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

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

1. School of Biology, University of St Andrews, St Andrews KY16 9TH, UK
2. Natural Science and Psychology, Liverpool John Moores University, Liverpool, L3 3AF, UK
3. Fachbereich Mathematik und Technik, RheinAhrCampus, Koblenz University of Applied Sciences, Joseph-Rovan-Allee 2, 53424 Remagen, Germany

Corresponding author: GDR tel +44 1334 654815; fax +44 1334 644801; email gr41@st-andrews.ac.uk

Abstract

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. Commonly-used statistical packages offer a range of alternative formal statistical tests of the null hypothesis of Normality, with inference being drawn on the basis of a calculated p-value. Here we aim to review the statistical literature on the performance of these tests, and briefly survey current usage of them in recently-published papers, with a view to offering advice on good practice. We find that authors in animal behaviour seem to be using such testing most commonly in situations where it is inadvisable (or at best unnecessary) involving pre-testing to select parametric or not-parametric analyses; and making little use of it in model-fitting situations where it might be of value. Of the many alternative tests, 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 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 more detail (providing both the calculated sample statistic and the associated p-value). Finally, 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, as is often the case in our field.

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

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?

The most recent general comparison of tests of Normality compared the power of eight tests that were available through commonly-used statistics software: Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, Cramer-von Mises, Anderson-Darling, D'Agosino-Pearson, Jarque-Bera, and chi-squared tests (Yap & Sim, 2011). Simulation results suggested that if the alternative hypothesis to Normality is not constrained then the Shapiro-Wilk test gives the highest power. If the alternative is constrained in some way (e.g. by assuming that the alternative will be symmetric but shorter tailed than a Normal distribution), then the Jarque-Bera, D'Agostino-Pearson and Anderson-Darling tests can offer similar power to the Shapiro-Wilk test under different constraints, but they never substantially outperform it. The other four tests (Kolmogorov-Smirnov, Lilliefors, Cramer-von Mises and chi-squared) never outperform Shapiro-Wilk. Yap and Sim (2011) found that power was generally low (less than 0.3 and often much less) for sample sizes lower than 50, but with a steep increase in power to values closer to 1 for sample sizes between 50 and 200. Yazici and Yolacan (2007) concluded that the Shapiro-Wilk test gave the best power when the alternative was unconstrained of the 12 tests they compared. Razali and Wah (2011) argued that across a broad range of circumstances the Shapiro-Wilk test was superior to the Anderson-Darling, Lilliefors and Kolmogorov-Smirnov tests, with the difference in power often being several-fold. However, power of this test was less than 0.5 for five of the six underlying distributions explored when sample sizes where less than 50. Ramao, Degado and Costa (2010) compared 33 different tests and concluded that the Schapiro-Wilk and Chen-Shapiro tests (see below) were the best choices against an unconstrained alternative, and could still be recommended when the form of the alternative was constrained. Keskin (2006) compared four commonly-used tests and concluded that Shapiro-Wilk offered greatest power, sometime seven times that of the other tests. Oztuna, Ethan and Tuccar(2006) reached similar conclusions; and of the various underlying distributions they investigated, only for a uniform distribution was the power of the Shapiro-Wilk test above 0.5 for a sample size of 50. Mendes and Pala (2003) again found the Shapiro-Wilk test to be the most powerful of those tested, sometimes having several-fold more power than commonly-used

alternatives, but still sometimes being low for even moderate samples sizes. Farrell and Rogers-Stuart (2006) again recommended the Shapiro-Wilk test after an extensive evaluation of 13 different tests across 48 different underlying distributions: across these distributions the power of Shapiro-Wilk test was 0.38 on average for *N*=20 if α was set to 0.1 to boost power.

Although (based on our survey above) the Shapiro-Wilk test seems to be the best performing of the commonly-used tests, the test of Chen and Shapiro (1995) was designed to be always at least as powerful and often more powerful than the Shapiro-Wilk test; and the available evidence suggests that it achieves this performance(Brzezinski, 2012; Marmolejo-Ramos & Gonzalez-Burgos,2013; Seier, 2002).

Thus, of the commonly-used and -available tests, the Shapiro-Wilk test can be recommended as having the best power, often significantly greater power than alternatives; but even for this test power can be low for even moderate sample sizes ($N < 50$). For those willing to use a less-familiar test, that of Chen and Shapiro (1995) can be recommended as having generally better performance even than Shapiro-Wilk. Since we recommend these two tests in particular, we now briefly describe how researchers can access them.

Implementation of recommended tests.

The Shapiro-Wilk test is available through many commonly used statistics packages: e.g. SAS, SPSS, Statistica, Stata, and via the *shapiro.test* function in the *stats* package of *R*.

For a sample size of *n*, if the sample values ordered from smallest to largest are $x_1,\ldots,x_n$, and their mean value is $\bar{x}$ then the test statistic is given by

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

for weights $a_1,\ldots,a_n$, that depend on the expected values and the covariance matrix of the order statistics (for details see for example Thode, 2002). The denominator can be seen as a measure of

the variance of the sample. The numerator is essentially a similar measure of the variance that would be the best estimator if the sample were drawn from an underlying Normal distribution. The null hypothesis of an underlying Normal distribution is rejected if *W* is below a critical value. The challenge in implementing this technique to obtain the weights ($a_1,\ldots,a_n$). The software packages listed above all use the algorithm provided by Royston (1995). Given its implementation in many standard packages, we would be surprised if many researchers chose to implement this test themselves.

The Chen-Shapiro test is not available in many commonly used statistical packages: to our knowledge it is only available through the the *PoweR* package in *R*. However, the implementation of this test is sufficiently straightforward that many researchers would be comfortable implementing it themselves. The test statistic QH* is calculated as below:

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

Where *QH* is obtained as

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

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

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

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

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

Where $\Phi^{-1}()$ is the inverse of the standard Normal cumulative distribution function. Values of *QH** greater that a critical value suggest significant deviation from a Normal distribution, and critical values are provided in Table 2 and Appendix 2 of Chen and Shapiro (1995).

When should testing for Normality be conducted?

The general consensus in the statistical literature is that preliminary testing for Normality as a means of selecting whether to take a parametric or non-parametric approach to testing the hypothesis of primary interest (e.g. whether to use a t-test or Mann-Whitney U-test to test for a difference in central tendency between two groups) should not be undertaken (e.gSRasch, Kubinger & Moder, 2011; Rochon, Gondan & Kieser, 2012; Schoder, Himmelman & Wilhel, 2006; Schucany & Ng 2006; Shuster, 2009; Wells & Hintze 2007; Zimmerman, 2004). This is counter to the advice given in many of the most widely-used introductory statistics texts used by biologists (e.g. Dytham, 2011; Fowler, Cohen & Jarvis, 1998). For example, textbooks generally recommend that when comparing central tendency across groups that the sample for each group is tested separately for Normality. If all groups seem to be drawn from Normal distributions then a t-test or ANOVA is recommended to compare means across groups; otherwise non-parametric equivalents are recommended. However it is often more practical to apply the Normality testing to the residuals generated under the null hypothesis, especially for more complex designs or in the case of a continuous covariate.

One argument against this widely-used approach is essentially philosophical: if the pre-test does not give reason to reject the null hypothesis then the scientist proceeds as if the null hypothesis of Normality is true. However the philosophy of null-hypothesis statistical testing is that failure to reject the null hypothesis does not imply that the null hypothesis holds. Essentially, the problem here is that the procedure rests on the implicit assumption that the preliminary test for Normality has very high power, but (as discussed above) this will often be a highly questionable assumption. Another philosophical concern is that the preliminary tests of Normality imply their own assumptions about the underlying distribution and it seems logically inconsistent to check the

assumption of Normality but not these other underlying assumptions. On a more practical level the Type I and Type 2 error rates of the key test of interest (e.g. the t-test or U-test in the example mentioned above) are strongly influenced by the detail of the preliminary-testing procedure, and most concerningly the Type I error rates can deviate strongly from the nominal levels.

It is also important to note that the reliability of parametric methods such as for example ANOVA and the classical version of the t-test are also sensitive to violation of the assumption of equal variance across groups. Indeed for large samples, methods are often more robust to violation of Normality assumption (Lumley, Diehr, Emerson & Chen, 2002). However, pretesting for homogeneity of variances before selecting an appropriate statistical test is similarly not recommended (Rasch et al., 2011; Zimmerman, 1998; 2004a&b). Some tests of homogeneity of variance make the assumption that the underlying distributions are Normal (Zimmerman 2004a); although the Brown-Forsythe modification of Levene's test was designed to avoid this assumption (Brown & Forsyth, 1974). Further, the robustness of methods to separate violations of either normality or homogeneity of variance assumptions are not a good guide to the robustness of these methods to both violations ocurring simultaneously (Zimmerman, 1998).

For the moment, it is safe to conclude that preliminary testing for Normality as a means to selecting whether to take a parametric or non-parametric approach to testing the hypothesis of primary interest should not be undertaken. There are other situations, however, where testing to see if a distribution seems to be Normal seems useful. These relate to evaluating quality of model-fit, rather than selection of parametric versus alternative statistical tests of a null hypothesis. For example, some model fitting procedures (e.g. general linear modelling) assume that residuals around the fitted model are Normally distributed, and it may sometimes be useful to test this as part of evaluation of how successful a model-fitting exercise has been. However, caution needs to be applied in the interpretation of such testing. The issue of low power when sample sizes are small remains; and when sample sizes are very big then the test may suggest rejection for departures from

Normality that are biologically trivial. Alternatively, it might sometimes be useful to test for Normality to help justify fitting a Normal model to data in order to make predictions from that model, taking advantage of the known properties of the Normal distribution. The central limit theorem suggests that we might reasonably often expect to find Normal distributions. The central limit theorem implies that if we draw a large number of independent samples from any underlying distribution, then the distribution of the means of those samples will be approximately Normal. Many test statistics, scores and estimators encountered in practice contain sums of random variables within them. For example, students' exam grades are generally weighted sums of scores on a number of individual questions. Further, many estimators can be represented as sums of random variables through the use of influence functions (Johnson 2004). The central limit theory indicates that these statistical parameters will have asymptotically Normal distributions. Finally, one could interpret the p-value of a test on Normality as a descriptive measure, rather than performing a formal test with a fixed significance level. That could be useful, for example, when trying to find a suitable transformation for a sample of data. Residual analysis including testing on Normality could be applied to decide between different possible transformations.

Current usage in Animal Behaviour

We found that formal testing of the null hypothesis of Normality was carried out in 23 papers published in Animal Behaviour during 2014. Of these 12 used the Shapiro-Wilk test, 9 the Kolmorogov-Smirnov test, and one each used chi-square goodness of fit and the Lilliefors tests. Sample sizes ranged from 7 to 401, however in 17 of the 23 papers the sample size was 30 or less for at least on test of Normality. For 20 of the 23 papers the Normality test was used in order to decide whether parametric or non-parametric analysis should be used to test the hypothesis of primary interest (our experience with other areas of whole organism biology such as ecology, microbiology and palaeontology suggests this is a very common usage). On the other three occasions the test was used to examine the distribution of residuals from a fitted model. Only one paper of the 23 gave the

calculated test statistic and exact *P*-value. All other papers simply reported whether the *P*-value was greater than 0.05 or not, or (presumably equivalently) in words, whether the null hypothesis of Normality was rejected or not. The null hypothesis of Normality was rejected in six papers (9 of the 31 test performed overall); the median sample size of tests that rejected the null hypothesis was 29; the median sample size of those that did not reject the null hypothesis was 18: this difference was statistically significant: we used a Brunner-Munzel test rather than a Mann-Whitney U-test because of strong difference in the variances (Neuhäuser, 2012) $W_{BF} = 17.45$, $P = 0.023$. This suggests that in many cases Normality may have been incorrectly assumed because the test used did not have the power to detect a significant departure from Normality because of low sample sizes.

Discussion and Conclusions

For very large samples the Shapiro-Wilk test cannot be applied. For example, the function shapiro.test in R does not work for n>5001. However, we would like to mention that any marginal and irrelevant deviation from Normality can be significant in the case of very large samples. Thus, if the sample size is large enough, every sample will be significantly non-Normal because the Normal distribution will never be exactly true with real data. Thus, we do not recommend testing for Normality when sample sizes are extremely large (over 250 as a rule of thumb).

Ties (identical values) can occur in a sample; even when the underlying distribution is continuous, rounding (as a result of graduations in a measuring device) leads to ties. Often, the possibility of ties is not considered in the comparison of Normality tests; for instance, Yap and Sim (2011) only investigated continuous distributions. However, the Shapiro-Wilk test is highly sensitive to the presence of ties (Royston, 1989). Royston (1989) presented a simple method of modifying the Shapiro-Wilks test statistic for non-continuous data and showed that the modified test has a high

power in comparison to the chi-squared test. In the absence of extensive investigation of the performance of alternative tests; we would recommend Royston's method be used whenever there are ties in a sample. Based on our review above, we think there are a number of ways that researchers in animal behaviour (and more widely) could take better advantage of formal tests of the null hypothesis that a sample is drawn from a Normal distribution.

Firstly, at present authors seem to be using such testing most commonly in situations where it is inadvisable (or at best unnecessary); and making little use of it in situations where it might be of value. Specifically, despite this being the most common use by far in our survey of 2014 *Animal Behaviour* papers, we do not recommend that authors use a formal test of Normality as a means to selecting whether to take a parametric or non-parametric approach to testing the hypothesis of interest. Rather we recommend that the statistical approach be determined prior to data collection on the basis of underlying knowledge of the system. Where this knowledge is not definitive, conservatively selecting a non-parametric approach can be recommended. Conversely, we recommend that authors make more use of Normality testing in other situations. Firstly, many models within the general linear model framework (including least-squares regression) assume that the residuals around the fitted model are Normally distributed. Thus diagnostic testing of the quality of model fit might often usefully involve testing this assumption (we found 47 papers in 2014 issues of Animal Behaviour where such testing might have been appropriate, of which only three presented or mentioned Normality tests). Secondly, we argue that many quantities of interest to researchers might be expected to be Normally distributed on theoretical grounds, and in such cases we would recommend testing this expectation. If a Normal distribution can be justified then fitting such a model to the data (estimation of the mean and variance) would allow the very well-understood properties of the Normal distribution to be utilised in order to explore expected properties of the 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.

Acknowledgment

We thank two anonymous reviewers for very helpful comments.

References

Brown M. B., & Forsythe, A. B. (1974). "Robust tests for equality of variances". *Journal of the American Statistical Association*, 69, 364–367.

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

Chen, L., & Shapiro, S. (1995). An alternative test for Normality based on Normalised spacings. *Journal of Statistical Computation and Simulation*, 53, 269-287.

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

Farrell, P.J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for Normality and symmetry: extending the Spiegehalter test. *Journal of Statistical Computation and Simulation*, 76, 803-816.

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

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

Keskin, S. (2006). Comparison of serval univariate Normality tests regarding Type I error-rate of the test in simulation based small samples. *Journal of Applied Science Research*, 2, 296-300.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002) The importance of the normality assumption in large public health data sets. *Annual Reviews in Public Health*, 23, 151-169.

Marmolejo-Ramos, F., & Gonzalez-Burgos, J. (2013). A power comparison of various tests of univariate Normality on ex-Gaussian distributions. *Methodology*, 94, 137-149.

Mendes, M., & Pala, A. (2003). Type I error rate and power of three Normality tests. *Pakistan Journal of Information and Technology*, 2, 135-139.

Formatted: Spanish (Spain)

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

Oztuna, D., Elhan, A. H., & Tuccar, E. (2006). Investigation of four different Normality tests in terms of Type I error rate and power under different distributions. *Turkish Journal of Medical Science*, 36, 171-176.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t-test: pre-testing its assumptions does not pay. *Statistical Papers*, 52, 219-231.

Razali, N. M., & Wah, Y. B. (2009). Power comparisons for Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modelling and Analytics*, 2, 21-33.

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: preliminary assessment of Normality when comparing two independent samples. *BMC Medical Research Methodology*, 12, 81.

Romao, X., Degado, R., & Costa, A. (2010). An empirical power calculation of univariate goodness-of-fit tests for Normality. *Journal of Statistical Computation and Simulation*, 80, 545-591.

Royston, P. (1989). Correcting the Shapiro-Wilk W for ties. *Journal of Statistical Computation and Simulation*, 31, 237-249.

Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W test for Normality. *Applied Statistics*, 44, 547-551.

Schoder, V., Himmelmann, A., & Wilhel, K. P. (2006), Preliminary testing for Normality: some statistical aspects of a common concept. *Clinical and Experimental Dermatology*, 31, 757-761.

Schucany, W.R., & Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for Normality do not validate the on-sample student t. *Communications in Statistics: theory and methods*, 35, 2275-2286,

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

Formatted: Spanish (Spain)

Formatted: Spanish (Spain)

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.