2023 ■ Vol. 19 ■ no. 4

Managing the Assumption of Normality within the General Linear Model with Small Samples: Guidelines for Researchers Regarding If, When and How.

Conrad Stanisław Zygmont [a]

[a]Helderberg College of Higher Education, Somerset West, Western Cape, South Africa and Stellenbosch University, Stellenbosch, South Africa

Abstract ■ Academic textbooks, statistical literature, and publication guidelines provide conflicting, ambiguous and often incomplete answers to the question of how researchers should handle the normality assumption for classical general linear model tests when conducting their analyses. Previous studies have shown that normality violations can impact on type I errors, power, parameter estimates and standard error estimates of classical tests. This paper reviews the arguments in favour and against normality testing, the role of the central limit theorem, types of violations that tests within the general linear model are susceptible to, methods for evaluating the normality assumption, and the paradox that normality tests have low power in small sample sizes where the influence of assumption violations are likely to be most profound. A Monte Carlo simulation study was used to evaluate the power of 18 normality tests across 18 alternative distributions, and the effect of normality deviations on estimates of centrality, scatter and regression coefficients. The results demonstrate that the type of normality test and distribution matters, and that a conditional testing procedure utilising normality tests to select between classic, non-parametric and robust tests should not be used. Instead, an alternative procedure for managing the normality assumption is advised, and demonstrated in the supplementary materials using R code and data that are provided.

Keywords ■ normality, parametric assumptions, Monte Carlo, normality test. **Tools** ■ R.

zygmontc@hche.ac.za

10.20982/tqmp.19.4.p302

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

Parametric tests within the General Linear Model (GLM), of which *t*-tests and ANOVA are special cases, have the assumption that the response variable (or residuals of the model) should follow a Gaussian distribution for unbiased parameter estimates and correct inferences. The best methods, and even whether testing for normality is necessary for such statistical tests, remain hotly debated topics in both scholarly literature (e.g. Bishara et al., 2021; Büyükuysal & Sümbüloğlu, 2021; Delacre et al., 2019; Knief & Forstmeier, 2021; Orcan, 2020; Shatz, 2023; Wilcox & Rousselet, 2023) and online discussions among academics (Anglim, 2016; Halvorsen, 2019; Silverfish, n.d.). Data following a non-Gaussian distribution, which violate the normality assumption, frequently feature in research in fields like psychology and education (Cain et al., 2017). In such disciplines real-life data are unlikely to ever be perfectly Gaussian (Micceri, 1989). At the centre of the normality testing debate is the paradox that formal null-hypothesis tests of normality have low power in small sample sizes, when detecting non-normality of the population distribution is most needed (this is when parametric tests are least robust); but in large samples they are too sensitive to immaterial deviations from normality that don't actually matter (where the Central Limit Theorem predicts parametric tests will be most robust to normality violations).

Discrepancies in Statistics Texts

Several popular introductory statistical textbooks, from various disciplines, recommend the assumption of normality should be tested in order to determine if parametric tests are appropriate for use (e.g. Keller, 2018; Montgomery & Runger, 2011; Paolella, 2018). Other textbooks caution against routine normality testing in favour of using non-parametric or robust methods like resampling, or emphasize the robust nature of parametric statistics in large sample sizes (e.g. Gravetter & Wallnau, 2014; Howell, 2013; Weisberg, 2014). For example, Field (2018) suggests "if your sample is large then don't use significance tests of normality, in fact don't worry too much about normality at all. In small samples pay attention if your significance tests are significant but resist being lulled into a false sense of security if they are not" (p. 346). Pek et al. (2018) reviewed 61 undergraduate and graduate level statistics textbooks regarding their recommendations of how to deal with non-normality. They found that most graduate textbooks recommend transformations (89%) while just over half (56%) propose classic parametric GLM tests would be robust due to the Central Limit Theorem (CLT). The two most common approaches discussed in undergraduate textbooks were to either ignore the normality assumption in light of CLT (78%), or use a rank-based method (76%) when data do not conform to normality (conditional testing). Another approach, largely missing from contemporary textbooks was proposed by Hogg (1977a, 1977b) and Tukey (1977), but has largely been ignored in contemporary textbooks. They suggested that non-parametric or robust methods should routinely be conducted simultaneously with classical parametric tests; when results concur across procedures no further analysis is necessary, but when they differ substantially data and theory should be used in the evaluation of which test is best suited. Zimmerman (2011) has recently revived discussion regarding this approach and has shown that it can protect Type I error rates and increase power in small sample sizes. The merits, cautions, and adaptations to this approach will be revisited in the discussion section of this paper. Conflicting suggestions in the literature can leave lecturers, students, and researchers uncertain about how to approach the assumption of normality, which testing processes should be used, and if there is a specific sample size at which normality assumptions are unnecessary.

Guidelines from Associations and Journals

Researchers often rely on association and journal guidelines to direct them in formulating and writing up their statistical analyses. The American Psychological Association (APA) task force on statistical inference recommends researchers "take efforts to assure that the underlying assumptions required for the analysis are reasonable given the data" (Wilkinson & Task Force on Statistical Inference, 1999, p. 601). The most recent APA reporting guidelines require researchers to describe how assumptions were checked and what accommodations were implemented if assumptions were violated (Applebaum et al., 2018). Statistical publishing guidelines for medical journals require that researchers must "verify that data conformed to the assumptions of the test used to analyse them" (Lang & Altman, 2016, p. 33). The American Statistical Association's statement on statistical significance and p-values notes that assumptions should be one of the contextual factors considered for proper statistical inference and that complete reporting and transparency are good scientific practice (Wasserstein & Lazar, 2016). Taken together, these guidelines seem to establish the requirement for thorough evaluation and reporting of how the normality assumption is handled, consideration of its potential effects, and appropriate adjustment in lieu of such considerations.

Arguments against Normality Testing

Many scholars propose that normality assumptions can effectively be ignored. Their main arguments include: (1) normality tests are unreliable in small samples; (2) that with a sufficient sample size commonly used hypothesis tests are sufficiently robust to normality violations given the CLT; or (3) preliminary testing of assumptions can inflate the conditional Type I error rate involved in the two-step testing process (e.g. Field, 2018; García-Pérez, 2012; Gelman & Hill, 2007; Hopper, 2014; Rochon et al., 2012; Rochon & Kieser, 2011). The first two arguments raise questions about what sample size is "good enough" for one to safety ignore the assumption of normality. Various authors have set threshold limits for small sample size, above which CLT will compensate for normality violations, and below which test statistics are likely to be adversely impacted. These have traditionally ranged from 25 to 100, but the actual sample size needed is impacted by the size and type of deviation from normality (Pek et al., 2018). Wilcox and Rousselet (2023) affirm that "contrary to the theorem, normality is not guaranteed in all situations for sample sizes that are often presented as safe in statistics textbooks" (p. 4). There is evidence that research conditions exist in which CLT will not protect for violations in normality with sample sizes of 100, 200, or even 300 when the error distribution is skewed with a thick tail (Lindstromberg, 2020). The third argument has shown to be true in some cases (Rochon & Kieser, 2011), but not in others, depending on which tests are paired together (Parra-Frutos, 2016). In other cases, such as when three or more group means are being compared, the two-step procedure has been shown to be advantageous (Lantz et al., 2016). Irrespective of how re-

searchers have chosen to deal with normality assumptions, their approach should be motivated in their report – something far too few researchers are currently doing (Delacre et al., 2019; Hu & Plonsky, 2019). It may be that some judge them to be unnecessary, but it is likely that many were insufficiently trained or confused regarding which assumptions to test, how to assess them, and what to do if the assumptions are violated (Hoekstra et al., 2012). Despite the controversy over normality testing there remain some researchers (e.g. Kim & Park, 2019; Mishra et al., 2019) and publishing guidelines (e.g. Applebaum et al., 2018) that continue to call for normality testing to be conducted and reported. And various scholars continue investing their time and expertise in trying to develop new (and hopefully improved) approaches to normality testing or modifications to existing tests (e.g. De la Rubia, 2022; Kellner & Cellise, 2019; Mória et al., 2021).

Are Normality Assumptions Really Necessary for GLM Tests?

Historically, pioneers such as Box, Fisher, Geary, Pearson, Please, and Pitman drew attention to the normality assumption and placed emphasis on skewness and kurtosis, demonstrating impact on parameter estimates of homogeneity of variance and measures of location and scatter. They also found classical parametric tests to be fairly robust if underlying distributions have similar shapes, homogeneity of variance, and sample sizes are not too small (Thode, 2002).

Normality and t-tests

Some convergence studies suggest normality may be an unnecessary assumption for t-tests and linear models relying on measures of central tendency even for large deviations from normality for samples as low as 100 (Lumley et al., 2002; Schröder & Yitzhaki, 2017). Others show that t-tests may lose power when normality is violated (e.g. Blair & Higgins, 1980a, 1980b, 1981, 1985; van den Brink & van den Brink, 1989). Wilcox (2001, 2022) demonstrated numerous examples of the vulnerability of parametric tests to subtle violations of distributional assumptions. Wilcox (2012) makes a strong argument that "resorting to the central limit theorem in order to justify the normality assumption can be highly unsatisfactory when working with means" (p. 328). Despite the central limit theorem assumption that the sampling distribution of the mean will be normal in large sample sizes, it is important to note that the empirical t distribution can deviate substantially from the asymptotic Student's t distribution when non-normality is present (Wilcox, 2022). For this reason theoretical and review articles often insist that for t-tests, which are based on the mean as an estimate of population central tendency, evaluating the normality assumption is essential (e.g. Kim & Park, 2019; Mishra et al., 2019). Generally the Student's t-test is much more sensitive to deviations from normality in the form of skewness than in the form of kurtosis, particularly in small samples (Sawilowsky & Blair, 1992; Wilcox, 1990; Zumbo & Jennings, 2002). Kurtosis impacts on the standard error of sample variance even at large sample sizes, with standard error underestimated in the case of leptokurtic distributions and overestimated with platykurtic distributions (Cain et al., 2017). But skewness and kurtosis on their own cannot be used to make a judgement if the t-test is appropriate, the whole structure of the distribution should be studied (Lee & Gurland, 1977; Orcan, 2020).

Normality and One-Way Fixed-Effects ANOVA

The ANOVA F statistic is generally considered more robust to moderate violations of distributional assumptions than the t-test (Blanca et al., 2017). ANOVA is more susceptible to heterogeneity combined with differences in group sample sizes than normality violations in the error distribution (Blanca et al., 2018; Delacre et al., 2019). In a review of early studies, Glass et al. (1972) concluded that the ANOVA F statistic is more susceptible to violations in kurtosis than skewness. Khan and Rayner (2003) came to the same conclusion in their simulation study. But research in this regard has not been consistent. Harwell et al. (1992) conducted a meta-analysis and found that skewness could impact ANOVA Type I error rates more than kurtosis. Findings vary based on the criteria used, and whether studies are more concerned with increases in Type I error rates, which are normally fairly robust, or decreases in power, which is more likely to be impacted (Wilcox & Rousselet, 2023).

Normality and Linear Regression

With regard to Ordinary Least Squares (OLS) linear regression, Gelman et al. (2021) suggest that normality is the least important assumption, and that it is not necessary to test for normality. Others argue that non-normality in regression residuals can distort regression parameter estimates and significance tests, and so residuals should be tested for normality (e.g. Beaujean, 2014; Das & Imon, 2016; Fox & Weisberg, 2011; Osborne & Waters, 2002). Monte Carlo simulations have demonstrated that when there are outliers on both the outcome and predictor variables, type I error rates are likely to be effected, particularly with small samples (Knief & Forstmeier, 2021). Type I errors and regression coefficients are more susceptible to bias from high skewness at low sample sizes than other types of Gaussian violations, mainly because of the possibility of the introduction of high leverage outliers. Parameter estimates may become nonsensical when assuming a normal dis-

tribution with non-Gaussian data. For this reason, Knief and Forstmeier (2021) suggest running linear models with a rank-based inverse normal (RIN) transformation of the data for hypothesis testing. But for parameter estimation, they propose using advanced approaches like generalized linear mixed models (with adequate checking of the assumptions of these techniques). Instead of transformations, Silva-Lugo et al. (2021) demonstrate that even at large sample sizes, non-parametric models can sometimes provide more accurate estimates. They proposed "instead of using the Central Limit Theorem to justify parametric linear regression analyses, we should use cross-validation or double cross-validation to make more objective and sound decisions as scientists" (Silva-Lugo et al., 2021, p. 15). Wilcox and Rousselet (2023) concur that whenever using parametric tests it is always "prudent to check the extent to which robust methods give similar results" (p. 28).

Normality and Linear Mixed Models

With linear mixed models, fixed effect model estimates have been found to be robust to violations of normality assumptions for Gaussian models (Warrington et al., 2014). Similarly, Schielzeth et al. (2020) found that even with violations of the distributional assumptions of random effects variances or residual variances the model estimates remained fairly robust. Parameters most closely linked to violations may result in greater variability in estimates from sample to sample when distributions are severely skewed, bimodal, or residuals are heteroskedastic.

Summary - Is It Necessary to Test the Normality Assumption?

In summary, normality may not be the most important assumption in general linear models (Gelman et al., 2021). However, evaluating normality is essential for detecting large deviations and outliers in small sample sizes, but this is precisely where formal normality tests tend to lose reliability and accuracy, and where choice of test becomes imperative (Razali & Wah, 2011). Outliers can influence both skewness and kurtosis, with deviations in skewness affecting estimates of location, and kurtosis affecting estimates of scatter. Different kinds of normality tests, and various tests in the General Linear Model, are differentially sensitive to such deviations. Classical *t*-tests are impacted more by deviations in skewness whereas ANOVA is influenced more by deviations in scatter, particularly if any of these models include groups with unequal sample sizes. It would be good practice to use normality tests that are sensitive to the kinds of variations that matter to your model, and always consider these in combination with other assumptions of your model.

How Should Researchers Evaluate the Normality Assumption?

Approaches for testing the parametric assumptions of data can be broadly categorized into (1) graphical methods, (2) descriptive statistics, and (3) goodness-of-fit (GOF) test statistics. Graphical approaches are useful because they allow for a quick, intuitive evaluation for the kinds of deviation from normality that is likely to be important for any specific model (e.g. outliers, skewness, or kurtosis). Various scholars and journal publishing guidelines have recommended that the normality assumption (of the data or residuals for ANOVA and regression) should be evaluated with discretion using quantile-quantile (Q-Q) plots (Grech & Calleja, 2018; Kozak & Piepo, 2017; Schucany & Ng, 2006; Shatz, 2023; Wilkinson & Task Force on Statistical Inference, 1999). The challenge with graphical evaluation of Q-Q plots is that interpretation is somewhat subjective resulting in fairly low inter-rater reliability (Aldor-Noiman et al., 2013; Loy et al., 2016). In order to address this limitation, Huang et al. (2019) developed an objective hypothesis testing process based on machine learning and computer vision evaluation of Q-Q plots. Descriptive statistics, including measures of skewness and kurtosis, provide quantitative indicators of deviations from normality in the sample. However, these sample statistics are seldom sufficient to detect normality violations that impact parameter estimates (Orcan, 2020). Finally, formal Goodness-of-fit hypothesis tests are available to evaluate omnibus hypothesis of normality. There are over 100 such tests available, with the Kolmogorov-Smirnov (Kolmogorov, 1933) test described as the most popular normality test (Arnold & Emerson, 2011). Its frequent use is likely because of its inclusion as a default in IBM SPSS Statistics (Pedrosa et al., 2015). It is often used despite the fact that "most people do not recommend its use" [italics in original] (Howell, 2013, p. 78). Any attempt to identify the best test is hampered by the infinite number of alternative distributions to test against. Typically, tests tailored to a specific distribution category have very low power for detecting divergence from normality on other categories (Farrell & Rogers-Stewart, 2006; Islam, 2017). For example, while the Jarque-Bera test has relatively high power with long-tailed distributions, it performs poorly for distributions with short tails, especially if the shape is bimodal (Thadewald & Büning, 2007). Tests based on the Vasicek (1976) entropy estimator have been found to outperform many other tests when distributions have little skew, negative kurtosis, high skew and kurtosis, or are bimodal but perform very poorly with log-normal distributions (Alizadeh Noughabi & Arghami, 2011, 2012; Yazici & Yolacan, 2007; Zamanzade & Arghami, 2012). Islam (2017, 2019) has attempted to overcome this limitation by

calculating a power envelope for various t-distributions using LR-tests based on Neyman-Pearson lemma, which can be used to calculate the stringencies for a number of normality tests. Islam (2017) found the Anderson-Darling test to have the highest power for all sample sizes on the complete selected class of alternatives, followed closely by the Chen-Shapiro (Chen & Shapiro, 1995) test, which performs particularly well with smaller and medium sample sizes. Numerous reviews of univariate goodness-of-fit (GOF) normality tests have been published (e.g., Adefisoye et al., 2016; Ahmad & Sherwani, 2015; Arnastauskaitė et al., 2021; Islam, 2017, 2019; Farrell & Rogers-Stewart, 2006; Pedrosa et al., 2015; Romão et al., 2010; Seier, 2002; Sürücü, 2008; Schick et al., 2011; Uyanto, 2022; Yap & Sim, 2011; Wijekularathna et al., 2022). With the exception of Adefisoye et al. (2016), Arnastauskaitė et al. (2021), Romão et al. (2010), and Uyanto (2022), most reviews have covered only a small number of well-known tests or focused on a specific class of tests, such as the entropy-based estimators (e.g., Alizadeh Noughabi & Arghami, 2011, 2012; Zamanzade & Arghami, 2012). Often the tests suggested are powerful only with larger sample sizes (e.g. Arnastauskaitė et al., 2021). None of these broad evaluations have included more than one type of empirical characteristic function class tests, which have recently received more attention (e.g., Bakshaev & Rudzkis, 2017; Lafaye de Micheaux & Tran, 2016; Van Zyl, 2017). Some reviews have included relatively little known tests, such as the Gel-Miao-Gastwirth (2007) test or Csörgő (1986) CS statistic, modifications of commonly used omnibus tests (e.g., Sürücü, 2008), or provided comparisons among different classes of tests and demonstrated the superiority of some little known tests for specific types of alternative distributions (e.g., Yazici & Yolacan, 2007). Even though some of these tests show merit they are not readily available to researchers in commonly used statistical software (Miecznikowski et al., 2013). As a result many researchers continue to use the default tests available in their software of choice, which is often a dubious choice (Engmann & Cousineau, 2011). This paper extends the work of previous reviews by evaluating the need for normality testing, comparing a large number of tests that have not all appeared together in a previous review, and making the procedures involved in normality testing more accessible to researchers by providing code for their implementation in appendix A and a demonstration of their use in appendix C.

Methodology

Design

A Monte Carlo simulation design was used with two primary goals: (a) to evaluate the power of eighteen normality tests in detecting departures from a population with a normal $N(\mu,\sigma^2)$ distribution at sample sizes ranging from 8 to 120; and (b) evaluate the accuracy of parameter estimates for location, scatter, and linear regression coefficients for each alternative distribution at different sample sizes. Zumbo and Jennings (2002) introduced the contamination index (CI) as a measure of the extent to which a sample distribution deviates from normal. Seeing as parameter values are available in a simulation study, the root mean square error (RMSE) was chosen as the preferred measure of the accuracy of parameter estimates for the mean, standard deviation, and regression beta coefficients. Higher RMSE values indicate that parameter estimates for that distribution are contaminated. Critical values for the Vasicek (1976) test, Zamanzade and Arghami (2012) TZ2mn test, Rahman and Govindarajulu (1997) modified Shapiro-Wilk test, the Chen and Shapiro (1995) test, the B_3^2 Coin (2008) test, the Data Driven Smooth Test (Janic & Ledwina, 2009), and D'Agostino (1971) omnibus test were calculated using 50,000 simulated samples from a standard normal distribution. For the Chen and Shapiro (1995) and Coin (2008) tests, which are right tailed, the 95th percentile was used; for the entropy tests and Rahman and Govindarajulu (1997) test, which are left tailed, the 5th percentile was used; and for the D'Agostino (1971) test values were calculated for the lower tail at the 2.5th percentile and upper tail at the 97.5th percentile. For the ECF test (Van Zyl, 2017) normality was rejected where $\left|v_n(1)/\sqrt{0.0431/n}\right| = \left|4.8168\sqrt{n}V_n(1)\right| > z_{1-\alpha/2}$. For all other tests the p values generated by the available functions in R were used in the analysis. Analyses were performed in R (Version 4.0.0, R Core Team, 2023) on a PC running Gentoo Linux using 10,000 simulations and a seed of 54321.

Normality tests used in this study

There were 18 different normality tests included in this study that represent over 5 different classes of univariate goodness-of-fit statistics. In order to aid interpretation later in the paper, the abbreviation used in the study, the test reference, equations, and R functions where the tests can be found are summarized in Table at the end of this article.

Population distributions included in this study

The performance of the various normality tests, and effects on parameter estimates, was evaluated using simulations across a range of different alternative distributions. There exist an infinite number of distributions one could test against and different approaches exist for selecting and characterizing distributions. At the most basic level distributions can be differentiated as either symmetrical or asymmetrical (Montenegro & Alonso, 2015). Yap

and Sim (2011), Farrell and Rogers-Stewart (2006), and Wijekularathna et al. (2022) extend this classification to symmetric short-tailed, symmetric long-tailed, and asymmetric distributions. Quessy and Mailhot (2011) suggest four kinds of alternative distributions; namely, bimodal alternatives, kurtosis alternatives, heavy-tailed alternatives, and mixture of skewness and kurtosis alternatives. Alizadeh Noughabi and Arghami (2011, 2012), Uyanto (2022) and Zamanzade and Arghami (2012) categorize distributions into four categories; namely, symmetric distributions with support for $(-\infty,\infty)$, asymmetric distributions with support for $(-\infty,\infty)$, distributions with support bounded at $(0,\infty)$, and distributions with support bounded at $(0,1)$. The greatest number of categories was defined by Seier (2002), who used a total of 9 categories. The approach taken in this paper is similar to that of Romão et al. (2010), who used symmetrical, asymmetrical and a number of mixed normal distributions with various shapes. Mixed normal distributions, or contamination models, have demonstrated their usefulness as population models to mimic outlier contamination or other properties of real non-normal distributions across a variety of disciplines (Blair & Higgins, 1980b; Zumbo & Jennings, 2002). Symmetrical and asymmetrical distributions were selected based on their utility in previous simulation studies and to allow for comparisons across studies. Symmetrical distributions included Student's t (df = 5), Logistic, Tukey (shape = -0.25), Tukey (shape = 0.75), Tukey (shape = 1.05), and Uniform. Asymmetrical distributions included Weibull (shape = 2, scale = 3), Generalized Pareto (location = 0, scale = 2, shape = 0), Gumbel (location = 0, scale = 2), Gamma (shape = 2, rate = 3), asymmetric Power (location = 0, asymmetry = 2, scale = 0.8, tail decay = 1.5), and asymmetric Laplace (location = 4, scale = 2, asymmetry = 2). Contaminated population models were created to mimic parameter estimates from real-life distributions from the social sciences known to deviate from normality. They included: time spent eating and drinking per day (cf., U. S. Bureau of Labor Statistics, 2015), income (cf., United State Census Bureau, 2013), age at death (cf., Australian Institute of Health and Wealfare, 2015), GPA scores (cf., University of Wisconsin-Madison, 2017), lawyer starting salaries (cf., National Association for Law Placement, 2015), and age of sexual debut (cf., Bakilana, 2005; Zuma et al., 2011). These distributions have high contamination indices justifying their utility for this study; income having the largest (CI = 1.00), followed by age at death (CI = 0.63), with only GPA scores having a CI below 0.1 (CI = 0.03).

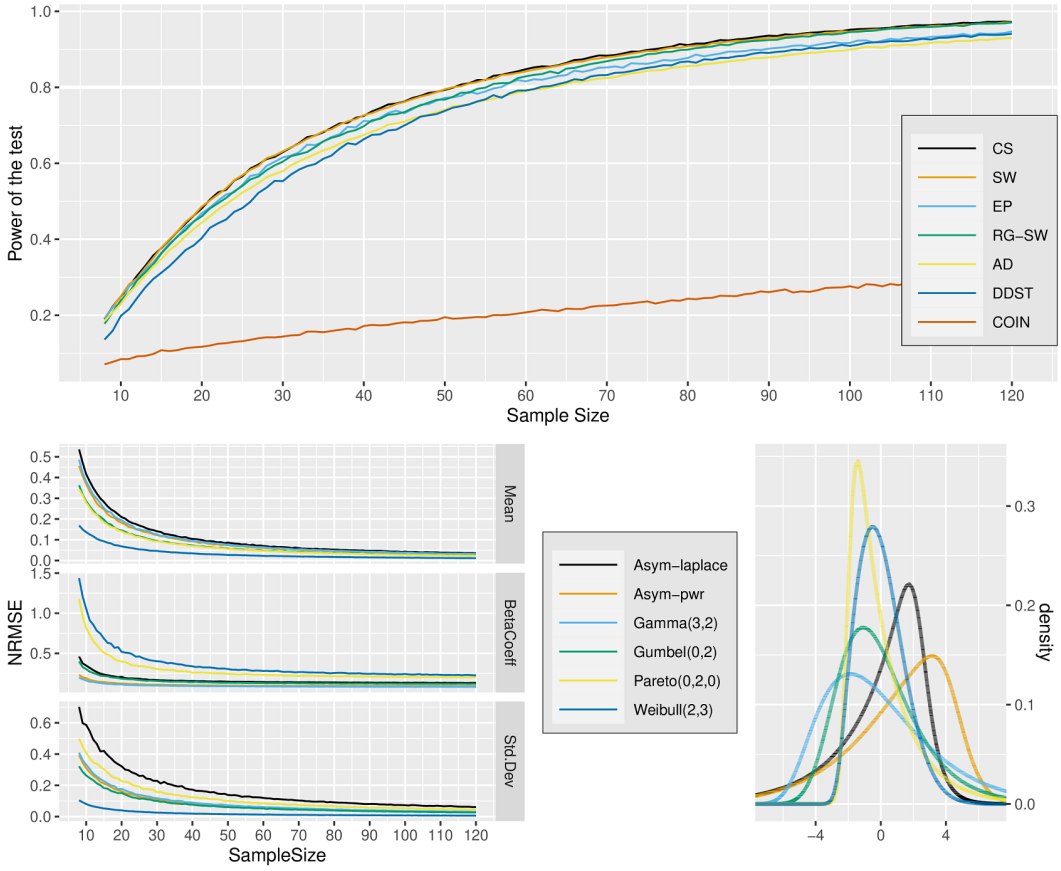
Results

The results of 10,000 Monte Carlo simulations at sample sizes from 8 to 120 on the power of normality tests across 18 distributions within three distribution groups is presented below. The findings regarding specific categories of distribution violations are presented first, followed by an overall synopsis of the findings. Readers who wish to study tables providing power for each normality test for each specific distribution are referred to Appendix B.

Relative Power of Normality Tests for Asymmetric Distributions

For asymmetric distributions the Chen-Shapiro and Shapiro-Wilk tests had the highest power at lower sample sizes across the distributions included in this category. Across all the distributions, Chen-Shapiro and Shapiro-Wilk statistics were able to correctly reject the null with an average power of .80 with samples as small as 53 and the Epps-Pulley 51. The Chen-Shapiro, Shapiro-Wilk, Vasicek and DBEG tests needed sample sizes as low as 19 in order to attain power of at least .80 for a Generalized Pareto (location = 0, scale = 2, shape = 0) distribution, whereas sample sizes of 93 and 95 were needed by Chen-Shapiro and Shapiro-Wilk respectively to attain the same level of power with a Weibull (shape = 2, scale = 3) distribution. The Generalized Pareto distribution has greater skew and kurtosis (skew = 2, kurtosis = 6.08) than the Weibull (shape = 2, scale = 3) distribution (skew = 0.63, kurtosis = 0.24), but both have a proportionally large impact on regression coefficients compared with the rest of the distributions in this category. The asymmetric Laplace (location = 4, scale = 2, asymmetry = 2) had the largest impact on estimates of the mean and standard deviation with its combination of a thick tail and skewness (skew= -1.81, kurtosis = 5.42). The asymmetric Laplace distribution required a sample size of 30 in order to attain a power of .80 with the Chen-Shapiro Statistic. On the other hand, the Coin B_3^2 statistic only obtained a power of .57 and Bonett-Seier 0.61 with a sample size of 100 for the same distribution. In order to evaluate the comparative power of normality tests across all the studied distributions within each category, the Normalized Root Mean Squared Error (NRMSE) was used as a metric of the error in classic measures of location, scale and regression beta coefficients in each distribution. This was then used to obtain the average power across the various asymmetric distributions, weighted by the degree to which the distribution would actually impact parameter estimates. The NRMSE is a robust measure of the accuracy of models, estimators, or predictions that allows for comparisons between distributions with different scales. RMSE is often normalized by the range, but in this case the MAD was used for a more robust estimate of scale, resulting in a measure from 0 to 1. Figure 1 presents the weighted average power of the six most powerful tests, as well as the least powerful test. Also represented are the distributions and NRMSE of parameter estimates for sample sizes from 8 to 120. Ta-

Figure 1 ■ Graphical representation of results for asymmetric distributions. The normality tests in the legend are ranked based on their weighted-average power across all distributions in this category for sample sizes below 60, based on their impact on parameter estimates (NRMSE). The six tests with the highest power in this sample size range, as well as the statistic with the lowest power are shown in descending order. Also provided are the error deviations in sample estimates of parameters for centrality, scatter and regression coefficients normalised by the MAD.

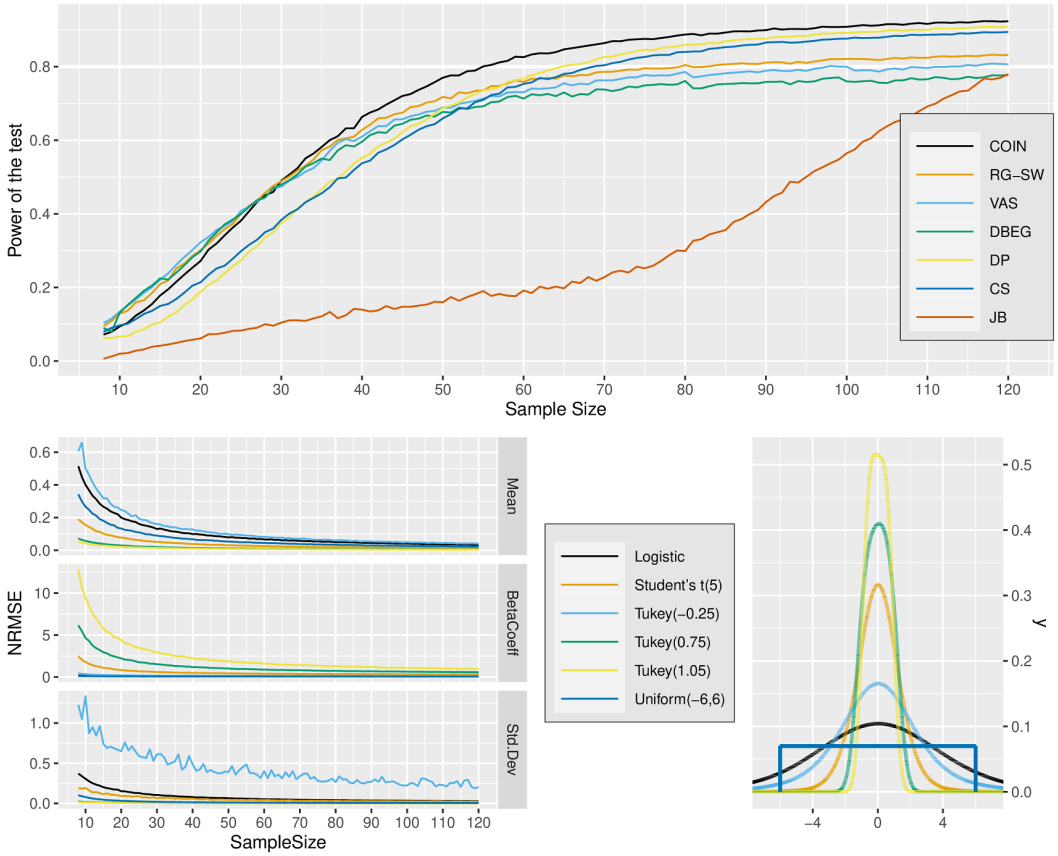


bles with simulated power across specific distributions are provided in the appendices.

Relative Power of Normality Tests for Symmetric Distributions

For symmetric distributions performance was differentiated between distributions with high and low kurtosis. In distributions with high kurtosis the Gel-Miao-Gastwirth and Robust Jarque-Bera tests performed best, whereas symmetric distributions with negative kurtosis were best differentiated from normal by the Vasicek and DBEG statistics. Conversely, the Robust Jarque-Bera, Jarque-Bera, and Lilliefors Kolmogorov-Smirnov tests performed poorly for platykurtic symmetric distributions, while the DBEG statistic and Vasicek statistic had the lowest power for leptokurtic symmetric distributions. In order to reach a power of .8 with the most powerful statistic (COIN), a sample size of 36 was sufficient for the Tukey (shape = 1.05) distribution. This distribution has a relatively small variation (sd = 0.55), but scores are distributed in the shoulders (kurtosis = -1.21), which may impact on regression coefficient estimates in small samples but does not bias estimates of the mean and standard deviation. However, even at a sample size of 120 the tests examined here were not able to correctly reject the null at better than .50 for Logistic and .74 for the Student's t(df = 5) distribution. The Logistic and t(df = 5) distributions are fairly leptokurtic (kurtosislog = 1.2, kurtosist(5) = 5.69), although not as much as the Tukey (shape = -0.25) distribution (kurtosis = 31.84); impacting on estimates of centrality and spread, but making their deviation from normal not as easily detectable as the Tukey (shape = -0.25) distribution. The Rahman and Govindara-

Figure 2 ■ Graphical representation of results for symmetric distributions. The normality tests in the legend are ranked based on their weighted-average power across all distributions in this category for sample sizes below 60, weighted by their impact on parameter estimates (NRMSE). The six tests with the highest power in this sample size range, as well as the statistic with the lowest power are shown in descending order. Also provided are the error deviations in sample estimates of parameters for centrality, scatter and regression coefficients normalized by the MAD.

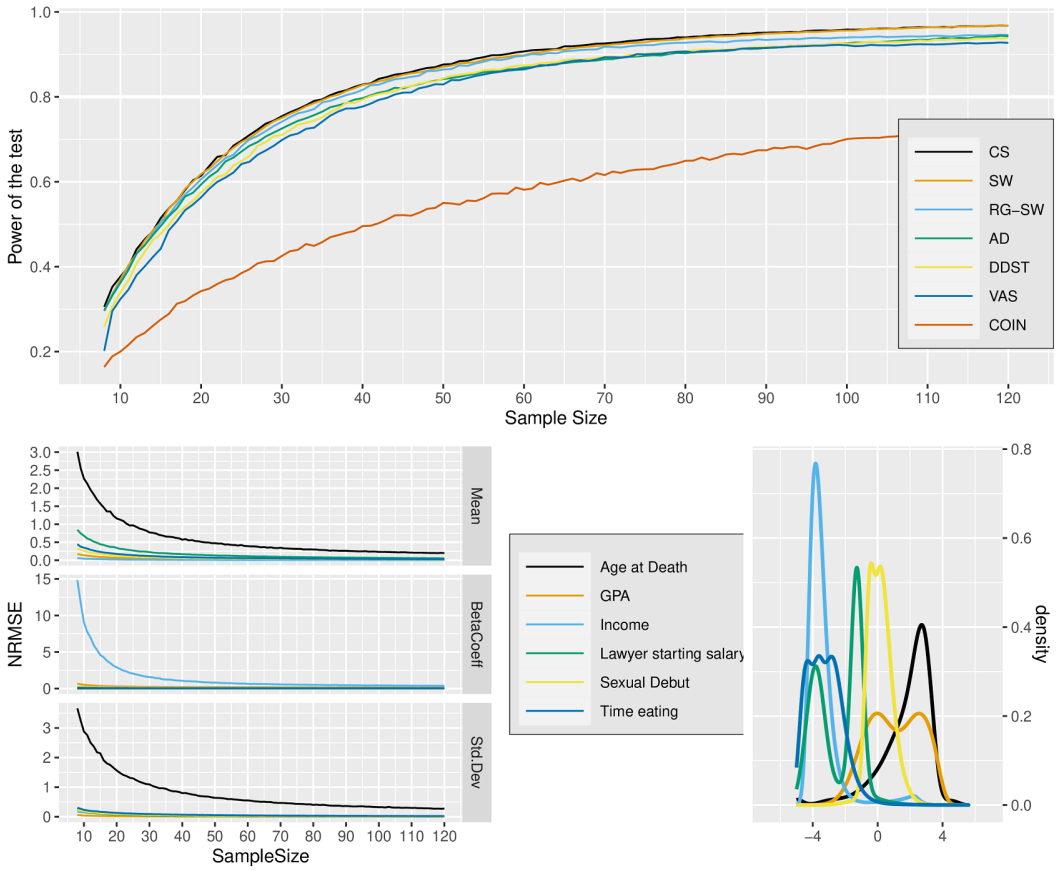


julu's Shapiro-Wilk test, Vasicek, and DBEG tests were able to achieve .80 power in sample sizes of 41 with low kurtosis symmetric distributions. Figure 2 presents the weighted-average power of the six most powerful tests, as well as the least powerful test, and the distribution representations and NRMSE of parameter estimates for sample sizes from 8 to 120.

Relative Power of Normality Tests for Mixed Distributions

Normality tests generally had higher power in detecting departures in the mixed distribution category, compared to the other two distribution categories, as distributions had more noticeable deviations from normality and this resulted in greater error in parameter estimates, particularly in small samples. Tests that consistently performed with high power in small sample sizes in this category included Chen-Shapiro, Shapiro-Wilk, Rahman and Govindarajulu's Shapiro-Wilk, and Anderson-Darling test statistics. Tests that performed poorly in this category included Coin B_3^2, Van Zyl's ECF test, Jarque-Bera test, Robust Jarque-Bera, Bonett-Seier, and Lilliefors tests. For sample sizes below 18 the Jarque-Bera test performed the worst on average, but it outperformed van Zyl's ECF test in samples above 18 and the Robust Jarque-Bera in sample sizes above 45. The van Zyl ECF test had lower power on average than the Robust Jarque-Bera for samples below 50, but had better power in larger samples. Performance varied depending on the idiosyncratic characteristics of each distribution. For example, the distribution of lawyer starting salaries was roughly bimodal with a negative kurtosis (-1.23) and slight skew (-0.11). With this distribution the Vasicek statistic was able to attain a power of .8 with a sample as small as 12, whereas

Figure 3 ■ Graphical representation of results for mixed distributions. The normality tests in the legend are ranked based on their average power across all distributions in this category for sample sizes below 60, weighted by their impact on parameter estimates (NRMSE). The six tests with the highest power in this sample size range, as well as the statistic with the lowest power are shown in descending order. Also provided are the error deviations in sample estimates of parameters for centrality, scatter and regression coefficients normalized by the MAD.



the Dagostino omnibus test had a power of .32 with a sample of 120 for the same distribution. On the other hand, time spent eating and drinking is long-tailed (kurtosis = 1.48) with a positive skew (0.78). This distribution required a sample size of 67 in order for the Vasicek statistic to reach a power of .8, with the Bonett-Seier test only reaching a power of .27 at a sample size of 120. The income distribution, with the majority of of the sample earning a modest income but a small proportion of the sample earning 80% of the total wealth (outliers) far to the right of the distribution, had a disproportionately large error in estimated regression coefficients. This was followed by GPA and sexual debut distributions. On the other hand, the distribution representing age at death that has a small peak during infancy a large spread and then strong negative skew (skew = -1.63, kurtosis = 3.57, MAD = 12.26), had high errors on estimates of central tendency and variability. Figure 3 displays plots that present the shape of the distributions, NRMSE, and the power of the 6 most powerful and the least powerful test based on their weighted-average covering sample sizes ranging from 8 to 120.

Discussion

The results of this study reiterate that GOF normality tests have severely limited utility for detecting departures from normality at very small sample sizes ($n < 35$), and that the choice of test is very important. There is no single normality test that performs best for every distribution and at every sample size. Using a single omnibus normality test consistently among different types of distributions, which has been reported to be the practice among many researchers (Engmann & Cousineau, 2011), will often result in an un-

Table 1 ■ Tests with high and low power at low sample size for various distributions

Category	High Power	Low Power	Vulnerable parametric test
Asymmteric	Chen-Shapiro, Shapiro-Wilk, Epps-Pulley, Rahman and Govindarajulu's SW, Anderson-Darling	Coin's B_3^2, Bonett-Seier, Gel-Miao-Gastwirth, ECF	Linear mixed models, OLS regression coefficients, and t-tests
Symmetric – high kurtosis	Gel-Miao-Gastwirth, Robust Jarque-Bera, Zamanzade, D'agostino K^2	DBEG statistic, Vasiczek, Rahman and Govindarajulu's SW	Regression coefficients and standard errors, and ANOVA
Symmetric – low kurtosis	DBEG, Vasicek, Rahman and Govindarajulu's SW, Coin's B_3^2	Robust Jarque-Bera, Jarque-Bera, and ECF	Regression coefficients & standard error, and t-tests
Mixed	Rahman and Govindarajulu's SW, Chen-Shapiro, Shapiro-Wilk, and Vasicek	ECF, Jarque-Bera, Robust Jarque-Bera, and D'Agostino omnibus	ANOVA, t-test, linear regression, and linear mixed models depending on specific distribution

acceptable loss of power (Quessy & Mailhot, 2011). In order to detect deviations from normality at an acceptable level of power, one generally needs sample sizes between 50-100 depending on the population characteristics. Normality tests should be selected based on their sensitivity to the distributional characteristics of the population distribution under investigation, which might be obtained from previous research. For asymmetric distributions, the Chen-Shapiro and Shapiro-Wilk tests offered the highest power at low sample sizes among the alternatives examined in this study. Previous studies found the Shapiro-Wilk and Anderson-Darling tests particularly useful at smaller samples, with the Jarque-Bera and van Zyl ECF tests performing well at larger sample sizes (Alizadeh Noughabi & Arghami, 2011, 2012; Lafaye de Micheaux & Tran, 2016; Romão et al., 2010; Seier, 2002; Van Zyl, 2017; Yap & Sim, 2011; Yazici & Yolacan, 2007). In general, for symmetric distributions the COIN B_3^2 test, Rahman and Govindarajulu's Shapiro-Wilk, Chen-Shapiro, Bonett-Seier, and Gel-Miao-Gastwirth statistics perform well in small samples (Islam, 2017; Romão et al., 2010; Quessy & Mailhot, 2011). However, the degree of skew and type of kurtosis are also important. For example, the B_3^2 test performs well with slightly skewed alternatives, but not when alternatives are highly skewed (Islam, 2019). When symmetric distributions have short tails and wide shoulders (low kurtosis), the D'Agostino-Pearson statistic, D'Agostino K^2 statistic, and Shapiro-Wilk tests have demonstrated high power (Jäntschi & Bolboacă, 2009; Razali & Wah, 2011; Seier, 2002; Yap & Sim, 2011). Within this study, the DBEG, Vasiczek, and Rahman and Govindarajulu's Shapiro-Wilk test performed best for platykurtic symmetric distributions. Among high kurtosis distributions the Robust Jarque-Bera, Shapiro-Wilk, Alizadeh Noughabi and Arghami (2010) entropy based estimator, and the Anderson-Darling tests have consistently performed well (Alizadeh Noughabi & Arghami, 2011, 2012; Islam, 2017; Thadewald & Büning, 2007; Yap & Sim, 2011). In this study the Gel-Miao-Gastwirth, Robust Jarque-Bera, and Zamanzade entropy-based tests performed best for detecting symmetric leptokurtic distributions. Interestingly, while Zamanzade and Arghami (2012) found the test based on their entropy estimator to be superior to other entropy-based tests for symmetric distributions, in this study Vasicek performed better on platykurtic distributions. Researchers may use Table 1 as a guide in order to select normality tests based on two criteria: (1) which is most powerful at detecting the types of violations from normality expected on theoretical grounds for the distribution under study, and/or (2) which is most powerful for detecting the types of violations the parametric statistic one is using is most susceptible to.

Overall, the results suggest that researchers should not perform conditional testing based on the results of a normality test at small sample size, as tests are generally not powerful, and this could inflate Type I errors. There are many non-parametric, resampling-based, or robust estimator-based tests available as alternatives to commonly used parametric tests (e.g., Cribbie et al., 2012; Fried & Dehling, 2011; Wilcox, 2012, 2022). For sample sizes between 50-100 the magnitude of errors in scale and location estimates diminishes dramatically and CLT does well to protect the asymptotic distribution of the test statistic. However, for some specific distributions (e.g. skew and heavy tailed), parameter estimates may still be considerably impaired even when sample sizes are 120 or more (Bradley, 1980; Field & Wilcox, 2017; Wilcox, 2012, 2022). At the same time, it is not necessary for normality departures to be dramatic for classical estimates of location and scale to be impaired (Lind & Zumbo, 1993). For this reason it is advised that researchers perform a sensitivity

analysis to investigate the empirical question of "how robust are my findings to violations of the normality assumption?" (e.g., Thabane et al., 2013). The Hogg-Tukey Procedure (so coined by Zumbo & Jennings, 2002) suggests that both classical and robust tests should be run and their results compared: when the suitably standardized values of the test statistics are similar, then the parametric tests results can be favoured, but when their p values differ substantially the robust test results should be favoured. In the context of groups with equal variances, Zimmerman (2011) proposed that if the t-test and a test based on ranks differ by more than 0.4, then non-parametric tests should be favoured. Under the constraints of publication bias and pressure to publish, this procedure could promote p-hacking, data-dredging, or fishing for a significant result (Gelman & Loken, 2014). To avoid this, all test processes and results should be reported and decisions clearly motivated by evaluation of the data and insights from previously published research. Normality tests selected for their sensitivity to specific deviations from normality in combination with graphical techniques can be used to evaluate if normality deviations are the cause for the differences between classical, non-parametric, or modern robust test results (Looney, 1995; Seier, 2002). Choosing which statistical test is best to answer one's research question and whether normality violations impact research outcomes are an empirical question best addressed through a review of the evidence using the optimal tools at one's disposal (Zygmont & Smith, 2014). This approach is demonstrated in Appendix C using a fictional study.

Conclusion

Choosing the correct statistic for one's research question at small sample sizes is critical as bias in parameter estimates is greatest at samples below 35, but reduces dramatically in samples between 30 and 50. As sample sizes increase the Central Limit Theorem may make normality testing less important, but not completely irrelevant, as some deviations may still affect parameter estimates and standard error calculations. The results of the present and previous research suggest that formal hypothesis tests of normality should not be used as a conditional check for selecting between parametric and non-parametric/robust test statistics. In small samples they lack power to detect deviations that may matter. Sample sizes of between 50 and 100 are generally needed for GOF tests to detect violations of the normality assumption with 80% power depending on the specific distribution. However, GOF tests need to be selected that are sensitive to the types of departures characteristic of the population distribution. Overall, the Chen-Shapiro, Rahman and Govindarajulu's Shapiro-Wilk, Shapiro-Wilk, Anderson-Darling and Vasicek tests have the highest power for the greatest range of distributions at lower sample sizes. When using normality tests, the choice should be guided by theoretical assumptions, information about data distributions from previous studies, and empirical data obtained using graphical methods (Wijekularathna et al., 2022). Possibly the best approach for evaluating if violation of normality assumptions is impacting on research outcomes is to use the approach suggested by Hogg (1977a, 1977b), Tukey (1977), Zumbo and Jennings (2002), and Zimmerman (2011), and supported in this paper. This involves making the question of which test statistic to use – in light of the normality assumption – an empirical question to be answered through sensitivity analysis. Parametric, non-parametric and robust alternatives should all be performed and reported. If they produce the same answer, normality is not a concern. If there are notable differences the possibility of normality being a causal mechanism for these differences should be evaluated using the available empirical and theoretical evidence at hand. Such evidence can be collected by utilizing graphical analyses and GOF normality tests known to be powerful in detecting specific distributional violations at small samples (e.g. 50 – 100). How to perform such an analysis is demonstrated in Appendix C. It is critical to note that outliers and idiosyncratic deviations in distributions, such as bimodality, are most probably more influential for parameter estimates than solely skewness or kurtosis at small sample sizes. All analyses should be reported to promote informed, transparent, reproducible, and ethical research. Future research should examine the utility of the Hogg-Tukey procedure in varied contexts and evaluate its impact on power and Type I error across the research processes.

Author's note

The author wishes to express his appreciation to MånsT on stats.stackexchange.com for spurning interest in the topic, Pierre Lafaye de Micheaux for your assistance and the PoweR package, and to Ehsan Zamanzade for R functions for his own and Vasiczek's entropy estimators.

References

Adefisoye, J. O., Kibria, G., M., B., & George, F. (2016). Performances of several univariate tests of normality: An empirical study. *Journal of Biometrics and Biostatistics, 07*(4), 1–8. doi: 10.4172/2155-6180.1000322.

Adler, D. (2005). *Vioplot: Violin plot [r package]* (Version 0.2). http://wsopuppenkiste.wiso.uni-goettingen.de/dadler

Ahmad, F., & Sherwani, R. A. K. (2015). Power comparison of various normality tests. *Pakistan Journal of Statistics and Operation Research, 11*(3), 331–345.

Aldor-Noiman, S., Brown, L. S., Buja, A., Rolke, W., & Stone, R. A. (2013). The power to see: A new graphical test of normality. *The American Statistician*, *67*(4), 249–260. doi: 10.1080/00031305.2013.847865.

Alimohammadi, I., Ahmadi Kanrash, F., Abolghasemi, J., Shahbazi, A., Afrazandeh, H., & Rahmani, K. (2019). Combined effect of noise and smoking on cognitive performance of automotive industry workers. *Basic and Clinical Neuroscience*, *10*(5), 515–526. doi: 10.32598/bcn.10.5.513.

Alizadeh Noughabi, H., & Arghami, N. R. (2010). A new estimator of entropy. *Journal of the Iranian Statistical Society*, *9*(1), 53–64.

Alizadeh Noughabi, H., & Arghami, N. R. (2011). Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation*, *81*(8), 965–972. doi: 10.1080/00949650903580047.

Alizadeh Noughabi, H., & Arghami, N. R. (2012). Goodness-of-fit tests based on correcting moments of entropy estimators. *Communications in Statistics – Simulation and Computation*, *42*(3), 499–513. doi: 10.1080/03610918.2011.634535.

Almeida, A., Loy, A., & Hoffman, H. (2018). GGplot2 compatible quantile-quantile plots in R. *The R Journal*, *10*(2), 248–261. doi: 10.32614/RJ-2018-051.

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, *49*(268), 765–769.

Anglim, J. (2016). *Is normality testing 'essentially useless'?* Retrieved January 16, 2021, from https://stats.stackexchange.com/questions/2492

Applebaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The apa publications and communications board task force report. *American Psychologist*, *73*, 3–25. doi: 10.1037/amp0000191.

Arnastauskaitė, J., Ruzgas, T., & Bražėnas, M. (2021). An exhaustive power comparison of normality tests. *Mathematics*, *9*(7), 788–788. doi: 10.3390/math9070788.

Arnold, T. B., & Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, *3*(2), 34–39.

Australian Institute of Health and Wealfare. (2015). *Age at death*. http://www.aihw.gov.au/deaths/age-at-death/

Bakilana, A. (2005). Age at sexual debut in South Africa. *African Journal of AIDS Research*, *4*(1), 1–5. doi: 10.2989/16085900509490335.

Bakshaev, A., & Rudzkis, R. (2017). Goodness-of-fit tests based on the empirical characteristic function. *Lithuanian Mathematical Journal*, *57*(2), 155–170. doi: 10.1007/s10986-017-9350-7.

Beaujean, A. A. (2014). Sample size determination for regression models using Monte Carlo methods in R. *Practical Assessment, Research and Evaluation*, *19*(12), 1–99. doi: 10.7275/d5pv-8v28.

Bishara, A. J., Li, J., & Conley, C. (2021). Informal versus formal judgement of statistical models: The case of normality assumptions. *Psychonomic Bulletin and Review*, *28*, 1164–1182. doi: 10.3758/s13423-021-01879-z.

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, *5*, 309–335. doi: 10.2307/1164905.

Blair, R. C., & Higgins, J. J. (1980b). The power of t and Wilcoxon statistics: A comparison. *Evaluation Review*, *4*(5), 645–656. doi: 10.1177/0193841X8000400506.

Blair, R. C., & Higgins, J. J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 124–128. doi: 10.1111/j.2044-8317.1981.tb00623.x.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119–128. doi: 10.1037/0033-2909.97.1.119.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557. doi: 10.7334/psicothema2016.383.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, *50*, 937–962. doi: 10.3758/s13428-017-0918-2.

Bonett, D. G., & Seier, E. (2002). A test of normality with high uniform power. *Computational Statistics and Data Analysis*, *40*(3), 435–445. doi: 10.1016/S0167-9473(02)00074-9.

Bradley, J. V. (1980). Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society*, *15*, 275–278.

Büyükuysal, M. C., & Sümbüloğlu, V. (2021). Comparison of normality tests in terms of Type-I error and power with different sample sizes and distributions. *International Journal of Basic and Clinical Studies*, *10*(2), 57–65.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*, 1716–1735. doi: 10.3758/s13428-016-0814-1.

Chen, L., & Shapiro, S. S. (1995). An alternative test for normality based on normalized spacings. *Journal of Statistical Computation and Simulation*, *53*(3), 269–287. doi: 10.1080/00949659508811711.

Coin, D. (2008). A goodness-of-fit test for normality based on polynomial regression. *Computational Statistics and Data Analysis*, *52*(4), 2185–2198. doi: 10.1016/j.csda.2007.07.012.

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology*, *65*(1), 56–73. doi: 10.1111/j.2044-8317.2011.02014.x.

Csörgő, S. (1986). Testing for normality in arbitrary dimension. *The Annals of Statistics*, *14*(2), 708–723.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*, 341–348.

D'Agostino, R. B., & Pearson, E. S. (1973). Tests for departure from normality: Empirical results for the distributions of b^2 and b^1. *Biometrika*, *60*(3), 613–622. doi: 10.1093/biomet/60.3.613.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. Marcel Dekker, Inc.

Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, *5*(1), 5–12. doi: 10.11648/j.ajtas.20160501.12.

De la Rubia, J. M. (2022). Testing for normality from the parametric seven-number summary. *Open Journal of Statistics*, *12*(1), 118–154. doi: 10.4236/ojs.2022.121009.

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of welch's F-test instead of the classical f-test in one-way anova. *International Review of Social Psychology*, *32*(1), 13–23. doi: 10.5334/irsp.198.

Engmann, S., & Cousineau, D. (2011). Comparing distributions: The two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitative Methods*, *6*(3), 1–17.

Epps, T. W., & Pulley, L. B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, *70*(3), 723–726. doi: 10.2307/2336512.

Farrell, P. J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, *76*(9), 803–816. doi: 10.1080/10629360500109023.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics (5th ed.)* Sage.

Field, A., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. doi: 10.1016/j.brat.2017.05.013.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression (2nd ed.)* Sage.

Fried, R., & Dehling, H. (2011). Robust nonparametric tests for the two-sample location problem. *Statistical Methods and Applications*, *20*, 409–422. doi: 10.1007/s10260-011-0164-1.

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, *3*(00325), 1–20. doi: 10.3389/fpsyg.2012.00325.

Gel, Y. R., & Gastwirth, J. L. (2008). A robust modification of the Jarque–Bera test of normality. *Economics Letters*, *99*(1), 30–32. doi: 10.1016/j.econlet.2007.05.022.

Gel, Y. R., Miao, W., & Gastwirth, J. L. (2007). Robust directed tests of normality against heavy-tailed alternatives. *Computational Statistics and Data Analysis*, *51*(5), 2734–2746. doi: 10.1016/j.csda.2006.08.022.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and other stories*. Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 420–432. doi: 10.1511/2014.111.460.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*, 237–288.

Gravetter, F. J., & Wallnau, L. B. (2014). *Essentials of statistics for the behavioral sciences (8th edtition)*. Wadsworth Cengage Learning.

Grech, V., & Calleja, N. (2018). WASP (write a scientific paper): Parametric vs. non-parametric tests. *Early Human Development*, *123*, 48–49. doi: 10.1016/j.earlhumdev.2018.04.014.

Halvorsen, K. B. (2019). *Are large data sets inappropriate for hypothesis testing?* Retrieved November 29, 2019, from https://stats.stackexchange.com/questions/2516

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*(4), 315–339. doi: 10.2307/1165127.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184. doi: 10.2307/2685478.

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Are assumptions of well-known statistical techniques

checked, and why (not)? *Frontiers in Psychology*, *3*(137), 1–99. doi: 10.3389/fpsyg.2012.00137.

Hogg, R. V. (1977a). An introduction to robust estimation. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 1–17). Academic Press.

Hogg, R. V. (1977b). An introduction to robust procedures. *Communications in Statistics – Theory and Methods*, *6*(9), 789–794. doi: 10.1080/03610927708827531.

Hopper, T. (2014). *Normality and testing for normality*. Retrieved March 21, 2014, from http://www.r-bloggers.com/normality-and-testing-for-normality/

Howell, D. C. (2013). *Statistical methods for psychology (8th)*. Wadsworth Cengage Learning.

Hu, Y., & Plonsky, L. (2019). Statistical assumptions in l2 research: A systematic review. *Second Language Research*, *37*, 171–184. doi: 10.1177/0267658319877433.

Huang, K.-W., Qiao, M., Liu, X., Liu, S., & Dai, M. (2019). An application of Q-Q Plot for normality test. https://arxiv.org/pdf/1901.07851.pdf

Islam, T. U. (2017). Stringency-based ranking of normality tests. *Communications in Statistics – Simulation and Computation*, *46*(1), 655–668. doi: 10.1080/03610918.2014.977916.

Islam, T. U. (2019). Ranking of normality tests: An appraisal through skewed alternative space. *Symmetry*, *11*(7), 872–899. doi: 10.3390/sym11070872.

Janic, A., & Ledwina, T. (2009). Data-driven smooth tests for a location-scale family revisited. *Journal of Statistical Theory and Practice*, *3*(3), 645–664. doi: 10.1080/15598608.2009.10411952.

Jäntschi, L., & Bolboacă, S. D. (2009). Distribution fitting 2: Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Cramer-von-Misses and Jarque-Bera statistics. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca*, *66*(2), 691–697. http://arxiv.org/pdf/0907.2832.pdf

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, *55*(2), 163–172. doi: 10.2307/1403192.

Keller, G. (2018). *Statistics for management and economics (11th ed.)* Cengage Learning.

Kellner, J., & Cellise, A. (2019). A one-sample test for normality with kernel methods. *Bernoulli*, *25*(3), 1816–1837. doi: 10.3150/18-BEJ1037.

Khan, A., & Rayner, G. D. (2003). Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics and Decision Sciences*, *7*(4), 187–206. doi: 10.1155/S1173912603000178.

Kim, T. K., & Park, J. H. (2019). More about the basic assumptions of t-test: Normality and sample size. *Korean Journal of Anaesthesiology*, *72*(4), 331–335. doi: 10.4097/kja.d.18.00292.

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*, 2576–2590. doi: 10.3758/s13428-021-01587-5.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale Istituti Attuari*, *4*, 883–891.

Kozak, M., & Piepo, H.-P. (2017). What's normal anyway? residual plots are more telling than significance tests when checking anova assumptions. *Journal of Agronomy and Crop Science*, *204*(1), 86–98. doi: 10.1111/jac.12220.

Lafaye de Micheaux, P., & Tran, V. A. (2016). PoweR: A reproducible research tool to ease Monte Carlo power simulation studies for goodness-of-fit tests in R. *Journal of Statistical Software*, *69*(3), 1–41. doi: 10.18637/jss.v069.i03.

Lang, T. A., & Altman, D. G. (2016). Statistical analyses and methods in the published literature: The SAMPL guidelines. *Medical Writing*, *25*(3), 31–36.

Lantz, B., Andersson, R., & Manfredsson, P. (2016). Preliminary tests of normality when comparing three independent samples. *Journal of Modern Applied Statistical Methods*, *15*(2), 11–99. doi: 10.22237/jmasm/1478002140.

Lee, A. F. S., & Gurland, J. (1977). One-sample t-test when sampling from a mixture of normal distributions. *Annals of Statistics*, *5*, 803–807.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 8. doi: 10.2307/2283970.

Lind, J. C., & Zumbo, B. D. (1993). The continuity principle in psychological research: An introduction to robust statistics. *Canadian Psychology / Psychologie Canadienne*, *34*, 407–414. doi: 10.1037/h0078861.

Lindstromberg, S. (2020). *The assumptions of normality and same-shape distributions in relation to commonly used tests of a difference between two samples* [ResearchGate Preprint]. doi: 10.13140/RG.2.2.12207.56481.

Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality. *The American Statistician*, *49*(1), 64–70.

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of Q-Q plots: The power of our eyes. *The American Statistician*, *70*(2), 202–214. doi: 10.1080/00031305.2015.1077728.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*(1),

151–169. doi: 10.1146/annurev.publhealth.23.100901.140546.

Masiero, M., Lucchiari, C., Maisonneuve, P., Pravettoni, G., Veronesi, G., & Mazzacco, K. (2019). The attentional bias in current and former smokers. *Frontiers in Behavioural Neuroscience, 13*, 154. doi: 10.3389/fnbeh.2019.00154.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166.

Miecznikowski, J. C., Vexler, A., & Shepherd, L. (2013). dbEmpLikeGOF: An R package for nonparametric likelihood ratio tests for goodness-of-fit and two-sample comparisons based on sample entropy. *Journal of Statistical Software, 54*(3), 1–99. doi: 10.18637/jss.v054.i03.

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia, 22*(1), 67–72. doi: 10.4103/aca.ACA_157_18.

Montenegro, S., & Alonso, J. C. (2015). Estudio de Monte Carlo para comparar pruebas de normalidad sobre residuos de mínimos cuadrados ordinarios en presencia de procesos autorregresivos de primer orden. *Estudios Gerenciales, 31*(136), 253–265.

Montgomery, D. C., & Runger, G. C. (2011). *Applied statistics and probability for engineers (5th edition)*. John Wiley & Sons inc.

Mória, T. F., Székelyba, G. J., & Rizzo, M. L. (2021). On energy tests of normality. *Journal of Statistical Planning and Inference, 213*, 1–15. doi: 10.1016/j.jspi.2020.11.001.

National Association for Law Placement. (2015). *Salary distribution curves*. Retrieved January 1, 2023, from http://www.nalp.org/salarydistrib

Orcan, F. (2020). Parametric or non-parametric: Skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education, 7*(2), 255–265. doi: 10.21449/ijate.656077.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation, 8*(2), 1–5. doi: 10.7275/r222-hv23.

Paolella, M. S. (2018). *Fundamental statistical inference: A computational approach*. John Wiley & Sons.

Parra-Frutos, I. (2016). Preliminary tests when comparing means. *Computational Statistics, 31*(4), 1607–1631. doi: 10.1007/s00180-016-0656-4.

Pedrosa, I., Juarros-Basterretxea, J., Robles-Fernández, A., Basteiro, J., & García-Cueto, E. (2015). Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar? *Universitas Psychologica, 14*(1), 245–254. doi: 10.11144/Javeriana.upsy14-1.pbad.

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology, 9*(02104), 1–20. doi: 10.3389/fpsyg.2018.02104.

Quessy, J.-F., & Mailhot, M. (2011). Asymptotic power of tests of normality under local alternatives. *Journal of Statistical Planning and Inference, 141*(8), 2787–2802. doi: 10.1016/j.jspi.2011.03.003.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rahman, M. M., & Govindarajulu, Z. (1997). A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics, 24*(2), 219–236. doi: 10.1080/02664769723828.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 22–33.

Revelle, W. (2016). *Psych: Procedures for psychological psychometric, and personality research [R package]* (Version 1.6.12). https://CRAN.R-project.org/package=psych

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*, 81–99. http://www.biomedcentral.com/1471-2288/12/81

Rochon, J., & Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *British Journal of Mathematical and Statistical Psychology, 64*, 410–426. doi: 10.1348/2044-8317.002003.

Romão, X., Delgado, R., & Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation, 80*(5), 545–591. doi: 10.1080/00949650902740824.

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2017). Beyond differences in means: Robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience, 46*(2), 1738–1748. doi: 10.1111/ejn.13610.

Rousselet, G. A., & Wilcox, R. R. (2020). Reaction times and other skewed distributions: Problems with the mean and median. *Meta-Psychology, 4*, 1630. doi: 10.15626/MP.2019.1630.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*(2), 352–360. doi: 10.1037/0033-2909.111.2.352.

Schick, A., Wang, Y., & Wefelmeyer, W. (2011). Tests for normality based on density estimators of convolutions.

Statistics and Probability Letters, *81*(2), 337–343. doi: 10.1016/j.spl.2010.10.022.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*, 1141–1152. doi: 10.1111/2041-210X.13434.

Schröder, C., & Yitzhaki, S. (2017). Reasonable sample sizes for convergence to normality. *Communications in Statistics - Simulation and Computation*, *46*(9), 7074–7087. doi: 10.1080/03610918.2016.1224347.

Schucany, W. R., & Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics - Theory and Methods*, *35*(12), 2275–2286. doi: 10.1080/03610920600853308.

Seier, E. (2002). *Comparison of tests for univariate normality*. Retrieved January 1, 2023, from http://interstat.statjournals.net/YEAR/2002/articles/0201001.pdf

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3), 591–599. doi: 10.2307/2333709.

Shatz, I. (2023). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behaviour Research Methods*, *online first*, 1–99. doi: 10.3758/s13428-023-02072-x.

Silva-Lugo, J. L., Warner, L. A., & Galindo, S. (2021). From parametric to non-parametric statistics in education and agricultural eduction research. *The Journal of Agricultural Education and Extension*, *28*(4), 393–413. doi: 10.1080/1389224X.2021.1936089.

Silverfish. (n.d.). *How to choose between t-test or non-parametric test e.g. wilcoxon in small samples*. Retrieved October 29, 2014, from https://stats.stackexchange.com/questions/121852

Sürücü, B. (2008). A power comparison and simulation study of goodness-of-fit tests. *Computers and Mathematics with Applications*, *56*(6), 1617–1625. doi: 10.1016/j.camwa.2008.03.010.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Chenglin, Y., & Goldsmith, C. H. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Medical Research Methodology*, *13*(92), 1–20. doi: 10.1186/1471-2288-13-92.

Thadewald, T., & Büning, H. (2007). Jarque–bera test and its competitors for testing normality – a power comparison. *Journal of Applied Statistics*, *34*(1), 87–105. doi: 10.1080/02664760600994539.

Thériault, R. (2022). *Rempsyc: Convenience functions for psychology [R package]* (Version 0.1.3). https://rempsyc.remi-theriault.com

Thode, H. C. J. (2002). *Testing for normality*. CRC Press.

Tukey, J. W. (1977). Robust techniques for the user. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 103–106). Academic Press.

U. S. Bureau of Labor Statistics. (2015). *American time use survey*. Retrieved January 1, 2023, from https://www.bls.gov/tus/

United State Census Bureau. (2013). Current population survey (CPS. http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

University of Wisconsin-Madison. (2017). *Course grade distributions*. Retrieved January 1, 2023, from https://registrar.wisc.edu/course_grade_distributions.htm

Uyanto, S. S. (2022). An extensive comparison of 50 univariate goodness-of-fit tests for normality. *Austrian Journal of Statistics*, *51*, 45–97. doi: 10.17713/ajs.v51i3.1279.

van den Brink, W. P., & van den Brink, S. J. G. (1989). A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *British Journal of Mathematical and Statistical Psychology*, *42*, 183–189. doi: 10.1111/j.2044-8317.1989.tb00907.x.

Van Zyl, J. M. (2017). The performance of univariate goodness-of-fit tests for normality based on the empirical characteristic function in large samples. *Communications in Statistics - Simulation and Computation*, *47*(4), 1146–1156. doi: 10.1080/03610918.2017.1307397.

van Zandt, T. (2002). Analysis of response time distributions. In J. Wixted (Ed.), *Steven's handbook of experimental psychology (3rd edition)* (pp. 461–516). Wiley.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society*, *38*, 54–59.

Warrington, N. M., Tilling, K., Howe, L. D., Paternoster, L., Pennell, C. E., Wu, Y. Y., & Briollais, L. (2014). Robustness of the linear mixed effects model to error distribution assumptions and the consequences for genome-wide association studies. *Statistical Applications in Genetics and Molecular Biology*, *13*, 567–587. doi: 10.1515/sagmb-2013-0066.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108.

Weisberg, S. (2014). *Applied linear regression (4th ed.)* John Wiley & Sons.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482. doi: 10.1007/BF03395630.

Wickham, H. (2016). *GGplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wijekularathna, D. K., Manage, A. B. W., & Scariano, S. M. (2022). Power analysis of several normality tests: A

Monte Carlo simulation study. *Communications in Statistics - Simulation and Computation*, *51*(3), 757–773. doi: 10.1080/03610918.2019.1658780.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, *32*(7), 771–780.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. Springer.

Wilcox, R. R. (2012). *Modern statistics for the social and behavioural sciences: A practical introduction*. CRC Press.

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing (5th edition)*. Academic Press.

Wilcox, R. R., & Rousselet, G. A. (2023). An updated guide to robust statistical methods in neuroscience. *Current Protocols*, *3*(3), 1–31. doi: 10.1002/cpz1.719.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155. doi: 10.1080/00949655.2010.520163.

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*(2), 175–183. doi: 10.1080/10629360600678310.

Zamanzade, E., & Arghami, N. R. (2012). Testing normality based on new entropy estimators. *Journal of Statistical Computation and Simulation*, *82*(11), 1701–1713. doi: 10.1080/00949655.2011.592984.

Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality. *British Journal of Mathematical and Statistical Psychology*, *64*, 388–409. doi: 10.1348/000711010X524739.

Zuma, K., Mzolo, T., & Makonko, E. (2011). Determinants of age at sexual debut and associated risks among South African youths. *African Journal of AIDS Research*, *10*(3), 189–194. doi: 10.2989/16085906.2011.626283.

Zumbo, B. D., & Jennings, M. J. (2002). The robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicológica*, *23*(2), 415–450. http://www.redalyc.org/articulo.oa?id=16923209

Zygmont, C. S., & Smith, M. R. (2014). Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *The Quantitative Methods for Psychology*, *10*(1), 40–55.

Table 2 ■ Goodness-of-fit test statistics used in this study

Abbr.	Reference Test Statistic and Notations R function and Package
Tests based on the empirical distribution function (EDF)	
AD	The Anderson-Darling test (Anderson & Darling, 1954; D'Agostino & Stephens, 1986) $A = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1][ln(p_{(i)}) + ln(1-p_{(n-i+1)})]$, where $p_{(i)} = \Phi([x_{(i)} - \overline{x}]/s)$ `ad.test()` in `nortest`
LKS	The Lilliefors (1967) modification of the Kolmogorov-Smirnov test $D = \max\{D^+, D^-\}$ with $D^+ = \max_{i=1,\ldots,n}\{i/n - p_{(i)}\}$, $D^- = \max_{i=1,\ldots,n}\{p_{(i)} - (i-1)/n\}$, where $p(i) = \Phi([x_{(i)} - \overline{x}]/s)$ `lillie.test()` in `nortest`
Tests based on measures of the moments	
B-S	The Bonett-Seier (2002) test $T_w = \frac{\sqrt{n+2}.(\hat{w}-3)}{3.54}$ where $\hat{w} = 13.29\left[ln\sqrt{m_2} - ln\left(n^{-1}\sum_{i=1}^{n}|x_i - \overline{x}|\right)\right]$ `statcompute(17, ...)` in `PoweR`
DP	The D'Agostino-Pearson (1973) K^2 test $K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)$ where $Z(\sqrt{b_1})$ and $Z(b_2)$ are the normal approximations for skewness and kurtosis `statcompute(6, ...)` in `PoweR`
JB	The Jarque-Bera (1987) test $LM_N = N\left\{\frac{\sqrt{\hat{b}_1}}{6} + \frac{(\hat{b}_2-3)^2}{24}\right\}$ where $\hat{b}_1 = \frac{\hat{u}_3^2}{\hat{u}_2^3}$ and $\hat{b}_2 = \frac{\hat{u}_4}{\hat{u}_2^2}$ `statcompute(7, ...)` in `PoweR`
RJB	The Robust Jarque-Bera test (Gel & Gastwirth, 2008). $\text{RJB} = \frac{n}{6}\left(\frac{m_3}{J_n^3}\right)^2 + \frac{n}{64}\left(\frac{m_4}{J_n^4} - 3\right)^2$ where $J_n = \frac{\sqrt{\pi/2}}{n}\sum_{i=1}^{n}\lvert x_i - M\rvert$ `statcompute(9, ...)` in `PoweR`
Regression and correlation tests	
COIN	The B_3^2 Coin (2008) test

	$z_{(i)} = \beta_1.\alpha_i + \beta_3.\alpha_i^3$, where β_1 and β_3 are fitting parameters and a_i are expected values of standard normal order statistics `statcompute(30, ...)` in `PoweR`
CS	The Chen-Shapiro (1995) test $CS = \frac{1}{(n-1).s}\sum_{i=1}^{n-1}\frac{x_{(i+1)}-x_{(i)}}{M_{i+1}-M_i}$ `statcompute(26, ...)` in `PoweR`
DAGO	D'Agostino (1971) omnibus test $D = \frac{\sum_{i=1}^{n}(i+\frac{n+1}{2})X_{(i)}}{n^2\sqrt{m_2}}$ `statcompute(24, ...)` in `PoweR`
RG-SW	The Rahman and Govindarajulu (1997) modification of the Shapiro-Wilk test $a_i = -(n+1)(n+2)\phi(m_i)[m_{i-1}\phi(m_{i-1}) - 2m_i\phi(m_i) + m_{i+1}\phi(m_{i+1})]$ where $m_0\phi(m_0) = m_{n+1}\phi(m_{n+1}) = 0$ `statcompute(23, ...)` in `PoweR`
SW	The Shapiro-Wilk (1965) test $W = \frac{\left\{\sum_{i=1}^{k} a_{(n-i+1)}(x_{(n-i+1)} - x_{(i)})\right\}^2}{\sum_{i-1}^{n}(x_i - \overline{x})^2}$ where $x_{(i)}$ s are the ordered statistics `shapiro.test()` in `nortest`
Empirical likelihood GOF tests based on sample entropy	
DBEG	Miecznikowski et al. (2013) $V_n = \min_{1\le m<n^{1-\delta}}(2\pi e s^2)^{n/2}\prod_{i=1}^{n}\frac{2m}{n(X_{(i+m)}-X_{(i-m)})}$ `DbEmpLikeGOF()` in package of same name
VAS	Vasicek (1976) test $TV_{mn} = \frac{exp\{HV_{mn}\}}{\hat{\sigma}}$, where $HV_{mn} = \frac{1}{n}\sum_{i=1}^{n} log\left\{\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\right\}$ `Entropy.Tests()` and `HV()` functions provided in Appendix A
ZAM	Zamanzade and Arghami (2012) $TZ2_{mn}$ test $TZ2_{mn} = \frac{exp\{HZ2_{mn}\}}{\hat{\sigma}}$, where $HZ2_{mn} = \sum_{i=1}^{n} w_i log\{b_i\}$ with $w_i = \begin{cases} \frac{m+i-1}{\sum_{i=1}^{n} w_i} & \text{if } 1 < i < m, \\ \frac{2m}{\sum_{i=1}^{n} w_i} & \text{if } m+1 < i < n-m \\ \frac{n-i+m}{\sum_{i=1}^{n} w_i} & \text{if } n-m+1 < i < n \end{cases}$ `Entropy.Tests()` and `HZ2()` functions provided in Appendix A
Empirical characteristic function class tests	
ECF	Van Zyl's (2017) empirical characteristic function test. $v_n(1) = \log\left(\left\|\widehat{\phi_S}(1)\exp(-1/2)\right\|\right)$, where $\sqrt{n}(v_n(1)) \sim N(0, 0.0431)$ `ecf()` function provided in Appendix A.
EP	The Epps-Pulley test (Epps & Pulley, 1983) $T(\alpha) = n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n}\exp\left\{-\frac{1}{2}\left(X_j - X_k\right)^2/\left(\alpha^2 S^2\right)\right\} - 2n^{-1}\left(1+\alpha^{-2}\right)^{-\frac{1}{2}}\sum_{j=1}^{n} exp\left[-\frac{1}{2}\left(X_j - \bar{X}\right)^2/\left\{S^2\left(1+\alpha^2\right)\right\}\right] + \left(1+2\alpha^{-2}\right)^{-\frac{1}{2}}$ `statcompute(31, ...)` in `PoweR`
Other tests	
GMGW	The Gel-Miao-Gastwirth R_{sJ} test (2007) $R_{sJ} = s/J_n$ where $J_n = \frac{\sqrt{\pi/2}}{n}\sum_{i=1}^{n}\lvert x_i - M\rvert$ `statcompute(33, ...)` in `PoweR`
DDST	Data driven smooth test (Janic & Ledwina, 2009). $W_k^*(\tilde{\gamma}) = \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell^*(Z_i;\tilde{\gamma})\right]\left[\tau^*(\tilde{\gamma})\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell^*(Z_i;\tilde{\gamma})\right]'$, where $\tilde{\gamma}$ is an approriate estimator of γ and $\tau^*(\tilde{\gamma}) = Cov_{\theta_0}\left[(\ell^*(Z_i;\tilde{\gamma})\right]'\left[\ell^*(Z_i;\tilde{\gamma})\right]$ `ddst.norm.test()` in `ddst`

Appendices follows.

Appendix A ■ Code used for functions not contained in R packages found in the CRAN

```
# Functions for entropy based tests
# Kernel density estimator
f <- function(a,x){
  n=length(x);h=((4/(3*n))^(1/5))*sd(x);
  p=(1/(n*h))*sum(dnorm((a-x)/h))
  return(p)
}

#Vasicek's Entropy estimator
HV <- function(x,m){
  n=length(x)
  x=sort(x)
  H=c()
  c=2
  for(i in 1:n) {
    a1=i-m
    a2=i+m
    if(a1<1)a1=1
    if(a2>n)a2=n
    H=c(H,log(n*(x[a2]-x[a1])/(c*m)))
  }
  return(mean(H))
}

#Zamanzade's second entropy estimator
HZ2=function(x,m){
  x<-sort(x)
  n<-length(x)
  Z2<-c()
  W<-c()
  for(i in 1:n){
    if(i<=m) W[i]<-m+i-1
    if(i>m&&i<=(n-m)) W[i]<-2*m
    if(i>(n-m)) W[i]<-n-i+m
    k1<-max(1,(i-m))
    k2<-min(n,(i+m))
    d<-0
    for(j in k1:(k2-1)) d<-(f(x[j+1],x)+f(x[j],x))/2*(x[(j+1)]-x[j])+d
    a<-x[k2]-x[k1]
    Z2[i]<-log(a/d)
  }
  W<-W/sum(W)
  return(sum(W*Z2))
}
#Function to run either Vasicek or Zamanzade's test
#arguments include: x = data, m = window size (integer smaller than sample size/2), and test is
     1 for Vasicek or 2 for Zamanzade
Entropy.Tests<-function(x,m,test){
  if(test==1) H<-HV(x,m)
  if(test==2) H<-HZ2(x,m)
```

```
  S<-sqrt(sum((x-mean(x))^2)/length(x))
  return(exp(H)/S)
}

#Function for van Zyl's ECF test
ecf <- function(t, x) {
  Vectorize( function(t) mean(exp(complex(real=0, imaginary=1)*t*x))  )(t)
}

zyl_test <- function(x) log( Mod(ecf(1, scale(x))/exp(-1/2))  )

# To run the test:
zyl_test(IQ$N)/(sqrt(0.0431/n)
```

Appendix B ■ Power tables

The following three tables (B.1 to B.3) provide the power for each normality test for six different distributions within each of the three categories at sample sizes of 20, 40 and 80.

Table B.1 ■ Power of 18 normality tests across different mixed distributions and sample sizes

Type	Size	LKS	AD	SW	DP	JB	RJB	BS	RG-SW	DAGO	CS	COIN	GMGW	VAS	ZAM	DBEG	DDST	EP	ECF
Age at death	20	.46	.61	.65	.52	.40	.53	.19	.62	.43	.65	.18	.45	.54	.56	.52	.57	.64	.37
	40	.78	.90	.92	.81	.79	.82	.33	.88	.71	.92	.29	.64	.80	.87	.79	.88	.92	.61
	80	.98	1	1	.98	.98	.98	.53	.99	.93	1	.47	.85	.96	.99	.95	.99	1	.88
Time spent eating	20	.10	.16	.19	.16	.10	.13	.12	.23	.11	.20	.11	.15	.24	.12	.22	.15	.15	.09
	40	.16	.30	.42	.32	.20	.20	.19	.46	.20	.44	.20	.24	.48	.23	.46	.31	.27	.15
	80	.28	.58	.80	.53	.40	.36	.27	.83	.29	.82	.26	.33	.83	.50	.83	.58	.48	.24
Income	20	.62	.69	.71	.69	.67	.69	.56	.68	.65	.69	.53	.66	.64	.68	.64	.68	.71	.63
	40	.87	.91	.93	.91	.90	.91	.83	.91	.88	.92	.81	.88	.88	.91	.88	.91	.92	.87
	80	.98	.99	1	.99	1	.99	.97	.99	.98	1	.96	.98	.99	.99	.99	1	.99	.98
Age of sexual debut	20	.10	.14	.18	.17	.10	.15	.09	.16	.11	.17	.09	.14	.13	.14	.12	.15	.15	.14
	40	.14	.25	.32	.27	.21	.24	.11	.25	.16	.32	.13	.18	.19	.24	.19	.27	.25	.21
	80	.25	.44	.56	.42	.39	.39	.13	.38	.26	.54	.19	.22	.29	.43	.25	.45	.43	.32
Lawyer starting salary	20	.92	.96	.93	.60	.01	.06	.78	.93	.12	.94	.58	.53	.95	.61	.92	.95	.80	.01
	40	1	1	1	.81	.03	.03	.96	1	.15	1	.74	.72	1	.99	1	1	.96	.08
	80	1	1	1	.83	.67	.38	1	1	.25	1	.89	.86	1	1	1	1	1	.70
Student GPA score	20	.12	.18	.20	.14	0	.01	.23	.34	.07	.22	.24	.19	.41	.04	.39	.17	0	0
	40	.26	.46	.52	.48	0	0	.52	.72	.21	.57	.61	.51	.74	.14	.75	.50	0	0
	80	.57	.85	.93	.85	.09	0	.87	.98	.58	.95	.94	.88	.98	.58	.98	.89	0	.17

Table B.2 ■ Power of 18 normality tests across different symmetric distributions and sample sizes

Type	*Size*	LKS	AD	SW	DP	JB	RJB	BS	RG-SW	DAGO	CS	COIN	GMGW	VAS	ZAM	DBEG	DDST	EP	ECF
Student's *t*(5)	20	.13	.17	.19	.23	.17	.27	.16	.13	.19	.18	.19	.27	.08	.22	.07	.20	.18	.17
	40	.18	.27	.31	.35	.34	.43	.30	.17	.32	.29	.35	.41	.11	.36	.08	.33	.28	.31
	80	.29	.42	.47	.54	.54	.62	.51	.25	.55	.46	.56	.61	.20	.55	.13	.54	.42	.53
Logistic	20	.08	.11	.12	.15	.09	.17	.10	.07	.11	.11	.12	.19	.05	.14	.04	.13	.12	.10
	40	.10	.14	.17	.21	.19	.27	.16	.08	.17	.16	.21	.27	.05	.21	.04	.19	.16	.17
	80	.14	.21	.27	.31	.32	.39	.28	.10	.31	.22	.32	.40	.08	.33	.05	.29	.21	.30
Tukey (−0.25)	20	.29	.36	.37	.42	.33	.49	.36	.25	.39	.37	.38	.50	.18	.44	.15	.41	.37	.36
	40	.48	.60	.60	.63	.62	.73	.65	.41	.65	.59	.66	.75	.32	.69	.25	.65	.60	.62
	80	.73	.84	.86	.85	.87	.93	.90	.66	.90	.84	.90	.94	.62	.90	.50	.89	.85	.88
Tukey (0.75)	20	.07	.11	.13	.11	0	0	.16	.26	.09	.15	.20	.14	.31	.02	.28	.11	.09	0
	40	.14	.30	.39	.48	0	0	.37	.66	.36	.45	.63	.40	.66	.06	.67	.38	.28	0
	80	.32	.69	.88	.93	.05	0	.75	.99	.85	.91	.98	.80	.97	.39	.98	.82	.68	.23
Tukey (1.05)	20	.10	.18	.22	.17	0	0	.22	.39	.09	.22	.31	.19	.45	.03	.42	.17	.13	0
	40	.20	.46	.61	.65	0	0	.51	.85	.36	.67	.81	.53	.86	.15	.86	.56	.44	0
	80	.48	.88	.98	.98	.17	0	.87	1	.88	.99	1	.90	1	.66	1	.94	.86	.47
Uniform	20	.10	.17	.20	.16	0	0	.22	.37	.09	.21	.30	.19	.43	.03	.40	.17	.13	0
	40	.19	.44	.58	.64	0	0	.50	.83	.38	.65	.80	.53	.84	.13	.84	.54	.42	0
	80	.45	.86	.97	.98	.16	0	.86	1	.88	.98	1	.89	1	.64	1	.93	.85	.45

Table B.3 ■ Power of 18 normality tests across different asymmetric distributions and sample sizes

Type	Size	LKS	AD	SW	DP	JB	RJB	BS	RG-SW	DAGO	CS	COIN	GMGW	VAS	ZAM	DBEG	DDST	EP	ECF
Weibull (2, 3)	20	.10	.13	.15	.13	.07	.12	.06	.15	.09	.15	.05	.11	.14	.11	.12	.11	.14	.05
	40	.17	.25	.32	.23	.16	.20	.07	.33	.11	.33	.06	.13	.26	.20	.26	.23	.30	.08
	80	.32	.50	.67	.44	.38	.38	.08	.68	.14	.68	.08	.14	.53	.42	.57	.53	.56	.11
Pareto (0, 2, 0)	20	.58	.78	.84	.60	.49	.59	.20	.84	.52	.83	.15	.47	.85	.68	.84	.71	.78	.40
	40	.90	.98	1	.90	.88	.89	.33	1	.80	.99	.23	.67	1	.96	1	.97	.98	.66
	80	1	1	1	1	1	1	.51	1	.97	1	.32	.87	1	1	1	1	1	.90
Gumbel (0, 2)	20	.20	.27	.31	.29	.21	.29	.12	.26	.19	.31	.10	.24	.21	.28	.19	.26	.31	.17
	40	.35	.50	.59	.50	.46	.51	.18	.52	.33	.57	.14	.35	.40	.48	.40	.52	.57	.29
	80	.63	.81	.88	.81	.80	.80	.27	.84	.55	.88	.21	.48	.70	.79	.69	.84	.85	.51
Gamma (2, 3)	20	.22	.32	.37	.31	.21	.29	.11	.35	.20	.37	.09	.22	.29	.30	.27	.30	.37	.18
	40	.42	.61	.71	.55	.48	.53	.14	.68	.33	.72	.11	.31	.57	.57	.58	.61	.68	.29
	80	.73	.92	.97	.87	.86	.84	.20	.96	.56	.97	.13	.42	.90	.88	.92	.92	.93	.48
Asymmetric power (0, 2, 0.8, 1.5)	20	.26	.36	.41	.31	.20	.30	.10	.37	.21	.40	.08	.23	.31	.31	.29	.32	.40	.17
	40	.49	.67	.72	.56	.49	.54	.14	.68	.35	.72	.11	.32	.56	.59	.55	.63	.71	.30
	80	.80	.94	.96	.88	.87	.86	.17	.94	.59	.96	.13	.44	.84	.90	.84	.94	.95	.49
Asymmetric laplace (4, 2, 2)	20	.47	.62	.64	.52	.41	.54	.20	.58	.43	.63	.19	.44	.49	.56	.49	.56	.64	.35
	40	.79	.90	.91	.82	.79	.82	.34	.86	.70	.91	.31	.65	.78	.87	.77	.88	.92	.60
	80	.98	1	1	.99	.99	.98	.54	.99	.94	1	.49	.86	.97	.99	.96	.99	1	.86

Appendix C ■ Demonstration

This supplementary material is intended to demonstrate the procedure advocated in the article manuscript for deciding between classical GLM models, non-parametric test statistics, and robust methods while taking into account the effect of possible violations of the normality assumption. The data used in this vignette are response time data from a hypothetical study of the effects of nicotine on the colour-Stroop task using smoking-related cues (e.g. tobacco, cigar, smoke). In the hypothetical study we are comparing the performance of a sample of 45 non-smokers (have not smoked within the last six months) and 62 active smokers (were actively smoking just before the trials start). The participants are otherwise similar in terms of demographic, medical, and cognitive performance metrics. Generally, research shows that active smokers should complete Stroop trials faster but have more errors; but when presented with smoking-related cues, smokers take longer than non-smokers on the Stroop task (Alimohammadi et al., 2019; Masiero et al., 2019). Instructions are provided for replicating these methods using the R statistical programming environment (R Core Team, 2023). If you do not already have R installed on your PC, follow the following generic instructions available online to install R and R-Studio, which I recommend as an IDE (e.g.: https://teacherscollege.screenstepslive.com/a/1108074-install-r-and-r-studio-for-windows). If you are new to R, an introductory course would be recommended (e.g. https://learndigital.withgoogle.com/digitalgarage/course/introduction-to-r, https://www.datacamp.com/courses/free-introduction-to-r or https://coursera.org/learn/r-programming). Readers can download from the journal's web site the dataset used in this supplementary document in order to replicate the analysis themselves. Load the data into R and look at the structure of the data using the following commands (substituting `/FileLocation/ResponseTimeData.Rdata` with the file location on your computer):

```
load("/FileLocation/ResponseTimeData.Rdata")
str(RTdata)
```

Next one should install the required packages containing the functions we will use later (if you do not already have them). The following commands will only install those packages that are not already installed on your computer from the Comprehensive R Archive Network (CRAN). You will also need to install `PoweR` and `dbEmpLikeGOF` from the archive, as they are no longer available on CRAN, and the `rogme` package from github as it is not available yet on CRAN:

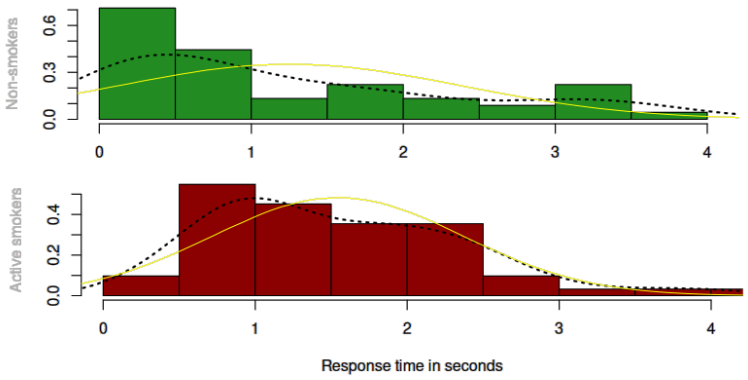
```
RP <- c("car","dplyr","rbtt","rempsyc","rstatix","devtools","nortest","vioplot")
install.packages(setdiff(RP, rownames(installed.packages())))
devtools::install_version("PoweR", version = "1.0.7")
devtools::install_version("dbEmpLikeGOF", version = "1.2.4")
devtools::install_github("GRousselet/rogme")
```

Research Question

We will be examining whether there is a difference in response times based on smoking group. Traditionally, if testing the normality assumption, most researchers would have run a Kolmogorov-Smirnov test (Arnold & Emerson, 2011; Engmann & Cousineau, 2011; Pedrosa et al., 2015). Having done so they would have concluded that both non-smoker ($D = 0.18$, $p = .09$) and active smoker distributions are normally distributed ($D = 0.10$, $p = .57$). They would then most likely have followed on with a parametric Welch t-test. However, using a more powerful normality test like the Shapiro-Wilk or Anderson-Darling would have resulted in rejecting normality for the distribution of response times for non-smokers ($W = 0.88$, $p < .001$; $A = 1.97$, $p < .001$) and active smokers ($W = 0.95$, $p = .017$; $A = 0.8$, $p = .036$). Rather than using this two-step process or ignoring the normality assumption and assuming that the central limit theorem (CLT) will protect against normality deviations, it was argued in the article manuscript that a sensitivity analysis would be preferable to make the influence of assumptions on our test statistic an empirical question. We will compare the results to our research question using four methods: the independent samples Welch t-test, Wilcoxon rank sum test, robust bootstrapped t-test, and t-test based on trimmed mean. If there is no significant difference in these results (Zimmerman, 2011), we can conclude that violation of assumptions was either negligible or protected by the CLT as a result of sufficient sample size. If the differences among these test statistics are notable, then we can use graphical methods and normality tests to diagnose the cause and guide our selection of results. By reporting all analyses and using all the information available to motivate an outcome we are engaging in transparent, informed, and ethical research.

We can run the parametric Welch t-test, the non-parametric Wilcoxon rank-sum test, robust bootstrapped t-test, and Yuen's t-test based on trimmed means and with bootstrapped effect size confidence intervals using the following R commands:

Figure C.1 ■ Histogram and density plots for sample distribution and normal distribution. The solid yellow line denotes a normal distribution with the same mean and standard deviation as the data, the dotted black line denotes the smoothed density distribution of the sample data




```
t.test(RT~group, alternative='two.sided', conf.level=.95, data=RTdata, var.equal=F)

wilcox.test(RT ~ group, alternative='two.sided', data=RTdata, exact=TRUE)

rstatix::wilcox_effsize(RT ~ group, paired=F, data=RTdata, ci=T, ci.type="bca")
rbtt::rbtt(x = RTdata$RT[RTdata$group == "Active smoker"],
           y = RTdata$RT[RTdata$group == "Non-smoker"], n.boot=50000, method=2)
WRS2::yuen(RT ~ group, data = RTdata, tr = 0.2)
WRS2::yuen.effect.ci(RT ~ group, data = RTdata, tr = 0.2, nboot = 10000)
```

The Welch two-sample t-test ($t(76.34) = 1.47$, $p = .146$, 95% CI [-0.10, 0.68]) and the robust bootstrapped t-test assuming unequal variance ($t = 1.47$, $p = .157$, 95% CI [-0.12, 0.70]) both suggest that there are no significant differences in response time between smokers and non-smokers. On the other hand, the Wilcoxon rank sum test ($W = 1761$, $p = .02$; $r = 0.22$, 95% CI [0.03, 0.41]) and Yuen's t-test based on 20% trimmed means ($t(df = 46.69) = 2.02$, $p = 0.049$, 95% CI [0.002, 0.92]; $d = 0.30$, 95% CI [0.04, 0.52]) suggest there is a significant difference in response times with a small effect size. The t-tests are aimed at testing if the populations means are equal. On the other hand, the Wilcoxon rank sum test is testing if the medians are equal, or more specifically, if the distribution of ranked scores of one group tend to have higher scores than the other. According to Zimmerman (2011), the robust findings should be favoured; and graphical methods, descriptive statistics, as well as normality tests can be used to ascertain why the test results vary and validate the choice of methods.

Graphical methods

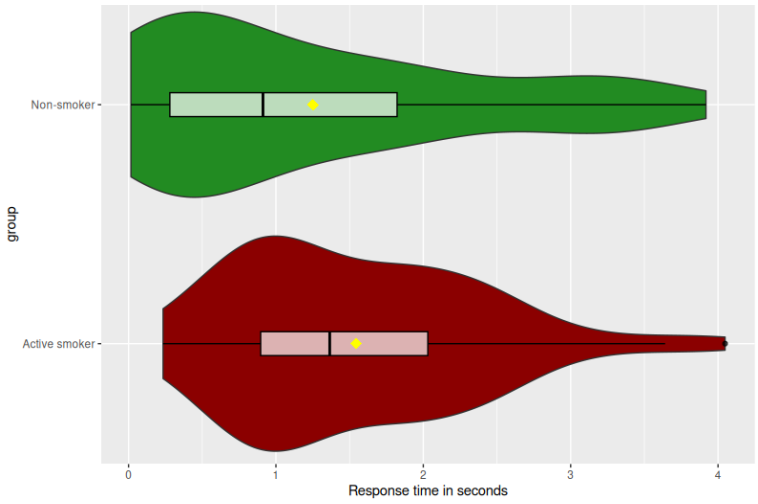
Graphical methods are particularly useful for identifying the causes of normality violations. Histograms and density plots allow for a quick intuitive evaluation of the general shape of any distribution of sample data. These are demonstrated in Figure C.1, and can be achieved using the following commands:

```
par(mfrow = c(2,1), bg="lightgrey")

hist(RTdata$RT[RTdata$group == "Non-smoker"] , prob=TRUE , breaks=8 , col="forestgreen",
     xlim=c(min(RTdata$RT),max(RTdata$RT)), freq=FALSE, xlab="" , ylab="Non-smokers",
     main="", col.lab="darkgrey")

lines(density(RTdata$RT[RTdata$group == "Non-smoker"] , adjust=1) , lty="dotted",
     col="black" , lwd=2)
```

Figure C.2 ■ Violin plot of response times compared by group. The yellow triangle denotes the sample mean for each group, a boxplot is overlaid in order to display the median, inter-quartile range, and any outliers (represented by the solid black circle in the active smoker group)




```
lines(density(rnorm(1e+07 , mean(RTdata$RT[RTdata$group == "Non-smoker"]),
    sd(RTdata$RT[RTdata$group == "Non-smoker"]))) , lty="solid" , col="yellow" , lwd=1)

hist(RTdata$RT[RTdata$group == "Active smoker"] , prob=TRUE , breaks=8 , col="darkred",
    xlim=c(min(RTdata$RT),max(RTdata$RT)), freq=FALSE, xlab="" , ylab="Active smokers",
    main="", col.lab="darkgrey")

title(xlab="Response time in seconds", col.lab="black")

lines(density(RTdata$RT[RTdata$group == "Active smoker"] , adjust=1) , lty="dotted",
    col="black" , lwd=2)

lines(density(rnorm(1e+07 , mean(RTdata$RT[RTdata$group == "Active smoker"]),
    sd(RTdata$RT[RTdata$group == "Active smoker"]))) , lty="solid" , col="yellow", lwd=1)
```

Another very useful method for plotting data, which makes the quartiles, median and outliers easily identifiable, as well as a quick evaluation of centrality, scatter and skewness, is the boxplot. Alternatively, the violin plot (see Figure C.2) combines the visual representation of shape from a (mirrored) kernel density estimate with the key summary statistics of a boxplot (Adler, 2005; Hintze & Nelson, 1998). Producing a violin plot of the response time data using the `ggplot2` package (Wickham, 2016) is possible with the following code:

```
ggplot2::ggplot(RTdata,aes(x=group, y=RT, fill=group)) +
    geom_violin() +
    scale_fill_manual(values=c("darkred", "forestgreen")) +
    geom_boxplot(width=0.1, fill="white",color="black",alpha=0.7) +
    stat_summary(fun = mean, geom = "point", shape = 18, size =4, color = "yellow") +
    coord_flip() + theme(legend.position = "none") + ylab("Response time in seconds")
```

Quantile-quantile and probability plots (see Figure C.3) are also very useful for detecting outliers and other deviations from normality. Skewness is visible when the slope of the data is steep initially and then levels out (negative skew) or is

fairly level initially and then increases exponentially upwards towards the end (positive skew). The Q-Q plot tends to have a stretched-out N shape (lower end drops below the line, higher end rises above the line) for long-tailed distributions and a stretched-out S shape (lower end above the line and higher end below the line) for short-tailed distributions. Outliers are visible as points that deviate notably from the line relative to other points. Probability plots have the advantage of discriminating in regions of high probability density, consequently deviations in the middle of a Gaussian distribution are more apparent (Das & Imon, 2016).

The `qqPlot` function in the car package (Fox & Weisberg, 2011) is particularly useful for interpretation as it includes a 95% confidence interval and allows the user to reference against a variety of distributions. This allows one to visually test whether there is a better fit to a Gaussian or another distribution. The same can be achieved using the `ggplot2` (Zimmerman, 2011) and `qqplotr` (Almeida et al., 2018) functions. The following code can be used to produce the Q-Q plot in Figure C.3 using the RT data:

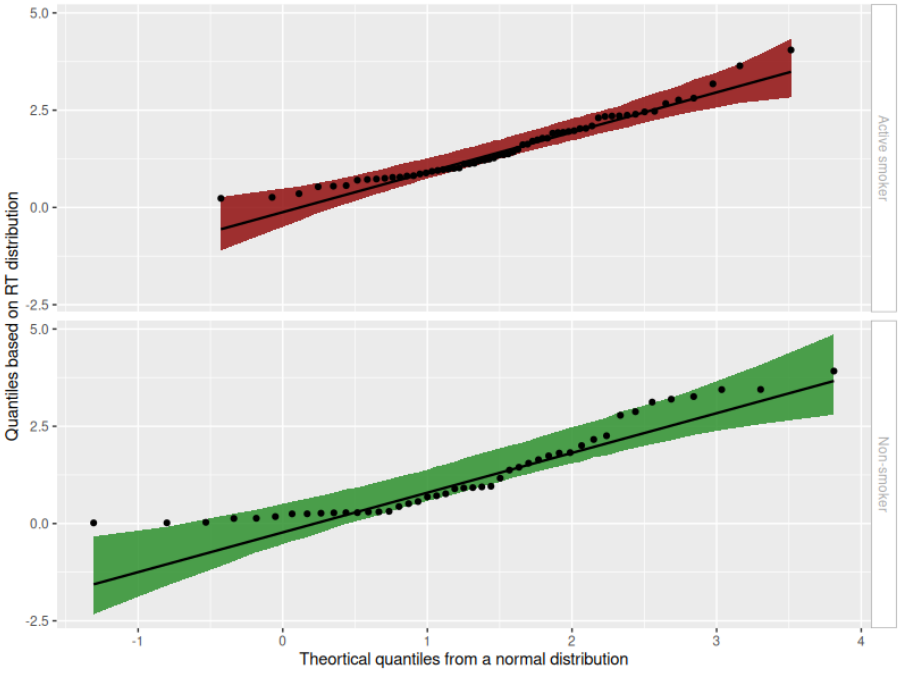
```
library(qqplotr)

ggplot2::ggplot(RTdata,aes(sample=RT, fill=group)) +
   stat_qq_band(bandType = "boot") + stat_qq_line() + stat_qq_point() +
   labs(x = "Theortical quantiles from a normal distribution", y="Quantiles based on RT
    distribution") +
   scale_fill_manual(values=c("darkred","forestgreen")) +
   facet_grid(rows= vars(group)) +
   theme(legend.position = "none", strip.background = element_rect(fill="white",
     color="darkgrey"), strip.text = element_text(color="darkgrey"))
```

The detrended probability plot (see Figure C.4) may be more intuitive for some people to quantify and compare deviations, it produces a plot with deviations along the range of quantiles presented horizontally. It can be produced using the code provided below:

```
# Detrended probability plot based on code by Vincent Zoonekynd
# Load the function using the following:
detrended.prob.plot <- function (dat, xlab="Standardised normal scores",
   ylab="Deviation from normal", main="", colordat="blue")
{
   x <- sort(na.omit(dat))
   a1 <- (quantile(x, .75) - quantile(x, .25)) / 1.34898
   a0 <- quantile(x, .25) - a1 * -0.6744898
   x <- x - (a0 + a1 * qnorm(ppoints(length(x))))
   y <- qnorm(ppoints(length(x)))
   plot(x ~ y, xlab=xlab, ylab=ylab, main=main, col=colordat, pch=19, col.main="darkgrey")
   abline(h=0, col="black")
}
```

Figure C.3 ■ Q-Q plot of numerical IQ scores




```
# Run the plot with the following:
par(mfrow = c(2,1), bg="lightgrey")

detrended.prob.plot(dat = RTdata$RT[RTdata$group == "Non-smoker"],
   colordat = "forestgreen", xlab="", main="Non-smoking group")

detrended.prob.plot(dat = RTdata$RT[RTdata$group == "Active smoker"],
   colordat = "darkred", main="Active smoking group")
```

Descriptive statistics

When data are normally distributed measures of central tendency (mean, median, and mode) should be equal. Measures of skewness and kurtosis provide valuable indicators of deviations from normal. The `describeBy` function that forms part of the excellent `psych` package (Revelle, 2016) can be used to obtain these statistics for each group using the following code:

```
psych::describeBy(RTdata$RT , RTdata$group, trim=0.2, type=3)
```

Descriptive statistics concur that response times for non-smokers have a slightly lower central tendency, flatter and wider distribution, and are more positively skewed with a slightly longer right tail ($\bar{X} = 1.25$, $\bar{X}_{trimmed(20\%)} = 0.99$, $md = 0.91$, $sd = 1.13$, $mad = 0.98$, $skew = 0.81$, $kurtosis = -0.63$) when compared to that of active smokers ($\bar{X} = 1.54$, $\bar{X}_{trimmed(20\%)} = 1.45$, $md = 1.36$, $sd = 0.83$, $mad = 0.85$, $skew = 0.72$, $kurtosis = 0.17$). The descriptive statistics and graphical analyses taken together suggest that active smokers tend to have less variation in response times and deviate less from a normal distribution, although both groups are positively skewed. This corresponds with the general understanding that response times are positively skewed in the population (van Zandt, 2002). We could test whether the deviation from a Gaussian distribution is statistically significant using the Anderson-Darling, DBEG, Gel-Miao-Gastwirth, and Shapiro-Wilk tests. The DBEG and Gel-Miao-Gastwirth statistics are particularly sensitive to deviations from normal within symmetrical distri-

Figure C.4 ■ Detrended normal probability plot

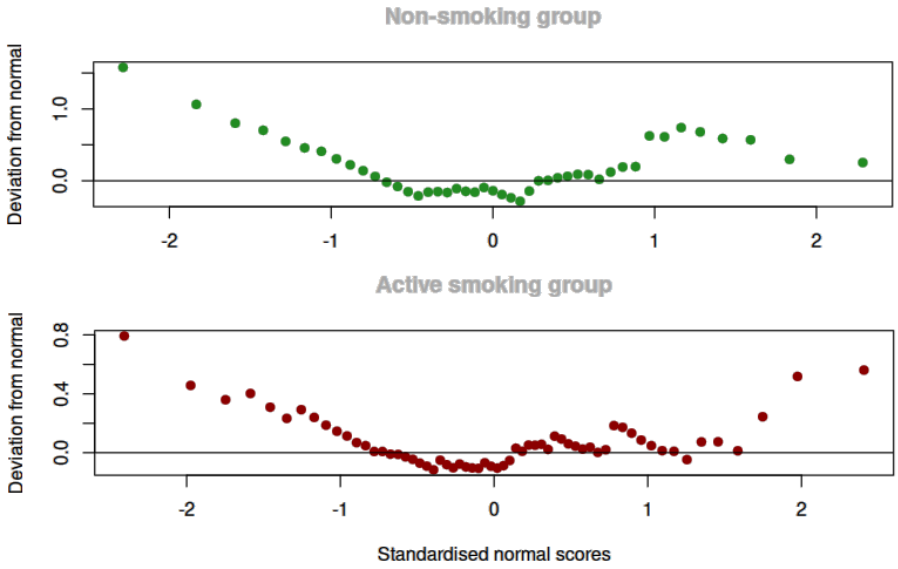


Table C.1 ■ The p value from various normality tests for each smoking condition

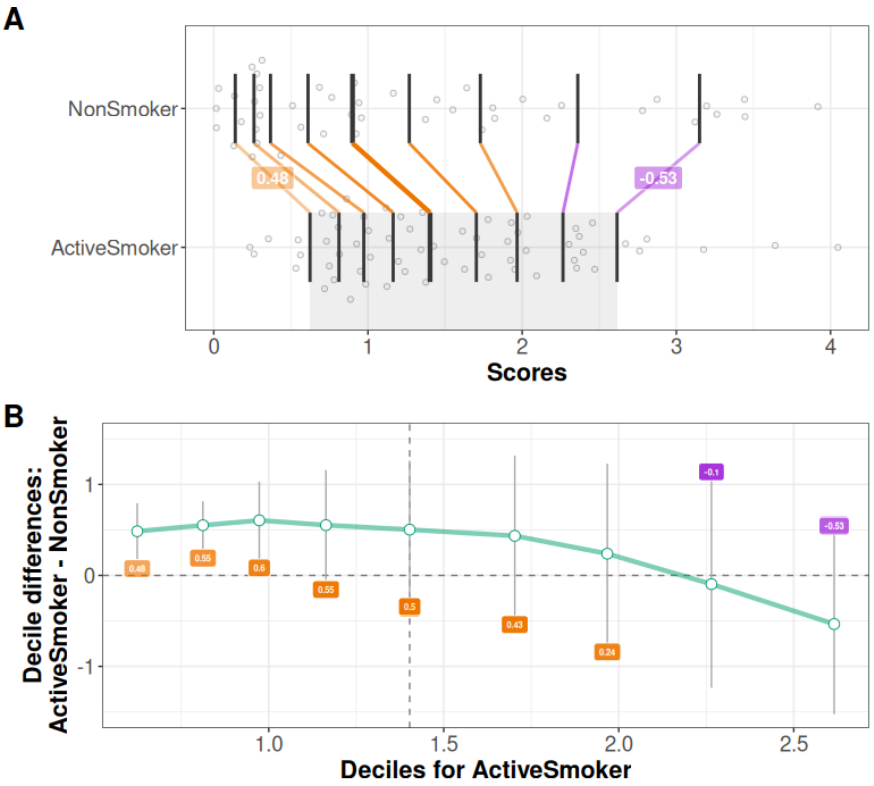
group	AndersonDarling	DBEG	GelMiaoGastwirth	ShapiroWilk
ActiveSmoker	.04	.06	.28	.02
NonSmoker	.00	.00	.49	.00

butions, whereas Shapiro-Wilk and Anderson-Darling are more sensitive to asymmetrical and mixed distributions. The results are formatted into Table 1, using the `nice_table()` function from the `rempsyc` package (Thériault, 2022) using the following code:

```
library(dplyr)
RTdata %>% group_by(group) %>%
   summarise(AndersonDarling = nortest::ad.test(RT)$p.value, DBEG = dbEmpLikeGOF::dbEmpLikeGOF(
    x=RT, testcall="normal", pvl.Table = F, num.mc=50000)$pvalue, GelMiaoGastwirth = PoweR::
    statcompute(33, RT)$pvalue, ShapiroWilk = shapiro.test(RT)$p.value) %>%
   rempsyc::nice_table(,title=c("Table 1","The p values from various normality tests across
    groups"))
```

These findings suggest that firstly, the choice of normality test is important (Gel-Miao-Gastwirth has low power in asymmetric distributions), and secondly that the RT data are not distributed according to a normal/Gaussian distribution in each group. In this case we should favour the results from the robust statistics reported earlier, supported by the graphical analysis and normality test results. We could conclude that while there may not be a significant difference in group means (based on standard and robust *t*-tests), there does seem to be a significant difference in response time distributions between groups (based on the Wilcoxon and Yuen's tests). When working with response time data, it is advisable to analyse data using distribution functions, which work well especially with larger sample sizes (Whelan, 2008). Hierarchical shift functions are particularly useful (Rousselet & Wilcox, 2020). This can be done using the rogme package (Rousselet et al., 2017), which shows that for the first three deciles non-smokers response times are statistically significantly faster than the active smokers, and with the difference becoming smaller from the 3rd to 7th decile, until active smokers have marginally faster response times in the last two deciles (see figure C.5). Taking into account the positive skew in the distributions, this would explain why we would expect a person picked at random from the non-smoking group to respond faster than

Figure C.5 ■ A comparison between active smokers and non-smoker response times across deciles. Plot A provides a 1-dimensional scatterplot for each group with differences between deciles colour-coded, and the median identified by a thicker marker. Plot B provides an indication of the differences between groups at each decile point with 95% bootstrap confidence intervals for each comparison.



someone picked at random from the active smoking group, even if the overall means are not significantly different. The code used to produce graphs of the decile comparison across groups is provided below:

```
library(ggplot2)
library(rogme)

shiftfunc <- shiftHD(RTdata, RT ~ group, nboot = 10000)
DecComPlot <- SFplot(shiftfunc, plot_theme = 1, symb_size = 3, diffint_col = "darkgrey")
DecComPlot <- add_sf_lab(psf, shiftfunc, y_lab_nudge = .1, text_size = 2)
ScatPlot <- plot_scat2(data = RTdata, formula = RT ~ group, xlabel = "", ylabel = "Scores",
    alpha = .3, shape = 21, colour = "grey10", fill = "grey90") + coord_flip()
DecPlot <- plot_hd_links(ScatPlot, shiftfunc[[1]], q_size = 1, md_size = 1.5, add_rect = TRUE,
    rect_alpha = 0.1, rect_col = "grey50", add_lab = TRUE, text_size = 5)
CompPlot <- cowplot::plot_grid(DecPlot, DecComPlot[[1]], labels=c("A", "B"), ncol = 1,
   nrow = 2, rel_heights = c(1, 1), label_size = 20, vjust = 1, scale=.95)
```

Conclusion

The analysis demonstrated here has illustrated the importance of informed, comprehensive, and transparent data analysis processes. Following the default two-step procedure -– of using the Kolmogorov-Smirnov test to check for normality

followed by the *t*-test — one would have concluded that there is no difference between groups. Ignoring the normality assumption would also have led to an incomplete understanding of the research outcomes. Using the approach originally suggested by Hogg (1977a, 1977b) and Tukey (1977), discussed by Zumbo and Jennings (2002), and further developed by Zimmerman (2011); this demonstration has shown the value of running both classical tests and robust tests together, and then using graphical and normality tests selected for their sensitivity to different deviations from normality, to explain any significant differences between classical and robust test outcomes. In most cases the results should be the same, but when they differ substantially all the results should be reported in the interest of ethical and transparent research, and the procedures demonstrated here used to guide interpretation.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on the journal's web site.

Citation

Zygmont, C. S. (2023). Managing the assumption of normality within the General Linear Model with small samples: Guidelines for researchers regarding if, when and how. *The Quantitative Methods for Psychology*, *19*(4), 302–332. doi: 10.20982/tqmp.19.4.p302.

Received: 20/07/2023 ~ Accepted: 03/11/2023