

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

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

3

4 In Press *Animal Behaviour*

5

6

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

8 Centre for Ecology & Conservation

9 University of Exeter, Cornwall

10 Penryn TR10 9EZ

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

12 d.j.hosken@exeter.ac.uk.

13

14

15 **Abstract.**

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

32

33

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

36 **Introduction**

37 “Never use an *F*-test to test equality of variances” – Van Valen 2005

38 “The effects of nonnormality on the distribution theories for the test
39 statistics...are catastrophic” – Miller 1998

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

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

55 Another common reason to compare sample variances is as a diagnostic
56 check for homogeneity of variance, prior to using ANOVA. Given the
57 importance of the question (“Do the variances differ?”), we seek a statistical
58 test that tells us the probability of detecting the observed signal were the null

59 hypothesis to be true. This P -value is commonly considered “significant” if it
60 lies below the conventional threshold of 0.05. So a test of variances must, if it
61 is to be accurate and effective, satisfy two statistical conditions. First, it should
62 have a low probability of concluding different variances when in fact the
63 samples are drawn from the same underlying population. This is the Type I (or
64 false positive) error rate, and conventionally it should be 0.05. Second, the
65 test should have a high probability of detecting a significant difference when
66 samples are drawn from populations with genuinely different variances. This
67 is called statistical “power”. Inevitably power decreases with decreasing
68 difference in variance between the underlying populations, such that small
69 differences in population variances can be hard to detect.

70 A standard statistical approach, among biologists at least, is to use the F -test
71 to ask whether variance ratios differ significantly from unity. However, as Van
72 Valen (1978; 2005), Miller (1998), and many other statisticians (e.g. Box,
73 1953) have noted, this is inappropriate. Unfortunately, biologists have not
74 heeded warnings from statisticians (as we have noted when serving as both
75 editors and reviewers), and incorrect testing keeps occurring. As part of the
76 continuing battle against inappropriate and anti-conservative (failure to control
77 Type I error) statistical analyses, we reiterate points raised by Van Valen
78 (2005) and Miller (1998) by bringing this issue to the attention of a larger
79 audience. We provide a comparison of statistical tests designed to compare
80 sample variances, and use numerical simulations to demonstrate risks of
81 false-positive and false-negative conclusions with increasingly severe
82 deviations from Normality. We focus on absolute variation in continuous

83 variables, but point readers to Van Valen (1974) for suggestions on discrete
84 variables.

85 Denouncement of the *F*-test might seem rather heretical, given its deep roots
86 in the statistical training of all biologists. The bad news is that *F*-tests of the
87 equality of variances are highly sensitive to deviations from Normality of the
88 underlying data distributions (Figure 1). Van Valen (2005) links this sensitivity
89 to violations of the Central Limit Theorem, but Miller (1998) attributes the
90 problem more properly to a direct mathematical dependence of the variance
91 of the sample variance on the kurtosis of the underlying probability distribution,
92 damped by the sample size. The *F*-test is quite insensitive to the data's third
93 moment, skew, but highly sensitive to its fourth, kurtosis (Miller 1998; Figure
94 1). Kurtosis measures the clustering of data around the mode, relative to
95 variance: leptokurtic distributions have most data clustered tightly around the
96 mode, coupled with very extreme values, and are therefore "heavy-tailed".
97 Platykurtic distributions are less clustered around the mode, coupled with a
98 paucity of extreme values, and are therefore "light-tailed". Heavy-tailed
99 distributions risk very high rates of falsely positive *F*-tests (i.e. Type I error
100 $\gg 0.05$), while light-tailed distributions can yield painfully conservative tests
101 (i.e. Type I error < 0.05). The good news is that *F*-tests used in standard
102 ANOVA are very robust to minor deviations from Normality, for two reasons.
103 First, the numerator of ANOVA tests represents variance among means,
104 hence kurtoses of the underlying distributions have been "averaged away".
105 Second, the denominator of ANOVA tests will (usually) have large degrees of
106 freedom that dampen the influence of kurtosis. Perversely though, the use of
107 *F*-tests (and their multi-sample extension, Bartlett's test) to check ANOVA's

108 assumption of homogeneous variance across treatments, remains highly
109 sensitive to departures from Normality. To quote Zar (1999), “Because of the
110 poor performance of tests for variance homogeneity.... it is not recommended
111 that [they] be performed as tests of the underlying assumptions of [ANOVA].”

112 Defenders of the *F*-test might cite the availability of statistical tests for the
113 Normality of data distributions. However, tests of normality have low power
114 (they incorrectly fail to reject H_0 except at very large sample sizes), and it is
115 particularly hard to detect the heavy distribution tails that can have so much
116 influence on both the magnitude of variance and the outcome of any *F*-test.
117 Affirmative results of Normality tests (e.g. non-significant goodness of fit tests)
118 should not be used to justify using the *F*-test to compare equality of variances
119 (Van Valen, 2005). Basically *F*-tests should be avoided, and since Bartlett’s
120 test is a generalization of the *F*-test to *k* samples, it should also be avoided or
121 at least used with extreme caution.

122

123 **A Comparison of Variance Comparisons**

124 So, what tests are appropriate to use in tests of equality of variance? For
125 univariate tests of absolute variation, Van Valen (2005) recommends three
126 relatively simple and appropriate tests: Jackknifing, Smith’s test and Levene’s
127 test. Miller (1998) does not scrutinize Smith’s test, but dissects a selection of
128 robust parametric (including Levene’s test and the Box-Anderson test) and
129 nonparametric options.

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

140

141 **Parametric Tests**

142 *Levene's test*

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

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

164 *Box-Anderson Test*

165 Box and Anderson (1955) developed an approximately robust test, based on
 166 permutation theory, which is discussed in Miller's (1998) review of variance
 167 comparisons. The test scales the numerator and denominator degrees of
 168 freedom of the standard F -test, to better match the theoretical variances
 169 under the Normal distribution and those under the permutation distribution.
 170 The significance of the F -ratio should be judged based on degrees of freedom

171 $df1 = \hat{d}(N_1 - 1)$ and $df2 = \hat{d}(N_2 - 1)$ where $\hat{d} = \left(1 + \frac{\hat{b}_2 - 3}{2}\right)^{-1}$ and

172
$$\hat{b}_2 = \frac{\left(\sum_{i=1}^2 N_i\right) \left(\sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^4\right)}{\left(\sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2\right)^2}.$$

173 In R, this significance can be queried using `pf(statistic, df1, df2)`. This test
 174 satisfies Type I error rates of 0.05 for all but the most extreme heavy-tailed
 175 distributions, for which it is anti-conservative (Figure 2). It ranks among the
 176 most powerful tests of equality of variance (Figures 3-5).

177

178 *Smith's test.*

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

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

$$185 \quad s_{s_j^2}^2 = \frac{\sum_{i=1}^N (x_{i,j} - \bar{x})^4 - s_j^4 \left(\frac{N_j - 3}{N_j} \right)}{(N_j - 2)(N_j - 3)}.$$

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

$$188 \quad \chi_{k-1}^2 = \sum_{j=1}^k \frac{s_j^4}{s_j^2} - \frac{\left(\sum_{j=1}^k \frac{s_j^2}{s_j^2} \right)^2}{\sum_{j=1}^k \frac{1}{s_j^2}},$$

189 and the significance of this statistic can be assessed using tables of
 190 significance or by querying the cumulative distribution function (e.g. using

191 pchisq(statistic, df) in software *R* (R Core Team, 2016)). Our simulations show
192 that Smith's test is hardly affected by even the most extreme skews and
193 kurtoses, but is extremely conservative, delivering Type I error (rejection of a
194 true null – a false positive) rates consistently and dramatically less than 5%
195 (i.e. Type I errors lie well below the critical threshold of 0.05) (Figure 2). It is
196 not commonly used in any of the empirical sciences, and this super-
197 conservatism also yields low power to detect real differences (Figures 3-5;
198 spectacularly low power with sample size $N=10$), which will probably not
199 improve its popularity.

200

201 **Resampling Tests**

202 *The Bootstrap*

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

210 *The Jackknife*

211 Jackknifing is another randomization technique and is now pretty standard. It
212 requires reasonable sample sizes (>20) and involves dropping one datum at a
213 time and calculating a variance for each group to be tested and for the total

214 variance, until each datum has been dropped in turn. The variance of the
215 variances can then be calculated and since these are distributed as t with $N-1$
216 degrees of freedom, they can be compared with t - or F -tests. The Jackknife is
217 robust to skew and to all but the most extreme kurtoses (Figure 2), is
218 conservative, but more so than Levene's test (i.e. the Type 1 error surface is
219 below 0.05). It is relatively powerful at reasonable sample sizes (Figures 3 &
220 5) but, being based on subsamples of the data, suffers low power at small
221 sample sizes (Figure 4). However, it is the only test that can provide
222 confidence intervals on variance estimates (also see Bissell & Ferguson,
223 1975).

224 *Permutation Tests*

225 The final test we consider here, Data Permutation, is completely data-driven,
226 relying entirely on the sample data to consider the evidence for or against
227 differences in variance between the two underlying populations. In other
228 words, it requires no distributional assumptions for the test statistic and
229 therefore loses power dramatically at small sample sizes. Data from the two
230 samples are shuffled (sampled without replacement) between two fake
231 samples, and the variance ratio is calculated. This is repeated many times
232 (here, 10K) to create an empirical distribution of variance ratios under the null
233 hypothesis of no difference. The observed variance ratio of the real samples
234 is compared to this null distribution, and significant differences are inferred
235 when this observation lies in the lower or upper 2.5% of the distribution of
236 outcomes. This test therefore uses the variance ratio, which might be called F ,
237 but it is not an F -test. Permutation tests are computationally expensive, but for
238 most real-world examples the power of the modern personal computer is

239 more than sufficient. See Rodríguez-Muñoz et al. (2010) for an application to
240 sex differences in reproductive variance in a wild insect. The Permutation Test
241 is robust to skew and kurtosis and, perhaps self-evidently, provides Type I
242 error rates of 0.05 or below (Figure 2). It is powerful at reasonable sample
243 sizes (Figures 3 & 5) but, being based on data shuffles, suffers low power at
244 small sample sizes (Figure 4). We note, however, that the permutation
245 approach is more powerful than the Jackknife at small sample sizes (Figure 4).

246

247 **Comparison of False Positives and Power**

248 *Simulations of Type I Error (false positive) rates*

249 For each test described here, including the *F*-test of sample variances, we
250 asked, “how often would we mistakenly conclude different variances when in
251 fact the samples are drawn from the same underlying population?” This is the
252 risk of false positive outcome, or the Type I error rate [i.e. $\Pr(\text{reject } H_0|H_0$
253 $\text{True})$]. We simulated populations of 10K measurements drawn from adapted
254 Normal distributions. We used the sinh-arcsinh family of distributions (Jones &
255 Pewsey, 2009) for which skew is manipulated using shape parameter ϵ
256 (positive values yield long tails above the mode, while negative values yield
257 long tails below the mode), and kurtosis using shape parameter δ (increasing
258 values move from leptokurtic (data clustered around the mode, but heavy-
259 tailed) to platykurtic (data spread around the mode, but light-tailed)
260 distributions, recreating the Normal distribution at $\delta=1$). We simulated
261 populations factorially across a range of skews and kurtoses, and scaled all
262 populations to have zero mean and unit standard deviation.

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

$$263 \quad y^* = \sinh\left(\left(\frac{1}{\delta}\right)(\operatorname{arcsinh}(y) + \varepsilon)\right)$$

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

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

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

277 *Simulations of Power*

278 The second valuable characteristic of a statistical test is its power, i.e. its
279 ability to detect signal when that signal is real. We only analyzed power of the
280 tests in relation to changes in kurtosis because all were relatively robust to
281 distributional skew (Figure 2). For these simulations we drew two samples of
282 $N = 30$ from distributions with mean zero, that shared kurtoses of $\delta = 0.5$
283 (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_0) 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_0 when H_0 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;

355 Sutherland, 1985). Unfortunately biologists often compare homogeneity of
356 variances incorrectly. Rather than name and shame here, we thought it would
357 be more helpful to point out this problem – reiterating Van Valen’s (1978,
358 2005) previous discourse – alert biologists to the pitfall, and provide simple
359 solutions. Our simulations of Type I error rates associated with various tests
360 confirm the sensitivity of *F*-test comparisons of variances to deviations from
361 Normality, particularly those associated with heavy-tailed data distributions.
362 Overall, Levene’s test tends to be the best means of comparing variances. It
363 is robust to deviations from Normality, is conservative but not painfully so, and
364 is powerful enough to detect signal when signal exists. For sufficiently large
365 sample sizes, Permutation Tests also seem to be robust and relatively
366 powerful. But whatever you do, when comparing variances, don’t use the *F*-
367 test.

368

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

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

378 **References**

- 379 Bissell, A. F., & Ferguson, R. A. (1975). The jackknife – toy, tool or two edged
380 weapon? *The Statistician*, 24, 79-100. <https://doi.org/10.2307/2987663>
- 381 Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40,
382 318-335. <https://doi.org/10.1093/biomet/40.3-4.318>
- 383 Box, G. E. P., & Anderson, S. L. (1955). Permutation theory in the derivation
384 of robust criteria and the study of departures from assumption. *Journal of*
385 *the Royal Statistical Society, Series B*, 17, 1-26.
- 386 Brown, A. L. & Robinson, B. W. (2016). Variation in behavioral plasticity
387 regulates consistent individual differences in *Enallagma* damselfly
388 larvae. *Animal Behaviour*, 112, 63-73.
389 <https://doi.org/10.1016/j.anbehav.2015.11.018>
- 390 Craft, B. B. (2016). Risk sensitive foraging: changes in choice due to reward
391 quality and delay. *Animal Behaviour*, 111, 41-47.
392 <https://doi.org/10.1016/j.anbehav.2015.09.030>
- 393 Dukas, R. & Real, L. A. (1993). Effects of nectar variance on learning by
394 bumble bees. *Animal Behaviour*, 45, 37-41.
395 <https://doi.org/10.1006/anbe.1993.1004>
- 396 Fox, J., & Weisberg, S. (2011). *An {R} Companion to Applied Regression*,
397 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. and 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](https://doi.org/10.1016/S0003-3472(85)80197-4)

- 442 Van Valen, L. (1974). Multivariate structural statistics in natural history. *Journal*
443 *of Theoretical Biology*, 45, 235-247. [https://doi.org/10.1016/0022-](https://doi.org/10.1016/0022-5193(74)90053-8)
444 [5193\(74\)90053-8](https://doi.org/10.1016/0022-5193(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 Hallgrímsson & BK Hall), pp 29-48. Elsevier
448 Academic Press, Burlington, MA. [https://doi.org/10.1016/B978-](https://doi.org/10.1016/B978-012088777-4/50005-3)
449 [012088777-4/50005-3](https://doi.org/10.1016/B978-012088777-4/50005-3)
- 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.

456 **Figure Captions**

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

474 **Figure 2.** Rates of false positive conclusions from tests of the equality of
475 variance of samples with $N = 30$, drawn from two populations. Type I error
476 rates are simulated from identical background populations of the sinh-arcsinh
477 family with mean 0, standard deviation 1, and kurtosis (on the x-axis) defined
478 by the delta parameter (small values = heavy-tailed; 1 = Normal; large values
479 = light-tailed). Line shadings represent different skews, described by the
480 epsilon parameter: black = unskewed (epsilon = 1); mid-grey = moderate

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

483

484 **Figure 3.** Simulations to determine the power (ability to detect real signal at
485 significance threshold = 0.05) of tests that compare sample variances.
486 Samples drawn with $N = 30$ from underlying populations following sinh-arcsinh
487 probability distributions, with mean zero, skew parameter zero, and sharing
488 different values of kurtosis parameter delta. For each test, the x-axis changes
489 the variance ratio of the two underlying populations, from 1 to 5. Dashed line
490 shows the threshold Type I error rate, which should ideally equal 0.05 for
491 variance ratio = 1 and should be recreated by “power” simulations at this
492 variance ratio. Line shadings represent: black = Normal (delta = 1); mid-gray =
493 moderately heavy-tailed (delta = 0.75); light-grey = heavy-tailed (delta = 0.5).
494 The “apparent” high power of the F -test for variance ratios close to 1 is in fact
495 due to Type I error (see Figure 2). Power trajectories converge to a maximum
496 of 1 with increasing variance ratio.

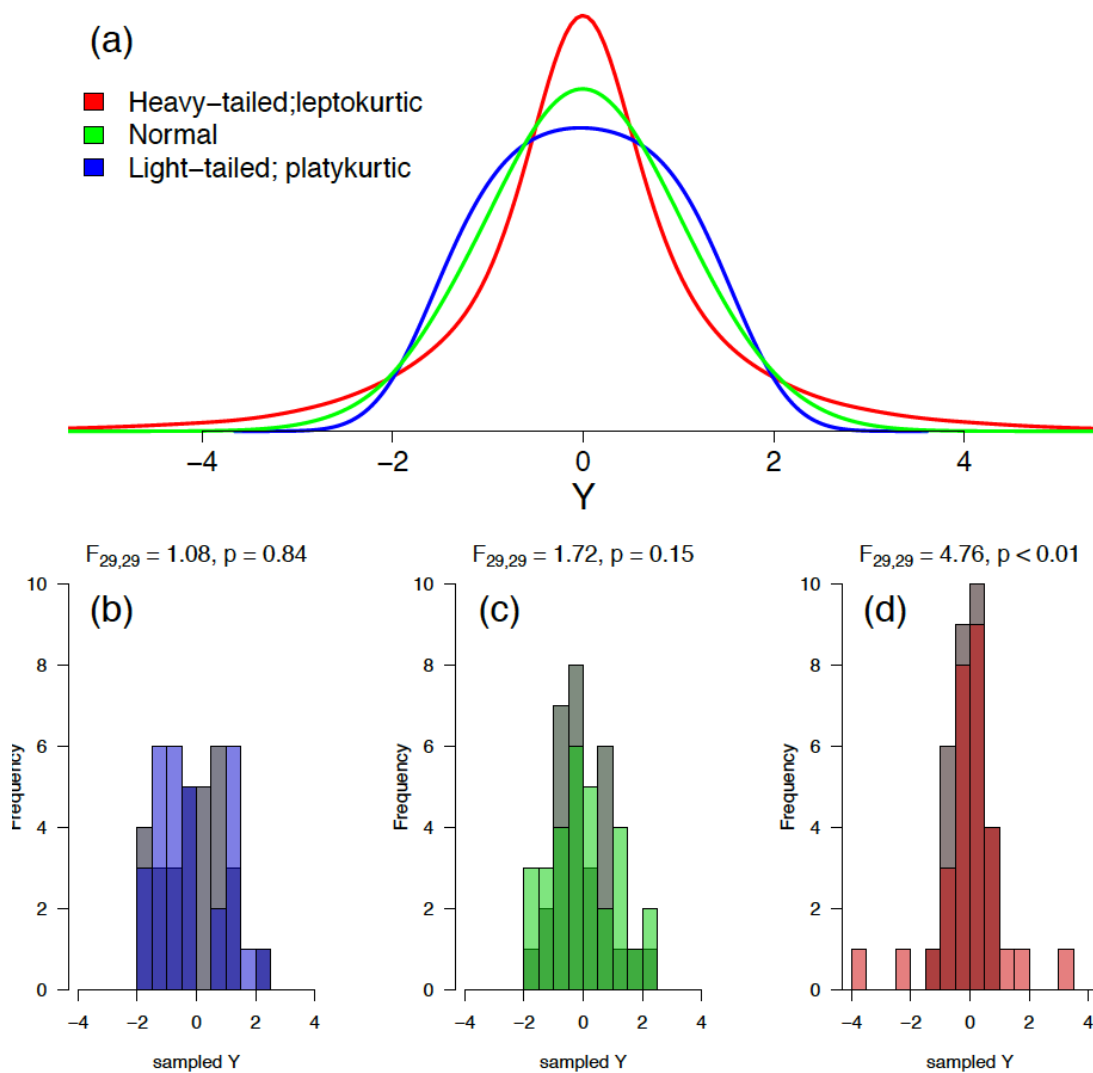
497

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

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

507

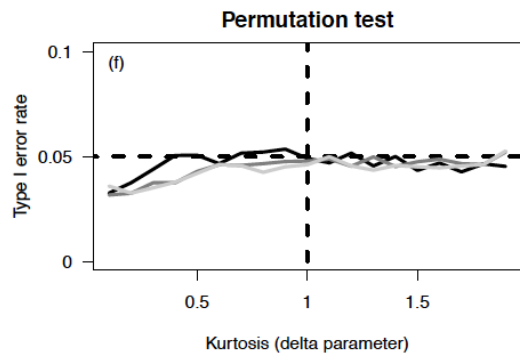
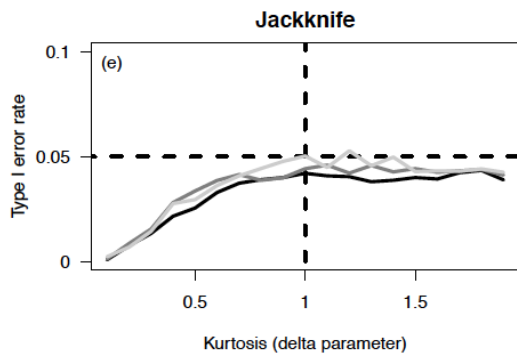
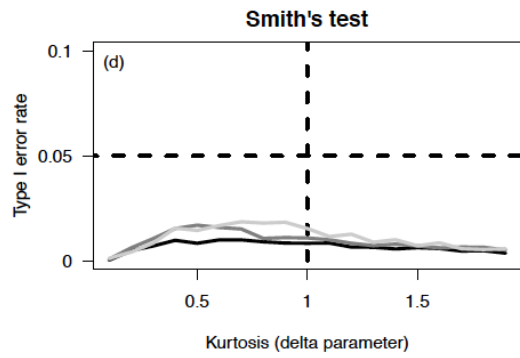
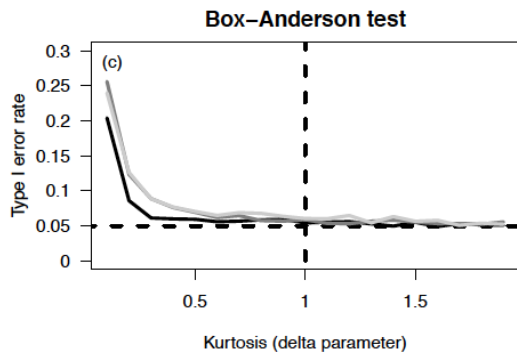
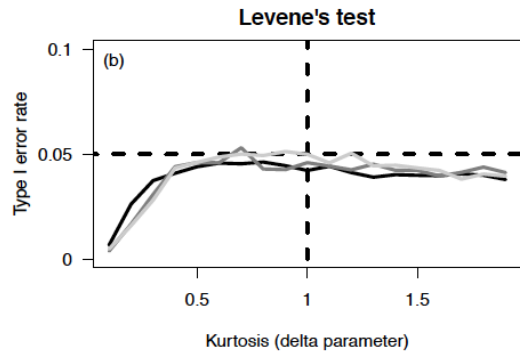
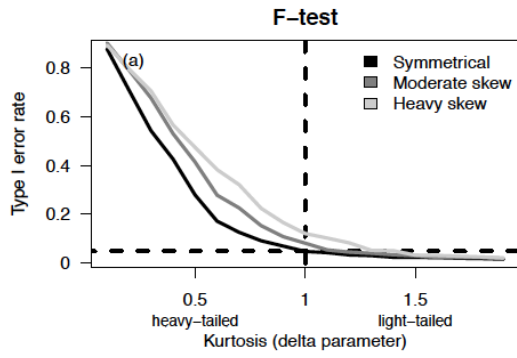
508



509

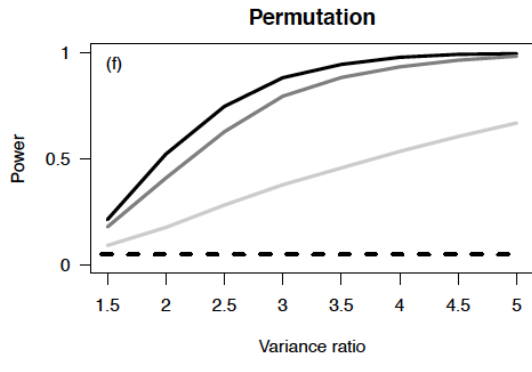
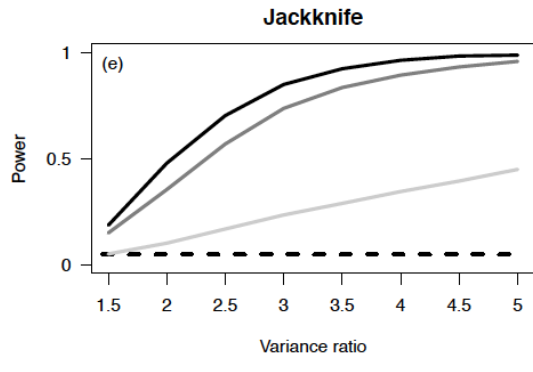
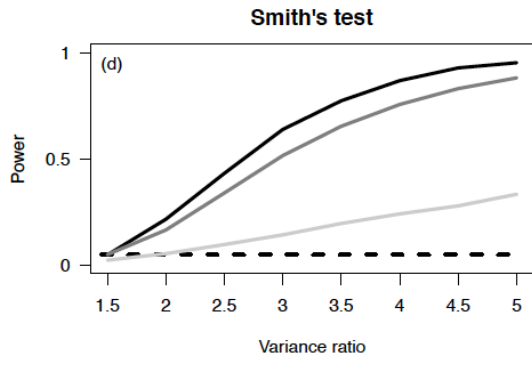
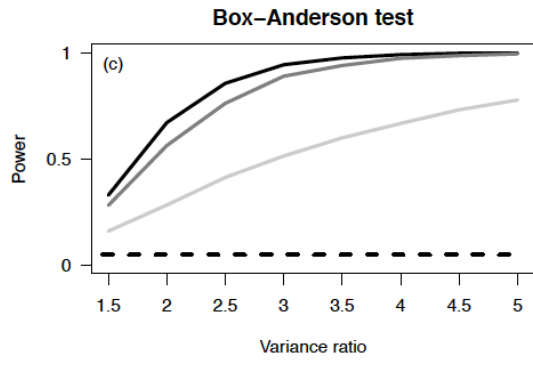
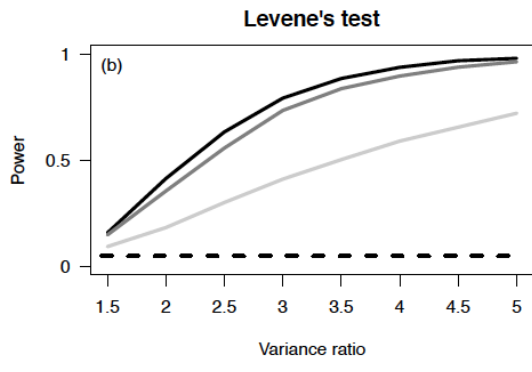
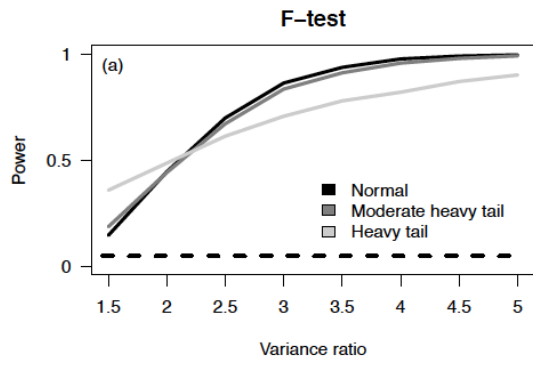
510

511



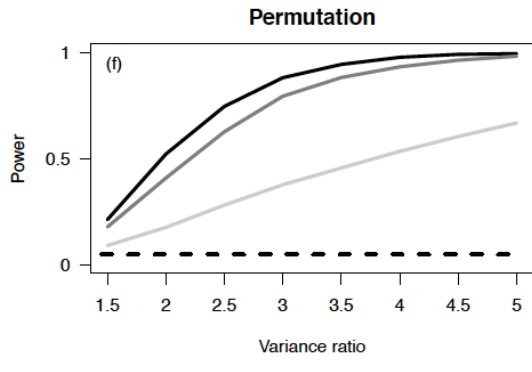
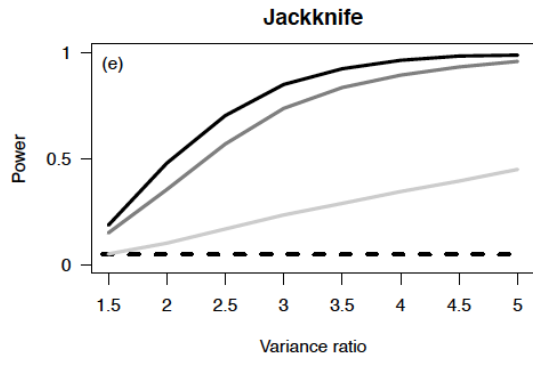
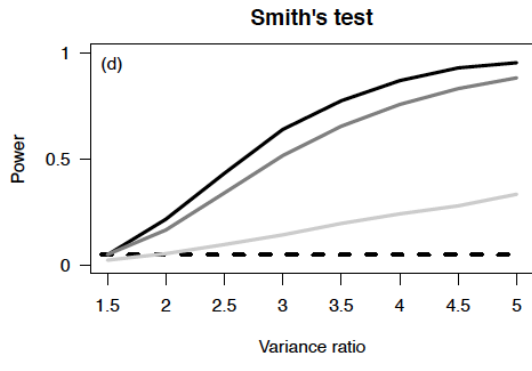
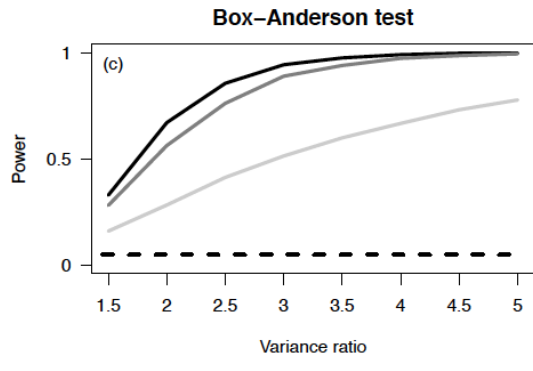
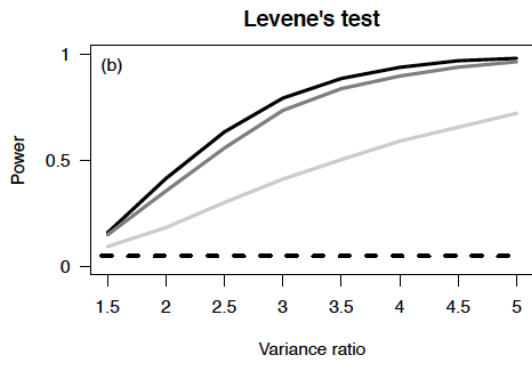
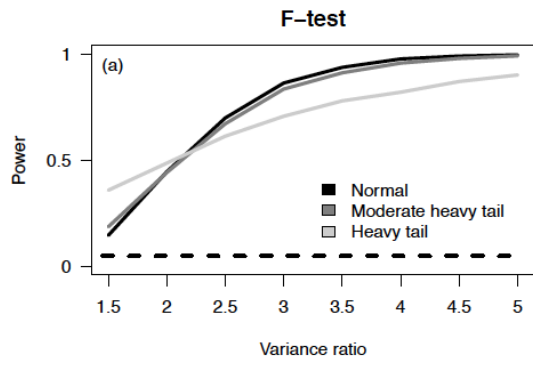
512

513



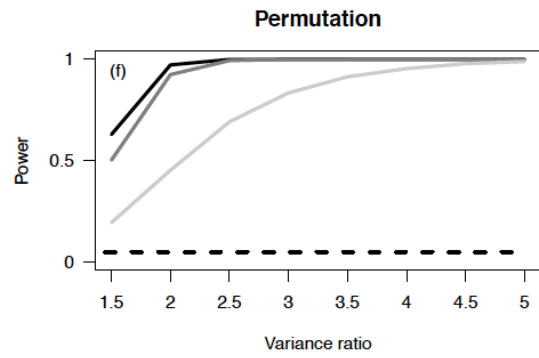
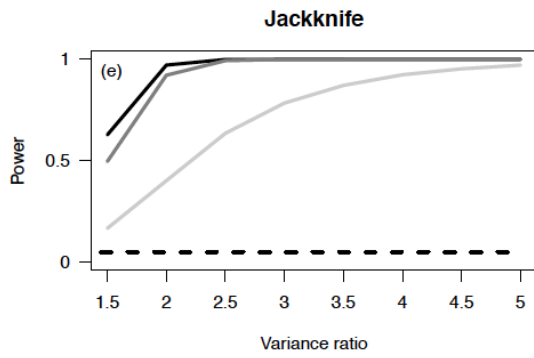
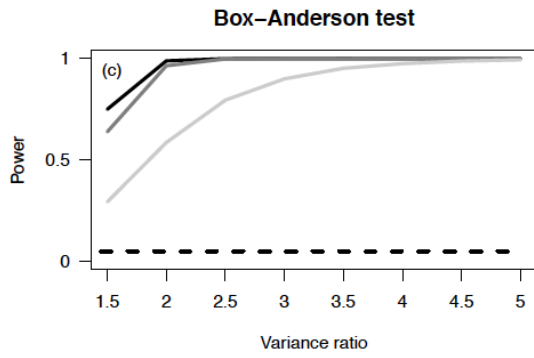
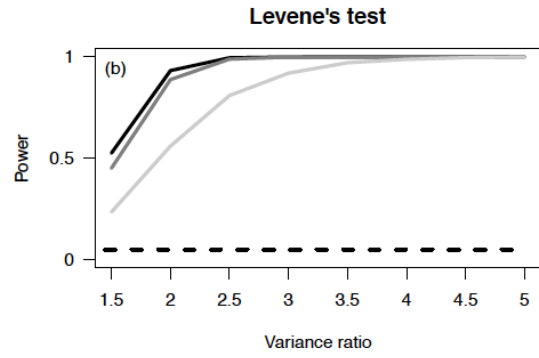
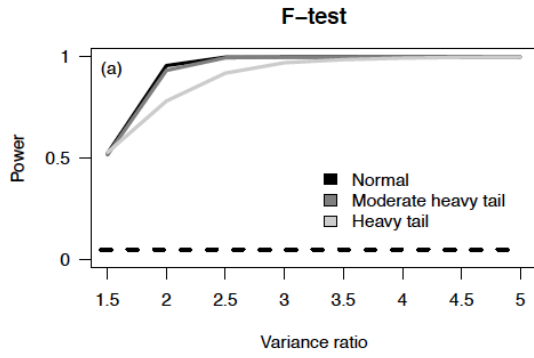
514

515



516

517



518