-sample variances. Samples drawn as in Figure 3 but with *N* = 10. Power trajectories fail to converge to 1, across the selected range of variance ratios, because of small sample size.

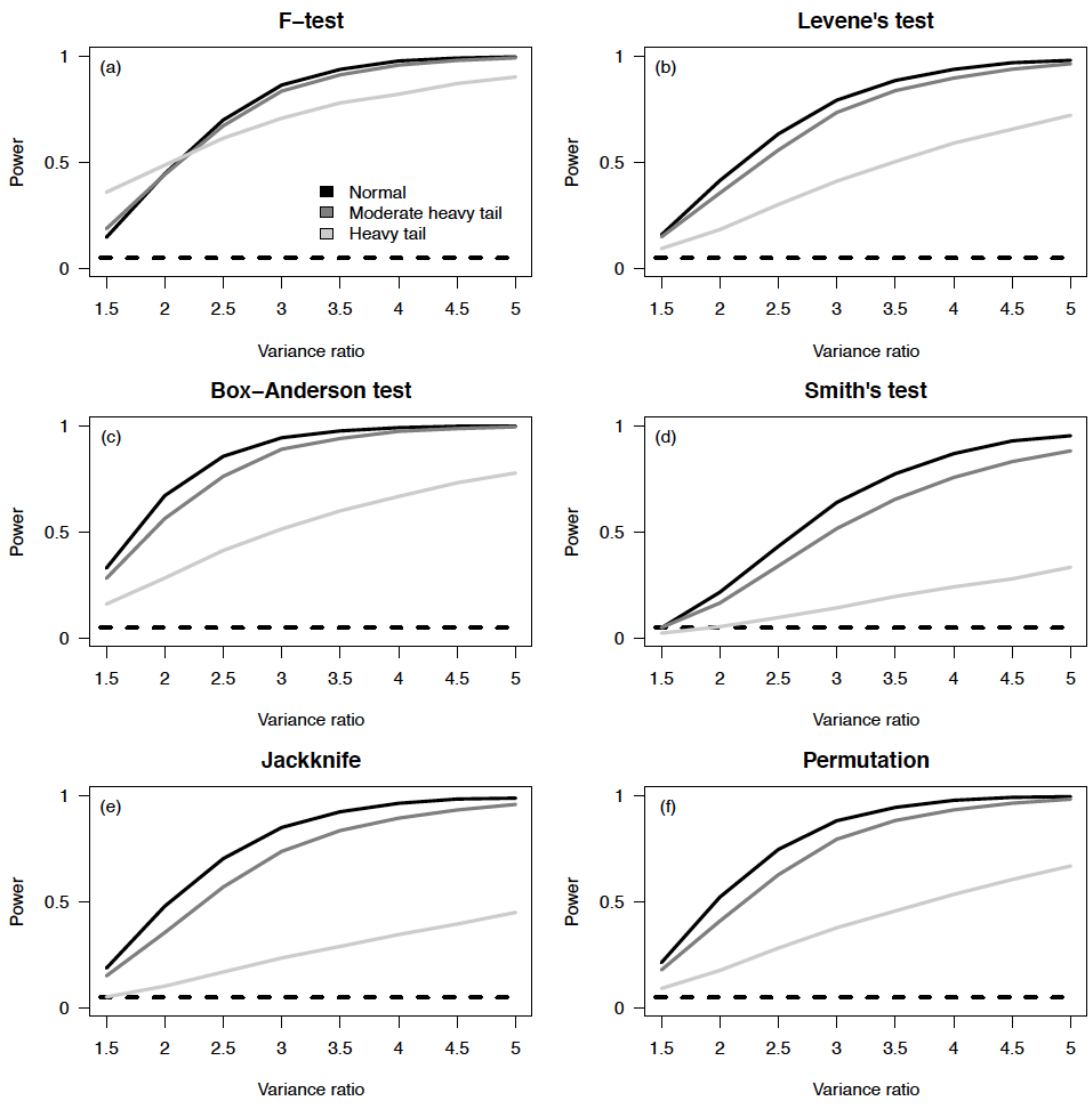
Figure 5. Simulations to determine the power (ability to detect real signal at significance threshold = 0.05) of tests that compare large-sample variances. Samples drawn as in Figure 3 but with *N* = 100. Power trajectories converge rapidly to 1 due to large sample sizes.




F–test
(a)
Symmetrical
Moderate skew
Heavy skew
Type I error rate
0
0.2
0.4
0.6
0.8
0.5
1
1.5
heavy–tailed
light–tailed
Kurtosis (delta parameter)
Levene's test
(b)
0.1
0.05
0
Box–Anderson test
(c)
0.3
0.25
0.2
0.15
Smith's test
(d)
Jackknife
(e)
Permutation test
(f)

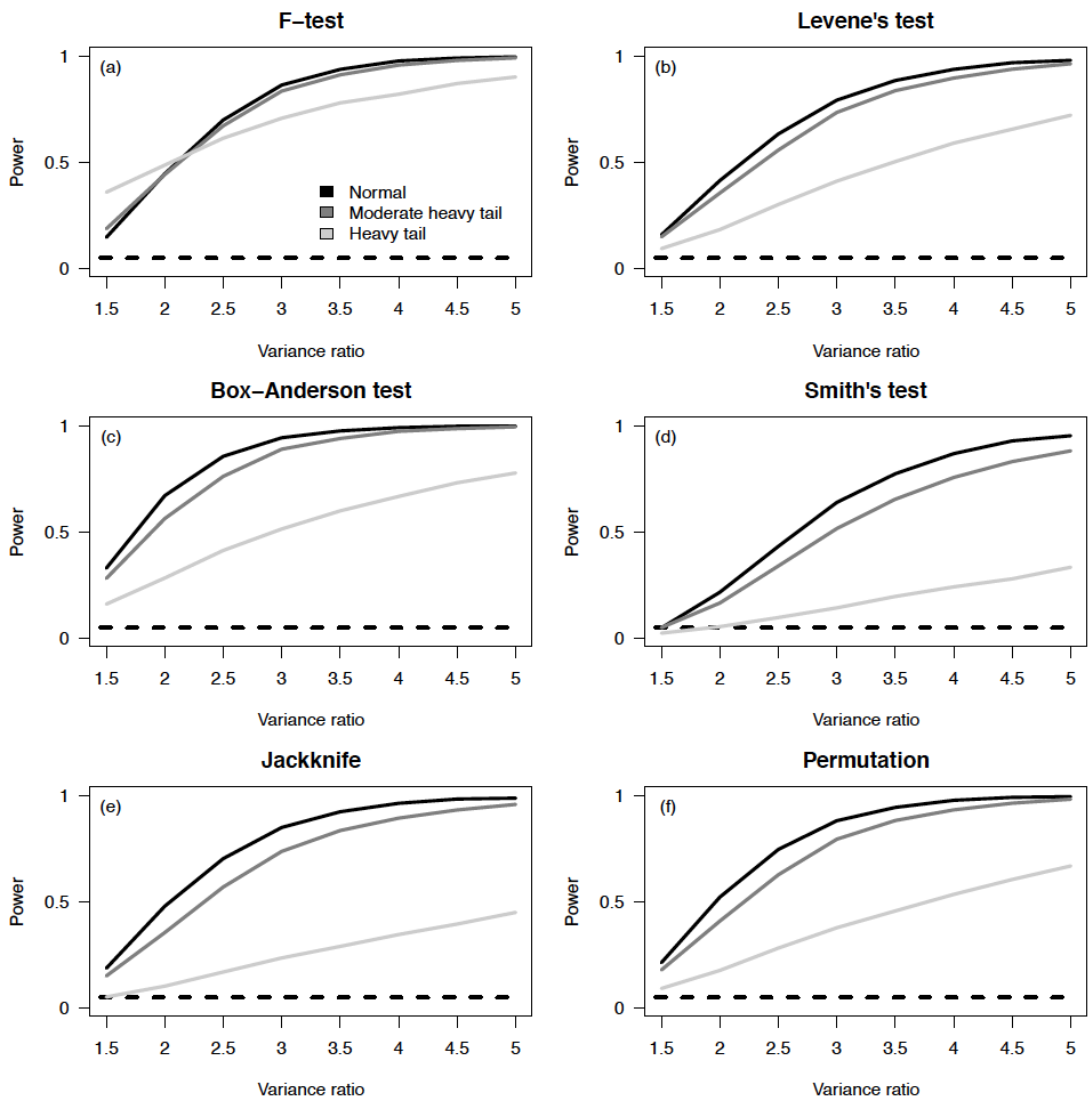
512

513


F−test
(a)
Normal
Moderate heavy tail
Heavy tail
Levene's test
(b)
Box−Anderson test
(c)
Smith's test
(d)
Jackknife
(e)
Permutation
(f)
Power
Variance ratio
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

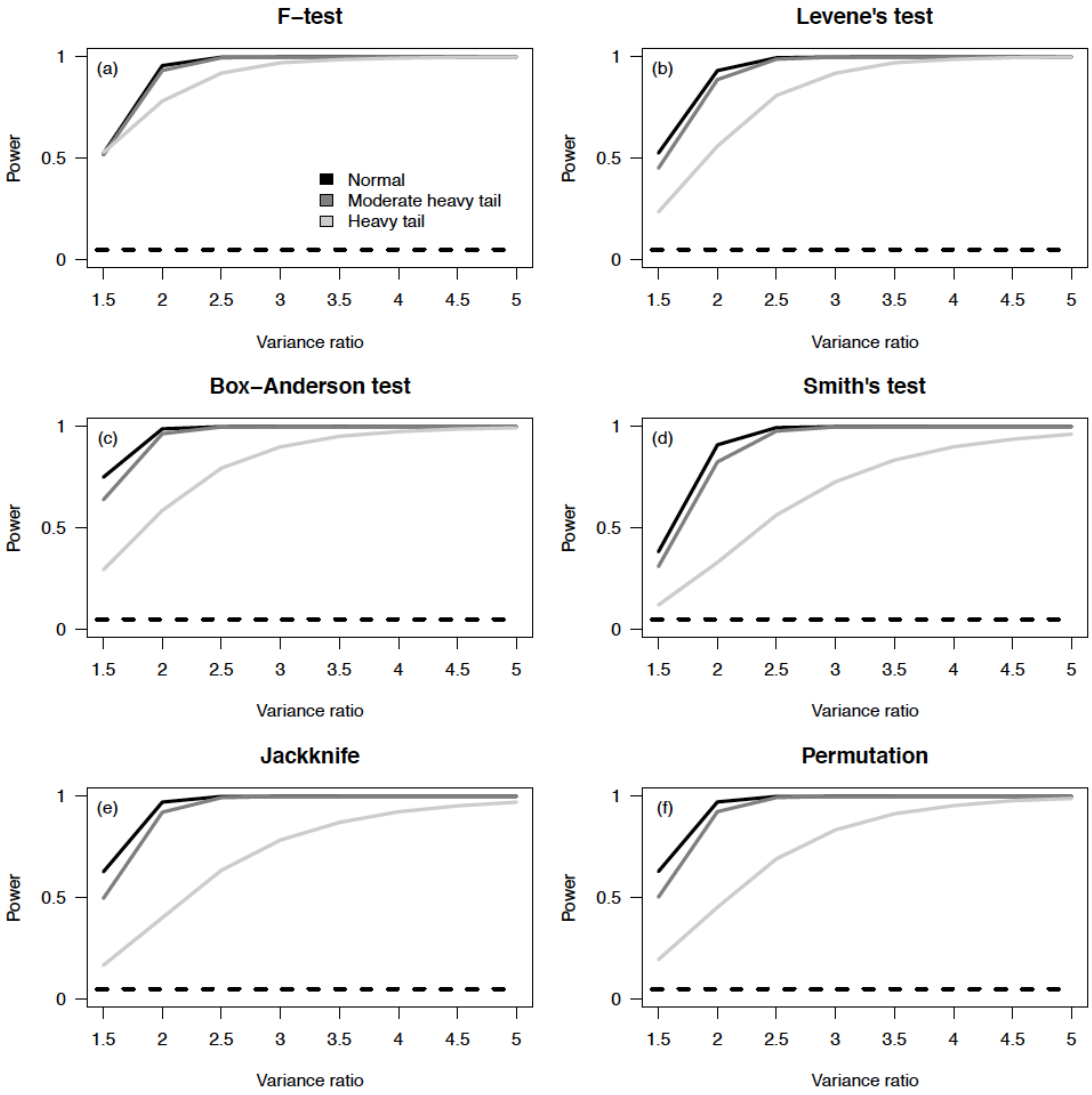
514

515


F−test
(a)
Levene's test
(b)
Box−Anderson test
(c)
Smith's test
(d)
Jackknife
(e)
Permutation
(f)
Normal
Moderate heavy tail
Heavy tail
Power
Variance ratio
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

516

517


F−test
(a)
Levene's test
(b)
Box−Anderson test
(c)
Smith's test
(d)
Jackknife
(e)
Permutation
(f)
Normal
Moderate heavy tail
Heavy tail
Power
Variance ratio
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

518