1   **Advice on testing the null hypothesis that a sample is drawn from a Normal distribution**

2

3   **Graeme D Ruxton[1], David M Wilkinson[2], Markus Neuhäuser[3]**

4

5   1. School of Biology, University of St Andrews, St Andrews KY16 9TH, UK

6   2. Natural Science and Psychology, Liverpool John Moores University, Liverpool, L3 3AF, UK

7   3. Fachbereich Mathematik und Technik, RheinAhrCampus, Koblenz University of Applied
8   Sciences, Joseph-Rovan-Allee 2, 53424 Remagen, Germany

9   Corresponding author: GDR tel +44 1334 654815; fax +44 1334 644801; email gr41@st-

10   andrews.ac.uk

11

12 **Abstract**

13 The Normal distribution remains the most widely-used statistical model, so it is only natural that

14 researchers will frequently be required to consider whether a sample of data appears to have been

15 drawn from a Normal distribution. Commonly-used statistical packages offer a range of alternative

16 formal statistical tests of the null hypothesis of Normality, with inference being drawn on the basis

17 of a calculated p-value. Here we aim to review the statistical literature on the performance of these

18 tests, and briefly survey current usage of them in recently-published papers, with a view to offering

19 advice on good practice. We find that authors in animal behaviour seem to be using such testing

20 most commonly in situations where it is inadvisable (or at best unnecessary) involving pre-testing to

21 select parametric or not-parametric analyses; and making little use of it in model-fitting situations

22 where it might be of value. Of the many alternative tests, we recommend the routine use of either

23 the Shapiro-Wilk or Chen-Shapiro tests; these are almost always superior to commonly-used

24 alternatives like the Kolmogorov-Smirnov test, often by a substantial margin. We describe how both

25 our recommend tests can be implemented. In contrast to current practice as indicated by our

26 survey, we recommend that the results of these tests are reported in more detail (providing both

27 the calculated sample statistic and the associated p-value). Finally, emphasize that even the higher-

28 performing tests of Normality have low power (generally below 0.5 and often much lower) when

29 sample sizes are less than 50, as is often the case in our field.

30

31 Keywords: Gaussian distribution, parametric statistics, Schapiro-Wilk test, statistics, statistical power

32 Word count: 3978

Introduction

The Normal distribution remains the most widely-used statistical model, so it is only natural that researchers will frequently be required to consider whether a sample of data appears to have been drawn from a Normal distribution. This can be done most simply by visual inspection of a histogram of the data, or a more specialised plot such as a Q-Q plot. However visual inspection of this nature on its own does not offer an objective means of decision making: potentially the same researcher could look at a graph on two different occasions and reach different conclusions as to whether the data was suggestive of an underlying Normal distribution or not; or two researchers could disagree when looking at the same graph without having an objective means to resolve their disagreement. Hence, an alternative would be a formal statistical test of the null hypothesis of Normality, with inference being drawn on the basis of a calculated p-value. Commonly-used statistical packages offer a range of different alternative tests (Yap & Sim, 2011). Here we review the statistical literature on the performance of these alternative tests, and briefly survey current usage of these tests in recently-published papers in Animal Behaviour, showing that current common usage departs from what is implied by the statistical literature. We also consider when such testing for Normality is most useful. This should allow us to offer clear advice to authors on how to apply such tests and to readers on how to interpret them.

Literature review

We reviewed the specialist statistics literature on Normality tests in order to explore the evidence in respect to the following issues:

1. Are there differences between alternative tests in terms of their power, and if so how substantial are these differences?
2. If there are substantial differences, can advice on selection of a test be offered?
3. How strongly is the power of such recommended tests affected by sample size?

57 The most recent general comparison of tests of Normality compared the power of eight tests that

58 were available through commonly-used statistics software: Shapiro-Wilk, Kolmogorov-Smirnov,

59 Lilliefors, Cramer-von Mises, Anderson-Darling, D'Agosino-Pearson, Jarque-Bera, and chi-squared

60 tests (Yap & Sim, 2011). Simulation results suggested that if the alternative hypothesis to Normality

61 is not constrained then the Shapiro-Wilk test gives the   highest power. If the alternative is

62 constrained in some way (e.g. by assuming that the alternative will be symmetric but shorter tailed

63 than a Normal distribution), then the Jarque-Bera, D'Agostino-Pearson and Anderson-Darling tests

64 can offer similar power to the Shapiro-Wilk test under different constraints, but they never

65 substantially outperform it. The other four tests (Kolmogorov-Smirnov, Lilliefors, Cramer-von Mises

66 and chi-squared) never outperform Shapiro-Wilk. Yap and Sim (2011) found that power was

67 generally low (less than 0.3 and often much less) for sample sizes lower than 50, but with a steep

68 increase in power to values closer to 1 for sample sizes between 50 and 200. Yazici and Yolacan

69 (2007) concluded that the Shapiro-Wilk test gave the best power when the alternative was

70 unconstrained of the 12 tests they compared. Razali and Wah (2011) argued that across a broad

71 range of circumstances the Shapiro-Wilk test was superior to the Anderson-Darling, Lilliefors and

72 Kolmogorov-Smirnov tests, with the difference in power often being several-fold. However, power of

73 this test was less than 0.5 for five of the six underlying distributions explored when sample sizes

74 where less than 50. Ramao, Degado and Costa (2010) compared 33 different tests and concluded

75 that the Schapiro-Wilk and Chen-Shapiro tests (see below) were the best choices against an

76 unconstrained alternative, and could still be recommended when the form of the alternative was

77 constrained. Keskin (2006) compared four commonly-used tests and concluded that Shapiro-Wilk

78 offered greatest power, sometime seven times that of the other tests. Oztuna, Ethan and

79 Tuccar(2006) reached similar conclusions; and of the various underlying distributions they

80 investigated, only for a uniform distribution was the power of the Shapiro-Wilk test above 0.5 for a

81 sample size of 50. Mendes and Pala (2003) again found the Shapiro-Wilk test to be the most

82 powerful of those tested, sometimes having several-fold more power than commonly-used

4

83   alternatives, but still sometimes being low for even moderate samples sizes. Farrell and Rogers-

84   Stuart (2006) again recommended the Shapiro-Wilk test after an extensive evaluation of 13 different

85   tests across 48 different underlying distributions: across these distributions the power of Shapiro-

86   Wilk test was 0.38 on average for N=20 if α was set to 0.1 to boost power.

87   Although (based on our survey above) the Shapiro-Wilk test seems to be the best performing of the

88   commonly-used tests, the test of Chen and Shapiro (1995) was designed to be always at least as

89   powerful and often more powerful than the Shapiro-Wilk test; and the available evidence suggests

90   that it achieves this performance(Brzezinski, 2012; Marmolejo-Ramos & Gonzalez-Burgos,2013;

91   Seier, 2002).

92   Thus, of the commonly-used and -available tests, the Shapiro-Wilk test can be recommended as

93   having the best power, often significantly greater power than alternatives; but even for this test

94   power can be low for even moderate sample sizes ($N < 50$). For those willing to use a less-familiar

95   test, that of Chen and Shapiro (1995) can be recommended as having generally better performance

96   even than Shapiro-Wilk. Since we recommend these two tests in particular, we now briefly describe

97   how researchers can access them.

98   **Implementation of recommended tests.**

99   The Shapiro-Wilk test is available through many commonly used statistics packages: e.g. SAS, SPSS,

100   Statistica, Stata, and via the *shapiro.test* function in the *stats* package of *R*.

101   For a sample size of *n*, if the sample values ordered from smallest to largest are $x_1,...,x_n$, and their

102   mean value is $\bar{x}$ then the test statistic is given by

103     $W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$,

104   for weights $a_1,...,a_n$, that depend on the expected values and the covariance matrix of the order

105   statistics (for details see for example Thode, 2002). The denominator can be seen as a measure of

106   the variance of the sample. The numerator is essentially a similar measure of the variance that

107   would be the best estimator if the sample were drawn from an underlying Normal distribution. The

108   null hypothesis of an underlying Normal distribution is rejected if W is below a critical value. The

109   challenge in implementing this technique to obtain the weights ($a_1,…,a_n$). The software packages

110   listed above all use the algorithm provided by Royston (1995). Given its implementation in many

111   standard packages, we would be surprised if many researchers chose to implement this test

112   themselves.

113   The Chen-Shapiro test is not available in many commonly used statistical packages: to our

114   knowledge it is only available through the the *PoweR* package in *R*. However, the implementation of

115   this test is sufficiently straightforward that many researchers would be comfortable implementing it

116   themselves. The test statistic QH* is calculated as below:

117
$$QH^* = \sqrt{n}(1 - QH)$$

118   Where *QH* is obtained as

119
$$QH = \frac{1}{s(n-1)} \sum_{i=1}^{n-1} \frac{x_{i+1} - x_i}{H_{i+1} - H_i}$$

120   Where *s* is the standard deviation of the sampled values:

121
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

122   Where $\bar{x}$ is the mean of the x_i values. H_i is given by

123
$$H_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

124   Where $\Phi^{-1}()$ is the inverse of the standard Normal cumulative distribution function. Values of QH^*

125   greater that a critical value suggest significant deviation from a Normal distribution, and critical

126   values are provided in Table 2 and Appendix 2 of Chen and Shapiro (1995).

127   **When should testing for Normality be conducted?**

128   The general consensus in the statistical literature is that preliminary testing for Normality as a means

129   of selecting whether to take a parametric or non-parametric approach to testing the hypothesis of

130   primary interest (e.g. whether to use a t-test or Mann-Whitney U-test to test for a difference in

131   central tendency between two groups) should not be undertaken (e.gSRasch, Kubinger & Moder,

132   2011; Rochon, Gondan & Kieser, 2012; Schoder, Himmelman & Wilhel, 2006; Schucany & Ng 2006;

133   Shuster, 2009; Wells & Hintze 2007; Zimmerman, 2004). This is counter to the advice given in many

134   of the most widely-used introductory statistics texts used by biologists (e.g. Dytham, 2011; Fowler,

135   Cohen & Jarvis, 1998). For example, textbooks generally recommend that when comparing central

136   tendency across groups that the sample for each group is tested separately for Normality. If all

137   groups seem to be drawn from Normal distributions then a t-test or ANOVA is recommended to

138   compare means across groups; otherwise non-parametric equivalents are recommended. However

139   it is often more practical to apply the Normality testing to the residuals generated under the null

140   hypothesis, especially for more complex designs or in the case of a continuous covariate.

141   One argument against this widely-used approach is essentially philosophical: if the pre-test does not

142   give reason to reject the null hypothesis then the scientist proceeds as if the null hypothesis of

143   Normality is true. However the philosophy of null-hypothesis statistical testing is that failure to

144   reject the null hypothesis does not imply that the null hypothesis holds. Essentially, the problem

145   here is that the procedure rests on the implicit assumption that the preliminary test for Normality

146   has very high power, but (as discussed above) this will often be a highly questionable assumption.

147   Another philosophical concern is that the preliminary tests of Normality imply their own

148   assumptions about the underlying distribution and it seems logically inconsistent to check the

149   assumption of Normality but not these other underlying assumptions. On a more practical level the

150   Type I and Type 2 error rates of the key test of interest (e.g. the t-test or U-test in the example

151   mentioned above) are strongly influenced by the detail of the preliminary-testing procedure, and

152   most concerningly the Type I error rates can deviate strongly from the nominal levels.

153   It is also important to note that the reliability of parametric methods such as for example ANOVA

154   and the classical version of the t-test are also sensitive to violation of the assumption of equal

155   variance across groups. Indeed for large samples, methods are often more robust to violation of

156   Normality assumption (Lumley, Diehr, Emerson & Chen, 2002). However, pretesting for

157   homogeneity of variances before selecting an appropriate statistical test is similarly not

158   recommended (Rasch et al., 2011; Zimmerman, 1998; 2004a&b). Some tests of homogeneity of

159   variance make the assumption that the underlying distributions are Normal (Zimmerman 2004a);

160   although the Brown-Forsythe modification of Levene's test was designed to avoid this assumption

161   (Brown & Forsyth, 1974). Further, the robustness of methods to separate violations of either

162   normality or homogeneity of variance assumptions are not a good guide to the robustness of these

163   methods to both violations ocurring simultaneously (Zimmerman, 1998).

164   For the moment, it is safe to conclude that preliminary testing for Normality as a means to selecting

165   whether to take a parametric or non-parametric approach to testing the hypothesis of primary

166   interest should not be undertaken. There are other situations, however, where testing to see if a

167   distribution seems to be Normal seems useful. These relate to evaluating quality of model-fit, rather

168   than selection of parametric versus alternative statistical tests of a null hypothesis. For example,

169   some model fitting procedures (e.g. general linear modelling) assume that residuals around the

170   fitted model are Normally distributed, and it may sometimes be useful to test this as part of

171   evaluation of how successful a model-fitting exercise has been. However, caution needs to be

172   applied in the interpretation of such testing. The issue of low power when sample sizes are small

173   remains; and when sample sizes are very big then the test may suggest rejection for departures from

174 Normality that are biologically trivial.   Alternatively, it might sometimes be useful to test for

175 Normality to help justify fitting a Normal model to data in order to make predictions from that

176 model, taking advantage of the known properties of the Normal distribution. The central limit

177 theorem suggests that we might reasonably often expect to find Normal distributions. The central

178 limit theorem implies that if we draw a large number of independent samples from any underlying

179 distribution, then the distribution of the means of those samples will be approximately Normal.

180 Many test statistics, scores and estimators encountered in practice contain sums of random

181 variables within them. For example, students' exam grades are generally weighted sums of scores on

182 a number of individual questions. Further, many estimators can be represented as sums of random

183 variables through the use of influence functions (Johnson 2004). The central limit theory indicates

184 that these statistical parameters will have asymptotically Normal distributions. Finally, one could

185 interpret the p-value of a test on Normality as a descriptive measure, rather than performing a

186 formal test with a fixed significance level. That could be useful, for example, when trying to find a

187 suitable transformation for a sample of data. Residual analysis including testing on Normality could

188 be applied to decide between different possible transformations.

189 ***Current usage in Animal Behaviour***

190 We found that formal testing of the null hypothesis of Normality was carried out in 23 papers

191 published in Animal Behaviour during 2014. Of these 12 used the Shapiro-Wilk test, 9 the

192 Kolmorogov-Smirnov test, and one each used chi-square goodness of fit and the Lilliefors tests.

193 Sample sizes ranged from 7 to 401, however in 17 of the 23 papers the sample size was 30 or less for

194 at least on test of Normality. For 20 of the 23 papers the Normality test was used in order to decide

195 whether parametric or non-parametric analysis should be used to test the hypothesis of primary

196 interest (our experience with other areas of whole organism biology such as ecology, microbiology

197 and palaeontology suggests this is a very common usage). On the other three occasions the test was

198 used to examine the distribution of residuals from a fitted model. Only one paper of the 23 gave the

199 calculated test statistic and exact *P*-value. All other papers simply reported whether the *P*-value was

200 greater than 0.05 or not, or (presumably equivalently) in words, whether the null hypothesis of

201 Normality was rejected or not. The null hypothesis of Normality was rejected in six papers (9 of the

202 31 test performed overall); the median sample size of tests that rejected the null hypothesis was 29;

203 the median sample size of those that did not reject the null hypothesis was 18: this difference was

204 statistically significant: we used a Brunner-Munzel test rather than a Mann-Whitney U-test because

205 of strong difference in the variances (Neuhäuser, 2012) W_{BF} = 17.45, *P* = 0.023. This suggests that in

206 many cases Normality may have been incorrectly assumed because the test used did not have the

207 power to detect a significant departure from Normality because of low sample sizes.

208

209

210 **Discussion and Conclusions**

211 For very large samples the Shapiro-Wilk test cannot be applied. For example, the function

212 shapiro.test in R does not work for n>5001. However, we would like to mention that any marginal

213 and irrelevant deviation from Normality can be significant in the case of very large samples. Thus, if

214 the sample size is large enough, every sample will be significantly non-Normal because the Normal

215 distribution will never be exactly true with real data. Thus, we do not recommend testing for

216 Normality when sample sizes are extremely large (over 250 as a rule of thumb).

217 Ties (identical values) can occur in a sample; even when the underlying distribution is continuous,

218 rounding (as a result of graduations in a measuring device) leads to ties. Often, the possibility of ties

219 is not considered in the comparison of Normality tests; for instance, Yap and Sim (2011) only

220 investigated continuous distributions. However, the Shapiro-Wilk test is highly sensitive to the

221 presence of ties (Royston, 1989). Royston (1989) presented a simple method of modifying the

222 Shapiro-Wilks test statistic for non-continuous data and showed that the modified test has a high

223 power in comparison to the chi-squared test. In the absence of extensive investigation of the

224 performance of alternative tests; we would recommend Royston's method be used whenever there

225 are ties in a sample. Based on our review above, we think there are a number of ways that

226 researchers in animal behaviour (and more widely) could take better advantage of formal tests of

227 the null hypothesis that a sample is drawn from a Normal distribution.

228 Firstly, at present authors seem to be using such testing most commonly in situations where it is

229 inadvisable (or at best unnecessary); and making little use of it in situations where it might be of

230 value. Specifically, despite this being the most common use by far in our survey of 2014 *Animal*

231 *Behaviour* papers, we do not recommend that authors use a formal test of Normality as a means to

232 selecting whether to take a parametric or non-parametric approach to testing the hypothesis of

233 interest. Rather we recommend that the statistical approach be determined prior to data collection

234 on the basis of underlying knowledge of the system. Where this knowledge is not definitive,

235 conservatively selecting a non-parametric approach can be recommended. Conversely, we

236 recommend that authors make more use of Normality testing in other situations. Firstly, many

237 models within the general linear model framework (including least-squares regression) assume that

238 the residuals around the fitted model are Normally distributed. Thus diagnostic testing of the quality

239 of model fit might often usefully involve testing this assumption (we found 47 papers in 2014 issues

240 of Animal Behaviour where such testing might have been appropriate, of which only three presented

241 or mentioned Normality tests). Secondly, we argue that many quantities of interest to researchers

242 might be expected to be Normally distributed on theoretical grounds, and in such cases we would

243 recommend testing this expectation. If a Normal distribution can be justified then fitting such a

244 model to the data (estimation of the mean and variance) would allow the very well-understood

245 properties of the Normal distribution to be utilised in order to explore expected properties of the

246 population of interest.

Secondly, there are considerable differences between the different tests available in terms of their statistical power. We recommend the routine use of either the Shapiro-Wilk or Chen-Shapiro tests; these are almost always superior to commonly-used alternatives like the Kolmogorov-Smirnov test, often by a substantial margin. We describe (above) how both our recommend tests can be implemented. In contrast to current practice as indicated by our survey, we recommend that the results of these tests are reported in detail (providing both the calculated sample statistic and the associated p-value).

Finally, we emphasize that even the higher-performing tests of Normality have low power (generally below 0.5 and often much lower) when sample sizes are less than 50. This small sample size situation is common in animal behaviour, as indicated by our survey above. Taborsky (2010) found that that the average sample size per treatment in laboratory experiments in the study of behavior was approximately 18, rising to 23 in field studies. In 17 of the 23 papers in our survey the sample size used in at least one test of Normality was less than 30; in such circumstances power to reject the null hypothesis will be low. However, of those 17 papers 14 failed to reject the null hypothesis and none of them discussed the issue of low power. We would recommend that such a discussion should be included any time sample size is less than 50 and the null hypothesis is not rejected.

We believe that these are easy-to-implement actions that together will significantly improve the usefulness of tests for Normality to authors, editors, reviewers and readers across whole-organism biology and beyond.

12

269   **References**

270   Brown M. B., & Forsythe, A. B. (1974). "Robust tests for equality of variances". *Journal of the*

271   *American Statistical Association*, 69, 364–367.

272   Brzezinski, M. (2012). The Chen-Shapiro test nor Normality. *The Stata Journal*, 12, 368-374.

273   Chen, L., & Shapiro, S. (1995). An alternative test for Normality based on Normalised spacings.

274   *Journal of Statistical Computation and Simulation*, 53, 269-287.

275   Dytham, C. (2011). *Choosing and using statistics; a biologists guide.* (3rd ed.). Chickester, U.K.: Wiley-

276   Blackwell.

277   Farrell, P.J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for Normality and symmetry:

278   extending the Spiegehalter test. *Journal of Statistical Computation and Simulation*, 76, 803-816.

279   Fowler, J.,, Cohen, L., & Jarvis, P. (1998). *Practical statistics for field biology.* Chichester, U.K.: John

280   Wiley.

281   Johnson, O. (2004). *Information Theory and the Central limit Theorem*. London, U.K.: Imperial College

282   Press.

283   Keskin, S. (2006). Comparison of serval univariate Normality tests regarding Type I error-rate of the

284   test in simulation based small samples. *Journal of Applied Science Research*, 2, 296-300.

285   Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002) The importance of the normality assumption in

286   large public health data sets. *Annual Reviews in Public Health*, 23, 151-169.

287   Marmolejo-Ramos, F., & Gonzalez-Burgos, J. (2013). A power comparison of various tests of

288   univariate Normality on ex-Gaussian distributions. *Methodology*, 94, 137-149.

289   Mendes, M., & Pala, A. (2003). Type I error rate and power of three Normality tests. *Pakistan Journal*

290   *of Information and Technology*, 2, 135-139.

13

291   Neuhäuser, M. (2012). *Nonparametric statistical tests: a computational approach*. New York: CRC

292   Press.

293   Oztuna, D., Elhan, A. H., & Tuccar, E. (2006). Investigation of four different Normality tests in terms

294   of Type I error rate and power under different distributions. *Turkish Journal of Medical Science*, 36,

295   171-176.

296   Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t-test: pre-testing its assumptions

297   does not pay. *Statistical Papers*, 52, 219-231.

298   Razali, N. M., & Wah, Y. B. (2009). Power comparisons for Shapiro-Wilk, Kolmogorov-Smirnov,

299   Lilliefors and Anderson-Darling tests. *Journal of Statistical Modelling and Analytics*, 2, 21-33.

300   Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: preliminary assessment of

301   Normality when comparing two independent samples. *BMC Medical Research Methodology*, 12, 81.

302   Romao, X., Degado, R., & Costa, A. (2010). An empirical power calculation of univariate goodness-of-

303   fit tests for Normality. *Journal of Statistical Computation and Simulation*, 80, 545-591.

304   Royston, P. (1989). Correcting the Shapiro-Wilk W for ties. *Journal of Statistical Computation and

305   Simulation*, 31, 237-249.

306   Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W test for Normality. *Applied

307   Statistics*, 44, 547-551.

308   Schoder, V., Himmelmann, A., & Wilhel, K. P. (2006), Preliminary testing for Normality: some

309   statistical aspects of a common concept. *Clinical and Experimental Dermatology*, 31, 757-761.

310   Schucany, W.R., & Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for Normality do not validate

311   the on-sample student t. *Communications in Statistics: theory and methods*, 35, 2275-2286,

312   Seier, E, (2002), Comparison of tests for univariate Normality. *Interstat*, 17, 1-17.

313   Shuster, J. (2009). Student t-tests for potentially abnormal data. *Statistics in Medicine*, 28, 2170-

314   2184.

315   Taborsky, M. (2010). Sample size in the Study of Behaviour. *Ethology*, 116, 185-202.

316   Thode. H. C. (2002). *Testing for Normality*. New York: Marcel Dekker.

317   Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology*

318   *in Schools*, 44, 495-502.

319   Yap, B. W., & Sim, C. H. (2011). Comparison of various Types of Normality tests. *Journal of Statistical*

320   *Computation and Simulation*, 81, 2141-2155.

321   Yazici, B., & Yolacan, S. (2007). A comparison of various tests of Normality. *Journal of Statistical*

322   *Computation and Simulation*, 77, 175-183.

323   Zimmerman, D. W. (1998). Invalidation of parametric and non-parametric statistical tests by

324   concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55-69.

325   Zimmerman, D.W. (2004a). A note on preliminary tests of equality of variances. *British Journal of*

326   *Mathematics and Statistics in Psychology*, 57, 173-181.

327   Zimmerman, D. W. (2004b). Conditional probabilities of rejecting Ho by Pooled- and Separate-

328   Variances t Tests Given Heterogenity of Sample Variances. *Communications in Statistics*, 33, 69-81.