The Power to See: A New Graphical Test of Normality
Author(s): Sivan Aldor-Noiman, Lawrence D. Brown, Andreas Buja, Wolfgang Rolke and Robert A. Stine
Source: *The American Statistician*, Vol. 67, No. 4 (NOVEMBER 2013), pp. 249-260
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: https://www.jstor.org/stable/24591489
Accessed: 05-05-2020 17:27 UTC

The Power to See: A New Graphical Test of Normality

Sivan ALDOR-NOIMAN, Lawrence D. BROWN, Andreas BUJA, Wolfgang ROLKE, and Robert A. STINE

Many statistical procedures assume that the underlying data-generating process involves Gaussian errors. Among the popular tests for normality, only the Kolmogorov–Smirnov test has a graphical representation. Alternative tests, such as the Shapiro–Wilk test, offer little insight as to how the observed data deviate from normality. In this article, we discuss a simple new graphical procedure which provides simultaneous confidence bands for a normal quantile–quantile plot. These bands define a test of normality and are narrower in the tails than those related to the Kolmogorov–Smirnov test. Correspondingly, the new procedure has greater power to detect deviations from normality in the tails. Supplementary materials for this article are available online.

KEY WORDS: Confidence bands; Graphical presentation; Normality test; Power analysis; Quantile–quantile plot.

1. INTRODUCTION

Statistical procedures often assume that the underlying data follow a Gaussian distribution. Among common tests for normality, only the Kolmogorov–Smirnov test has a graphical representation. Alternatives, such as the Shapiro–Wilk test, offer little insight as to how the observed data deviate from normality. This article introduces a simple method that provides simultaneous confidence bands for a normal quantile–quantile (Q–Q) plot. These bands define a test of normality and are narrower in the tails than those associated with the Kolmogorov–Smirnov test. Correspondingly, this new procedure has greater power to detect deviations from normality in the tails. To motivate this procedure, we begin with several examples in which we test for normality. In each case, our new Tail-Sensitive (TS) simultaneous confidence bands detect a departure from normality that is missed by the classical Kolmogorov–Smirnov (KS) confidence bands.

The first example concerns measurements from experiments that test the effectiveness of body armors. The data were collected as part of a National Academies report requested by the U.S. Army ("Testing of Body Armor" 2012). The Army wanted to investigate the difference between two methods that assess how deep a bullet penetrates ceramic body armor. In the standard test, a cylindrical clay model is layered under the armor vest. A projectile is then fired against the vest, causing an indentation in the clay. The deepest impression in the clay is measured as an indication of the survivability of the soldier using this armor. The traditional method of measuring the depth of this impression uses a manually controlled digital caliper. A recently adopted method measures the impression using a computer-controlled laser. The two methods were compared in a calibration experiment involving a series of test firings measured by each method. Figure 1 shows the Q–Q plot of measurements from the experiments: the upper plots show the measurements using the digital caliper and the lower plots show the results using the laser-based approach. These plots also present the KS and proposed TS 95% confidence bands. The KS bands imply that observations from both methods are consistent with normality. In contrast, the TS bands identify a suspicious outlier in the right tail of the caliper-based measurements and several deviations from normality in the laser-based data. The TS confidence bands indicate that if the Army adopts the laser-based method it should not rely on normality to establish its safety standards.

The second data series record monthly log returns of IBM stock from March 1967 to December 2008. This series was examined in Tsay (2010), and we obtained it from the corresponding website. There are many reasons to suspect that such stock returns are not normally distributed, such as conditional volatility that produces fat tails in a distribution. Inspecting a sequence plot of these data suggests that the volatility of these returns increased in the late 1990s. The familiar KS bands, however, fail to detect this departure from normality. Figure 2 shows the time plot and the Q–Q plot of the data with both our TS 95% confidence bands and the KS confidence bands. Five of the points in the left tail fall outside the TS confidence bands but lie well inside the KS confidence bands. Therefore, the TS confidence bands show that these data do not follow a normal distribution; the log returns for this stock have a

Sivan Aldor-Noiman, The Climate Corporation, 201 Third Street, Suite 1100, San Francisco, CA 94103 (E-mail: *sivan.aldor@gmail.com*). Lawrence D. Brown (E-mail: *lbrown@wharton.upenn.edu*), Andreas Buja (E-mail: *buja.at.wharton@gmail.com*), and Robert A Stine (E-mail: *stine@wharton.upenn.edu*), Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104. Dr. Wolfgang A Rolke, Department of Mathematical Sciences, University of Puerto Rico–Mayaguez (E-mail: *wolfgang.rolke@upr.edu*). The research of Sivan Aldor-Noiman and Lawrence D. Brown was supported in part by the NSF Grant DMS-1007657.

 DOI: 10.1080/00031305.2013.847865

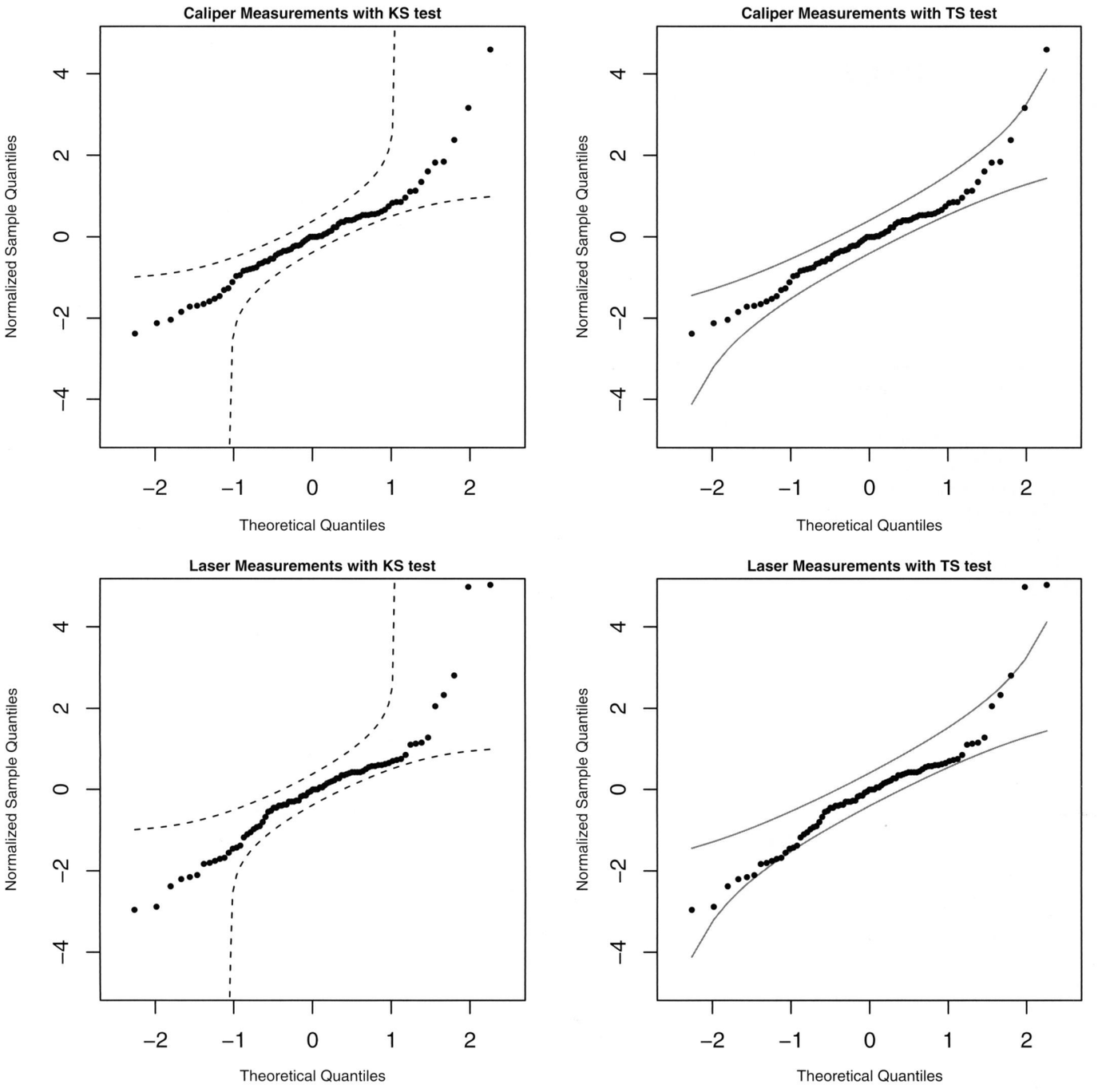

Figure 1. Measurements of bullet impressions on ceramic body armor. The upper plots shows the normal Q–Q plot of measurements taken using the digital caliper. The lower plots shows the quantile–quantile plot of measurements taken using the laser-based device. Both plots show the proposed 95% TS confidence bands and the corresponding Kolmogorov–Smirnov confidence bands.

heavier left-tail compared to the normal distribution. The KS confidence bands do not detect this deviation, and as such the researcher may wrongly conclude that the data are normally distributed.

The final example uses the smaller dataset of wave records from Bickel and Doksum (1977, p 384, Table 9.6.3). These data record the time spent above a high level for 66 wave records in San Francisco Bay. The analysis of these times in Bickel and Doksum (1977) does not reject the claim of normality at level $\alpha = 0.10$. Rosenkrantz (2000) used these data to demonstrate an improved method for constructing simultaneous confidence bands for the quantiles of a distribution. His procedure is also graphical, providing bands for the graph of the quantiles of the distribution versus the associated probabilities. His method rejects normality at level $\alpha = 0.10$. Figure 3 shows the normal Q–Q plot of the wave times with the KS bands and our TS bands at level $\alpha = 0.05$. The KS bands almost reject H_0 near the center of the data. In contrast, the TS bands detect a significant departure from normality in the right tail that the KS bands miss.

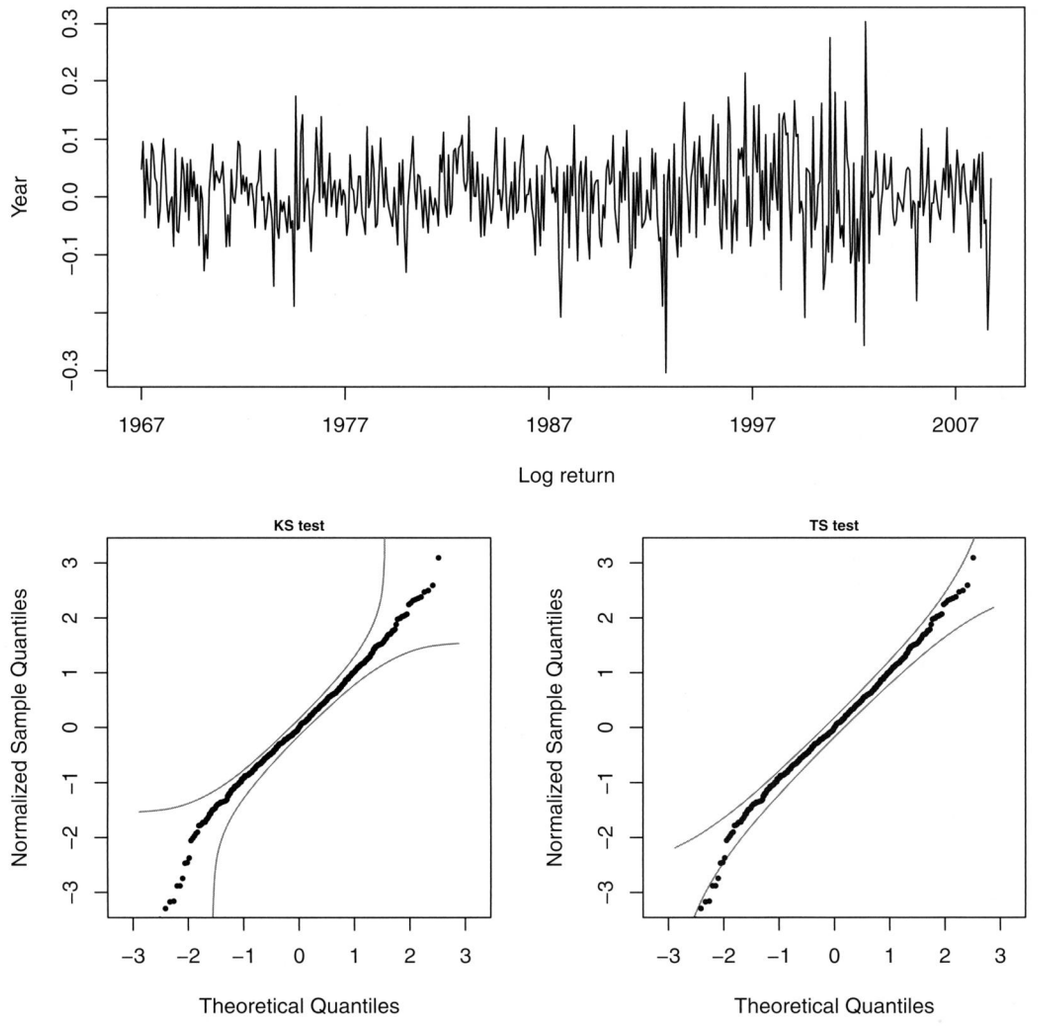

Figure 2. The monthly log returns for IBM stock from March 1967 to December 2008. The upper plot shows the time plot of the data against the time index. The lower plots shows the 95% Kolmogorov–Smirnov confidence bands and the corresponding TS confidence bands, respectively.

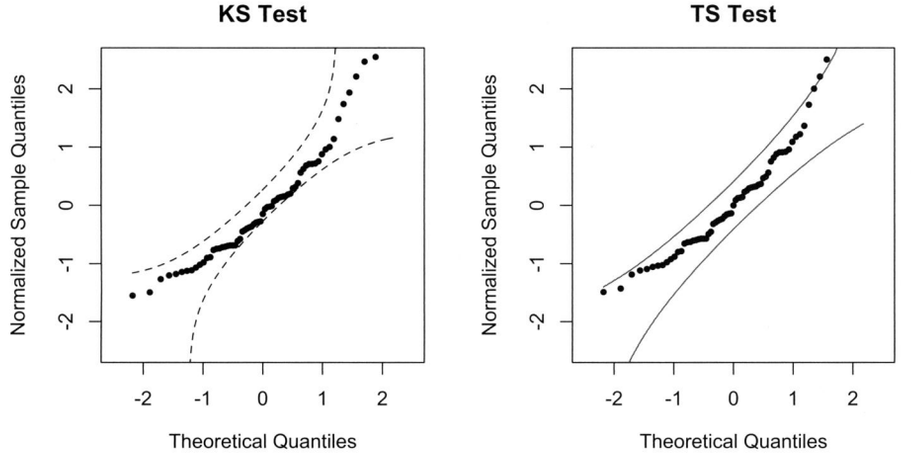

Figure 3. Wave data from Bickel and Doksum (1977). Plots show the 95% Kolmogorov–Smirnov confidence bands and the corresponding TS confidence bands, respectively. The online version of this figure is in color.

In the next section, we review common procedures that are used to test the normality assumption, both graphically and numerically. We later compare the power of these tests to that of the TS bands.

1.1 Tests for Normality

Given a sample $X_1, \ldots, X_n$, numerous statistics have been proposed to test the claim that the data are normally distributed with a given mean μ and variance σ^2. The hypotheses in question can be written as:

$$\begin{aligned} H_0 &: X_i \overset{\text{iid}}{\sim} F_0 = N(\mu, \sigma^2), \\ H_1 &: X_i \overset{\text{iid}}{\sim} F_1, \end{aligned} \tag{1}$$

where F_1 denotes an arbitrary continuous cumulative distribution function (CDF) different from that specified by H_0. We initially concentrate on the basic problem in which μ, σ^2 are specified in advance. Then, in Section 2.3, we turn to the more frequently encountered practical problem in which the mean and variance are unknown and must be estimated from the data. If μ, σ^2 are known, there is no loss of generality in assuming $\mu = 0, \sigma^2 = 1$, and we do so when considering this problem. We proceed by reviewing the more common testing procedures. Given μ, σ^2, most tests rely on a function of the deviation between the normal CDF Φ and the sample CDF $F_n(x) = \sum_i \mathbb{I}_{X_i \le x}/n$, where $\mathbb{I}$ denotes the indicator function.

In 1930, Cramèr and Von Mises (Darling 1957) presented a procedure to test the hypotheses (1). Their test statistic has the following form:

$$\omega_n = n \int_{-\infty}^{\infty} (F_n(t) - F_0(t))^2 dF_0(t).$$

Faraway and Csorgo (1996) studied the asymptotic distribution of ω_n.

A few years later, Kolmogorov (1933) derived the distribution of the maximum deviation between an empirical distribution and the distribution of the underlying population. The resulting test statistic, denoted by D_n, can be written as:

$$D_n = \sqrt{n} \sup_{-\infty < t < \infty} |F_n(t) - F_0(t)|.$$

For applications, Smirnov (1939) (and more widely available in Smirnov 1948) provided tables of the asymptotic null distribution of D_n under H_0. For example, if the data are a sample from F_0, these tables give the value c_α such that $\lim_n P(\sup_t |F_n(t) - F_0(t)| \le c_\alpha/\sqrt{n}) \approx 1 - \alpha$. Massey (1951) and later Birnbaum (1952) computed critical values for D_n for finite sample sizes. These limits presume, however, that F_0 is known whereas in practice it is more common for H_0 to specify the shape of the population up to one or two unknown parameters. For instance, H_0 might specify a normal population with some mean μ and variance σ^2. Estimating these parameters from data requires adjusting the critical values for D_n. The necessary adjustments were unknown until computers became more widely available; Lilliefors (1967) used a simulation to obtain a revised table of percentiles. The adjustments for estimation are larger than one might expect and explain why many had found a lack of power using the KS test. If μ and σ^2 are estimated by $\overline{X}$ and s^2, then the asymptotic critical value for $\sup |F_n - F_0|$ is about $0.89/\sqrt{n}$ rather than $1.36/\sqrt{n}$ for $\alpha = 0.05$. Without this adjustment for estimation, the KS test lacks power.

Tests of H_0 based on D_n have a visual rendering that displays the corresponding confidence bands. For example, Massey (1951) illustrate D_n by displaying bands of the form $\{x, F_0(x) \pm c_\alpha/\sqrt{n}\}$ around the hypothesized cumulative distribution. The test rejects H_0 unless the empirical distribution F_n lies fully inside these limits. Alternatively, one can place bands in a Q–Q plot. This link to graphical presentation probably contributes to the enduring popularity of the KS test among practitioners. The limits for D_n are distribution-free, so they imply limits on samples from a uniform distribution for which $F_0(x) = x$. In this case, the bands on F_0 are parallel diagonal lines $\{x, x \pm c_\alpha/\sqrt{n}\}$ that are truncated at 0 and 1. To obtain bands for the quantiles of the observed sample, the limits are transformed by the inverse distribution function back to the scale of the data. Because of the truncation of the bands near 0 and 1, these limits are useless in the tails of the data. The effects of this truncation are apparent in Figures 1–3. For example, the lower bound tends to $-\infty$ well before the minimum of the data. For further details on how these bands are constructed, the reader is referred to DasGupta (2011).

More recent tests for normality return to the style of the test proposed by Cramèr and von Mises. Not long after the tables of D_n became available, Anderson and Darling (1954) suggested the following test statistic:

$$A_n = n \int_{-\infty}^{\infty} \frac{(F_n(t) - F_0(t))^2}{F_0(t) \cdot (1 - F_0(t))} dF_0(t),$$

where A_n measures the weighted average squared deviation between the empirical CDF and the hypothesized CDF. Its distribution was documented in Anderson and Darling (1954). We can view A_n as a weighted version of the Cramèr-von Mises statistic which introduces the weight function $[F_0(t) \cdot (1 - F_0(t))]^{-1}$. By using this weight function, Anderson and Darling placed more emphasis on the deviation at the tails of the distribution, precisely where the KS test lacked power. Similar to the Kolmogorov–Smirnov test, the properties of the Anderson–Darling statistic were also examined for the case of unknown parameters, and Stephens (1974) gave tables to compute the adjusted p-values.

Tests based on A_n, ω_n, and D_n can be used for any specified null distribution, not just the normal. In contrast, Shapiro and Wilk (1965) derived a test statistic specifically designed to test whether the observed values are generated from a normal distribution with unknown parameters. Their test statistic takes the form

$$\begin{aligned} W_n &= \frac{b^2}{s^2} = \frac{(c\,\tilde{\sigma})^2}{s^2}, \\ \tilde{\sigma} &= \frac{\boldsymbol{m}^t \boldsymbol{V}^{-1} \boldsymbol{y}}{\boldsymbol{m}^t \boldsymbol{V}^{-1} \boldsymbol{m}}, \\ s^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \end{aligned} \tag{2}$$

where $\boldsymbol{y} = [Y_{(1)}, \ldots, Y_{(n)}]$ is the vector of the sample order statistics, and $\boldsymbol{m} = [m_1, \ldots, m_n]$ and $\boldsymbol{V} = [v_{ij}]$ are the corresponding expected values and covariance matrix of the standard normal order statistics. (In later computations, we use the

approximation suggested in Wilk and Gnanadesikan (1968) to evaluate $\boldsymbol{m}$ and $\boldsymbol{V}$.) The statistic b is, up to the normalizing constant denoted c, the estimated slope of the generalized least squares regression of the ordered observations on the expected values of the standard normal order statistics. Both b and s estimate the population standard deviation σ, but b is robust. Wilk and Gnanadesikan (1968) list critical values of W_n for various sample sizes.

Additional studies, such as Stephens (1974) and Razali and Wah (2011), showed that the Kolmogorov–Smirnov test is generally the least powerful test among those previously described, while the Shapiro–Wilk test is generally the most powerful of the group. The only clear advantage the Kolmogorov–Smirnov test possesses is its visual presentation. By visually inspecting the deviations on this plot, the researcher may be able to better understand and possibly correct the nonnormality in the data by a simple transformation or might be able to propose some different underlying process.

In the next section, we describe our new TS procedure. Just as simulation methods improved the power of KS statistics for models with estimated parameters, our TS procedure again leverages computing to improve the power of tests based on D_n by providing more useful bounds in the tails of the distribution. For most alternatives this procedure is as powerful as the Shapiro–Wilk test and can be depicted in a Q–Q plot like KS, as illustrated in Figures 1–3. We not only describe how to apply our method to examine the normal distribution hypothesis, but we also note that the testing procedure can be modified to test any other continuous distribution.

2. TAIL-SENSITIVE CONFIDENCE BANDS

In this section, we describe in detail our graphical method and the corresponding test. We introduce the procedure assuming F_0 is fully specified, and then show how to incorporate estimated parameters.

2.1 Bands for Fully Specified Null Distributions

To test whether n observations $X_1, \ldots, X_n$ are normally distributed with known mean μ_0 and standard deviation σ_0, we first standardize the data to have mean 0 and standard deviation 1 as $Z_i = (X_i - \mu_0)/\sigma_0$. We then construct the TS confidence bands and check whether the Q–Q plot of the normalized sample falls entirely inside the confidence bands.

We construct the TS confidence bands using the uniform distribution and then invert them to the normal scale by using the inverse standard normal CDF, Φ^{-1}. Forming the appropriate confidence bands for the uniform distribution requires two steps.

1. Build individual $1 - \gamma$ confidence intervals for each desired quantile of the uniform distribution. The quantiles of a uniform distribution follow a known Beta distribution, simplifying this task.
2. Adjust the confidence level, γ, to account for multiplicity so as to obtain simultaneous confidence bands with coverage $1 - \alpha$.

Step 1 defines the shape of a nested collection of pointwise confidence intervals indexed by the coverage parameter γ. Step 2 then identifies the nominal confidence level γ that defines a size α test. Establishing the necessary nominal coverage γ is analytically intractable, and therefore we determine it by simulation. Atkinson (1981) similarly used simulation to establish an envelope for a half-normal plot of regression residuals. Our approach also resembles the use of double bootstrap resampling to calibrate the coverage of a confidence interval (Beran 1987). We can avoid a second layer of simulation because the distribution of $Y_{(i)}$ is known. For plots, we apply the inverse normal CDF to obtain bounds for the normal Q–Q plot. This approach to the construction of the TS bands is a special case of a general framework for simultaneous inference called Calibration for Simultaneity (*CfS*) introduced in Buja and Rolke (2006).

Here are the details of the computational algorithm:

Step 1. Form prediction intervals for order statistics. Let $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$ denote the order statistics of a sample of size n from the uniform distribution on [0, 1]. From elementary probability theory, $Y_{(i)}$ follows a Beta distribution with shape parameters $\alpha = i$ and $\beta = n + 1 - i$; we denote the associated CDF by $B_{(i,n+1-i)}$. To construct a $1 - \gamma$ level prediction interval for $Y_{(i)}$, choose Beta quantiles $L_i(\gamma)$ and $U_i(\gamma)$ such that $P(L_i(\gamma) \le Y_{(i)} \le U_i(\gamma)) = 1 - \gamma$. For example, the equal-tail quantiles are $L_i(\gamma) = B^{-1}_{(i,n+1-i)}(\frac{\gamma}{2})$ and $U_i(\gamma) = B^{-1}_{(i,n+1-i)}(1 - \frac{\gamma}{2})$.

Step 2. Adjust the coverage level γ for simultaneity. Step 1 produces a nested collection of intervals because $[L_i(\gamma_1), U_i(\gamma_1)] \subset [L_i(\gamma_2), U_i(\gamma_2)]$ if $\gamma_1 > \gamma_2$. Reducing the Type I error γ expands the region covered by the intervals and thus increases the simultaneous coverage of the collection. All that we need to do is to find γ such that

$$P(L_i(\gamma) \le Y_{(i)} \le U_i(\gamma),\ \forall i) = 1 - \alpha.$$

For any choice of γ, we could use a simulation to find the simultaneous coverage to whatever precision we require. Simply draw many samples $Y_1, \ldots, Y_n \sim U[0, 1]$ and find the proportion of samples whose order statistics lie within the bounds. Though conceptually simple, computing is not yet so fast to allow such profligate number crunching, and so we rely on a statistical shortcut that takes advantage of the known distribution of $Y_{(i)}$.

To find γ, we draw M samples from the uniform distribution. Let $Y^m_{(i)}$ denote the order statistics of the mth sample $Y^m_1, \ldots, Y^m_n \sim U[0, 1]$, with $m = 1, \ldots, M$. For each of the M samples, we find the smallest value of γ for which the collection of intervals produced in Step 1 covers all of the quantiles. Assume that we are using intervals with error rates $\gamma/2$ in each tail. The value of γ that covers the mth sample is

$$\begin{aligned} C^m &= \min\left\{c : L_i(c) \le Y^m_{(i)} \le U_i(c),\ \forall i\right\} \\ &= \min\left\{c : B^{-1}_{i,n+1-i}(c/2) \le Y^m_{(i)} \le B^{-1}_{i,n+1-i}(1 - c/2),\ \forall i\right\} \end{aligned}$$

$$= \min\left\{c : c/2 \le B_{i,n+1-i}\left(Y^m_{(i)}\right) \le 1 - c/2,\ \forall i\right\}$$
$$= 2\min_i\left\{\min(B_{i,n+1-i}(Y^m_{(i)}), 1 - B_{i,n+1-i}(Y_{(i)})\right\}. \quad (3)$$

C_m is the smallest two-sided p-value for the mth simulated sample. To obtain simultaneous coverage $1-\alpha$ (up to simulation error), we set γ to the 100α-percentile over $[C^1, C^2, \ldots, C^M]$. The procedure guarantees that only a fraction α of the datasets have an order statistic outside the confidence bands. Buja and Rolke (2006) suggested alternative calculations needed for settings in which less is known about the sampling distributions.

The following algorithm summarizes the calculation of the coverage parameter γ and bands that would be used to test for normality in a Q–Q plot. The specific null hypothesis is $H_0 : F_0 = N(\mu_0, \sigma_0^2)$, with both μ_0 and σ_0^2 known.

1. For $m = 1, \ldots, M$:
   (a) Simulate $X^m_1, \ldots, X^m_n \sim N(\mu_0, \sigma_0^2)$.
   (b) Standardize the sample, computing $Z^m_i = (X^m_i - \mu_0)/\sigma_0$.
   (c) Apply the inverse probability transformation, with $Y^m_i = \Phi^{-1}(Z^m_i)$.
   (d) Sort the data, obtaining $Y^m_{(1)} \le \cdots \le Y^m_{(n)}$.
   (e) Compute the tail probability $A^m_i = B^{-1}_{(i,n+1-i)}(Y^m_{(i)})$, $i = 1, \ldots, n$.
   (f) Find the minimum probability $C^m = 2\min_i\{\min(A^m_i, 1 - A^m_i)\}$.
2. Set γ to the $100\,\alpha$ percentile of $C^1, \ldots, C^M$.
3. Form the limits $[\Phi^{-1}(B^{-1}_{i,n+1-i}(\gamma/2)),\ \Phi^{-1}(B^{-1}_{i,n+1-i}(1 - \gamma/2))]$ for $Z_{(i)}$.

For all of our examples, we use $M = 5000$. Notice that when F_0 is fully specified, we can omit Steps (a)–(c) and begin with a sample from a uniform distribution, $Y_1, \ldots, Y_n \sim U[0, 1]$. We include these steps here for convenience when describing the algorithm used when parameters are estimated.

An *R* function that creates the TS confidence bands is available online at *http://www-stat.wharton.upenn.edu/sivana/QConBands.r http://www-stat.wharton.upenn.edu/~sivana/QConBands.r*. It takes about 10 s to produce results based on $M = 5000$ simulations with $n = 100$ and $\alpha = 0.05$ on a common laptop. The code uses the built-in sorting, sampling, and distribution functions in R. Next, we investigate the shape of the resulting TS confidence bands and compare them with the KS confidence bands.

2.2 Comparison of the TS and KS Confidence Bands

It is useful to contrast the TS bands to the KS bands for a uniform distribution. Figure 4 shows the KS and TS simultaneous 95% confidence bands for $n = 50$. The 95% KS bands for sampling a uniform distribution are two parallel lines around the diagonal. The KS bands need to be truncated to lie between 0 and 1 (since the standard uniform distribution cannot exceed these values). This truncation to 0 and 1 on the uniform scale produces the wide, diverging bands seen in the Q–Q plots of data shown in Figures 1–3.

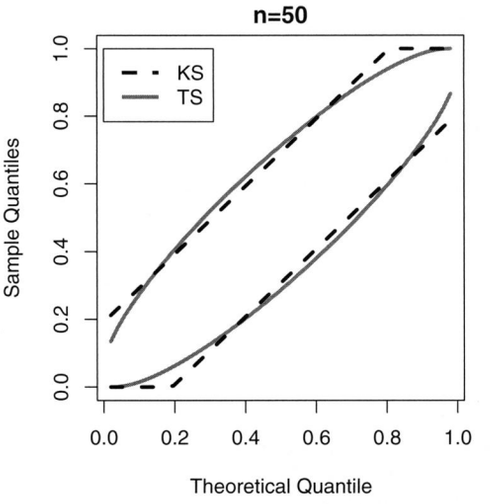

Figure 4. The 95% TS confidence bands versus the corresponding KS confidence bands (dashed).

In comparison, the TS bands in Figure 4 never reach beyond the boundaries. Instead, the TS bands are football-shaped (narrower at the extremes). This is to be expected because the variance of the extreme order statistics is much smaller than those at the center of the distribution. For example, suppose that n is odd and $m = (n+1)/2$. Then, $Y_{(1)}$ and $Y_{(n)}$ have a variance of $\frac{n}{(1+n)^2(2+n)} = O(1/n^2)$ whereas the median $Y_{(m)}$ has higher variance $\frac{n-1}{4n^2} = O(1/n)$. Also, the distributions of $Y_{(1)}$ and $Y_{(n)}$ are skewed to the right and left, respectively, while that of $Y_{(m)}$ is a symmetric unimodal distribution. As a result, the KS bands generally produce a less powerful test compared to the TS bands. The differences in form and performance between the TS and KS bands, and corresponding tests, becomes much clearer when we discuss their use for testing normality in the following section.

The differences between the KS and TS bands are also evident when we contrast them graphically for samples from a normal distribution. Figure 5 shows the 95% KS and TS confidence bands for $n = 50,\ 100,\ 1000$. Compared to the KS bands, the proposed TS bands are considerably tighter at the tails of the distribution. The TS bands are not, however, shorter everywhere. Figure 6 zooms in on the central portion of the normal distribution between $[-1, 1]$. By tightening the bands at the extremes, we increase their width slightly in the center of the distribution.

To further convey the difference between the two procedures, we look at the locations where the tests falsely reject the null hypothesis. By locations, we refer to the index i of the quantiles $Y_{(i)}$ and the frequency in which they lie outside the confidence bands. To examine this, we simulated $N = 50{,}000$ random samples from the standard normal distribution and recorded whether either of the tests falsely rejects the null hypothesis. If a sample is rejected by one of these tests, then the quantile positions where the sample exceeds the bands are recorded.

Figure 7 shows the histogram of the quantile indices where the test rejects for the KS and TS procedures and sample sizes $n = 50$ and $n = 100$. The histograms of the rejection locations for the KS test reveal a unimodal symmetric shape whereas those corresponding to the TS test resemble the uniform distribution.

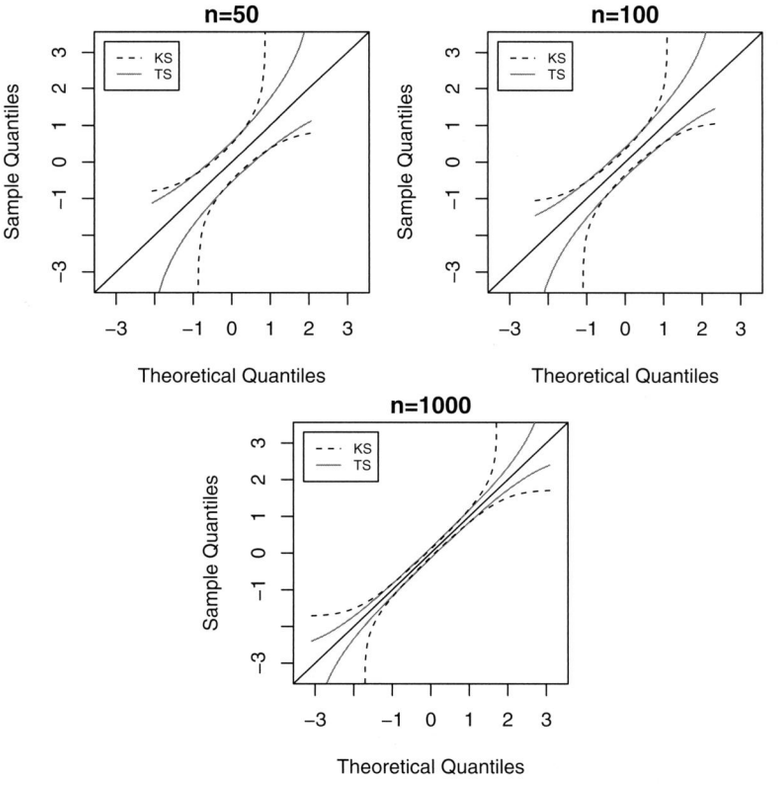

Figure 5. The 95% TS confidence bands versus the corresponding KS confidence bands (dash).

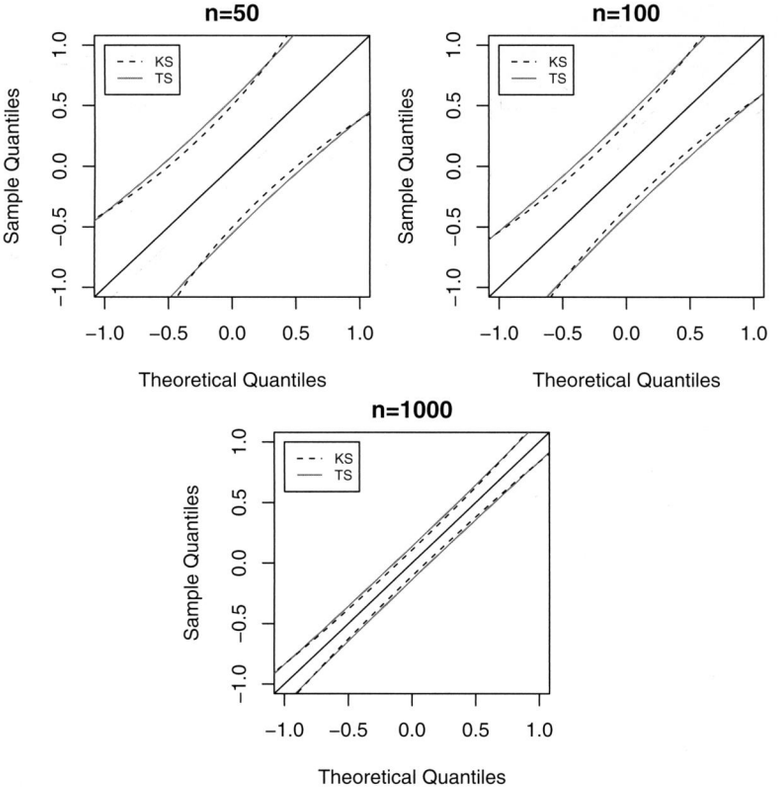

Figure 6. Zoomed-in plots of the 95% TS confidence bands versus the corresponding KS confidence bands (dash). These plots focus on the center of the domain and show the two tests are nearly identical over that range.

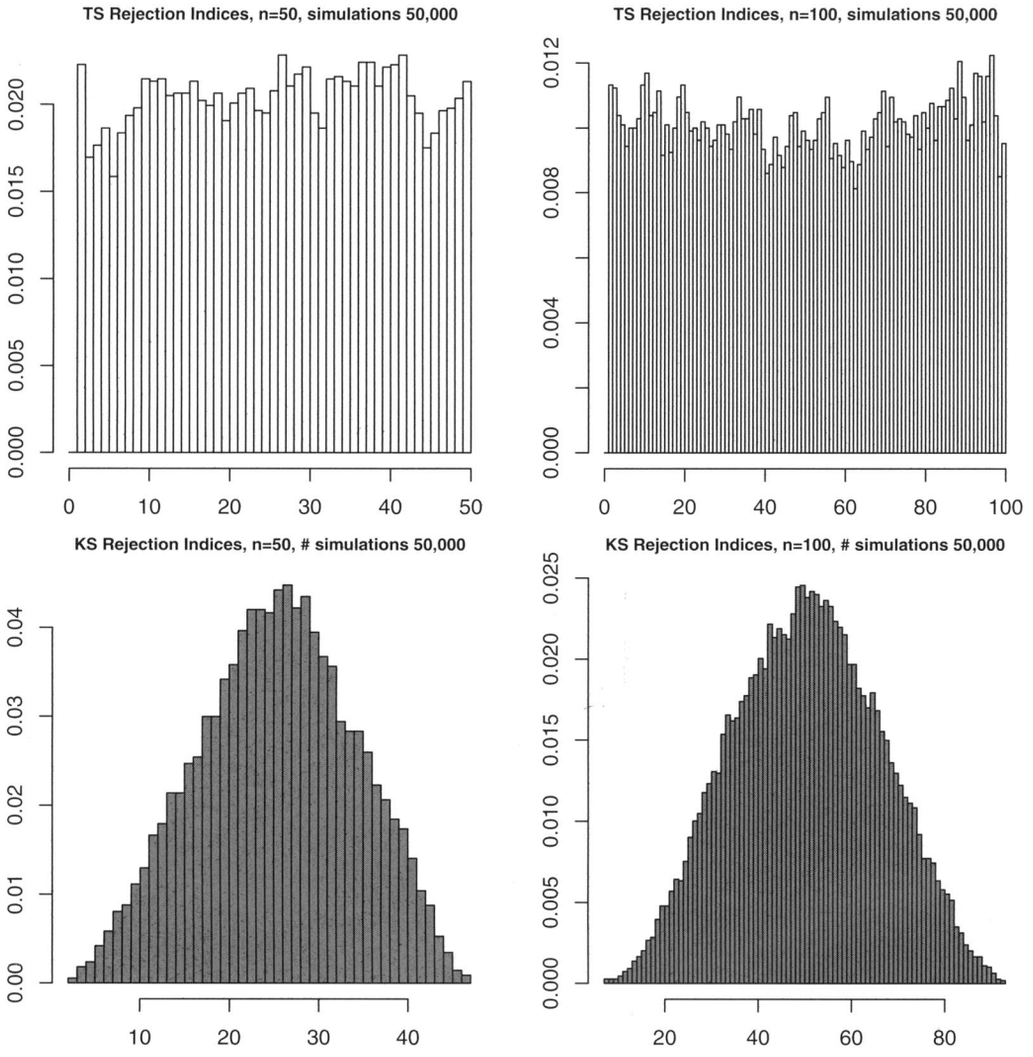

Figure 7. The histograms for the locations where the KS and TS tests falsely reject the null hypothesis (left, $n = 50$; right, $n = 100$).

These imply that the KS test is more likely to reject based on deviations in the center of the null distribution than deviations in the tails; the TS test rejects whether the deviations are at the tails or center of the null distribution.

The results in Figure 7 also suggest why the TS procedure performs better than the KS test against common nonnormal alternatives. Typically, when suitably scaled and centered, such alternatives have nearly normal behavior near their center but deviate from normality in the tails. Figure 7 suggests that TS is more sensitive in the tails of the distribution than the KS test. We especially see the difference between the two procedures when the alternatives are symmetric but heavier tailed compared to a normal distribution. In Section 3, we conduct a simulation study to investigate the power of these tests and show that the KS test is less powerful in detecting symmetric heavier tailed alternatives.

2.3 Testing Distributions With Unknown Parameters

The algorithm described in Section 2.1 presumes the null distribution is fully specified. In many applications, however, the researcher only knows the family of the underlying distribution, but not its identifying parameters. Simultaneous confidence bands must capture the effects of estimating these parameters. Our procedure can be modified to handle estimated parameters. We will demonstrate the needed adjustment for the normal distribution, but as we previously mentioned, this procedure can be easily modified to handle other families of continuous distributions.

To test whether $X_1, \ldots, X_n$, are normally distributed, we first estimate the population mean and standard deviation using the pair of estimators $(\breve{\mu}, \breve{\sigma})$, respectively. We discuss desirable choices for $\breve{\mu}$ and $\breve{\sigma}$ in Section 3.2. It is generally advisable to avoid the MLEs. We proceed by normalizing the sample by letting $Z_i = \frac{X_i - \breve{\mu}}{\breve{\sigma}}$. Then, we create the relevant confidence bands using a modified version of the procedure previously described and display these in the Q–Q plot of the normalized sample Z_i. We reject the null and conclude that the observed values are not normally distributed if any of the Z_i's lie outside the TS confidence bands. The steps required to construct the confidence bands are similar to those described in Section 2.1.

The only changes are the following modifications to Steps 1(a) and 1(b):

1(a) Simulate $X_1^m, \ldots, X_n^m \sim N(0, 1)$.

1(b) Standardize the sample using estimates $\breve{\mu}^m$ and $\breve{\sigma}^m$; that is, set $Z_i^m = \frac{X_i^m - \breve{\mu}^m}{\breve{\sigma}^m}$.

For example, one could use $\breve{\mu}^m = \bar{X}^m$ and $\breve{\sigma}^m = \sqrt{\frac{\sum_{i=1}^n (X_i^m - \bar{X}^m)^2}{n}}$. As shown in the following section, other choices generally provide more power.

3. POWER ANALYSIS

We use simulations to investigate the behavior of our testing procedure. More specifically, we examine the power of our procedure by calculating the percentage of times our testing procedure rejects the normal null distribution given that the simulated data are sampled from one of several alternative distributions. We study the power under two scenarios: (i) the mean μ and the standard deviation σ are prespecified and known. (ii) the mean and the standard deviation are unknown. In the first scenario, we employ the confidence bands described in Section 2.1, and for the second we use those from Section 2.3 to construct the appropriate confidence bands.

We set the significance level to $\alpha = 0.05$ and $n = 100$. We have conducted similar studies with sample sizes ranging from $n = 20$ to $n = 1000$, and the general pattern of results holds throughout this range. For alternative distributions, we use several that were previously studied in similar power studies in Wilk and Gnanadesikan (1968) and Rogers and Tukey (1972). Table 1 lists the alternatives, most of which are either skewed or heavy tailed.

3.1 Known Parameters

In this section, we use the theoretical mean and standard deviations of each of the alternative distributions listed in Table 1 to normalize the sample. We compare the performance of the TS procedure to the Kolmogorov–Smirnov (KS) and the Anderson–Darling (AD) tests, both of which originally required the mean and standard deviation to be known. All of these tests approximately achieve the desired size.

Table 1. Power analysis at level $\alpha = 0.05$ for $n = 100$

Alternative Distribution	TS test	KS test	AD test
Log Normal	1.000	1.000	1.000
$\chi^2(1)$	1.000	1.000	1.000
$\chi^2(5)$	**0.953**	0.346	0.496
$\chi^2(100)$	**0.089**	0.056	0.065
T(4)	**0.498**	0.136	0.187
Logistic	**0.186**	0.053	0.046
Poisson($\lambda = 15$)	0.192	**0.214**	0.076
Uniform(1,18)	**0.813**	0.614	0.727
Laplace($\mu = 0, b = 50$)	**0.486**	0.244	0.228
Norm Mix1 $\mu_1 = \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 10, p = 0.001$	**0.112**	0.042	0.079
Norm Mix2 $\mu_1 = \mu_2 = 0, \sigma_1 = 2, \sigma_2 = 1, p = 0.1$	**1.000**	0.758	0.996
Slash $\sigma = 1, a = 0.5, b = 0.9$	**0.1**	0.047	0.047
Wild $a = 12, p = 0.1$	**0.095**	0.059	0.054

Note: Boldface numbers in each row highlight the largest power against the given alternative distribution. Simulation std. errors range between 0.0000 and 0.0168.

Table 1 and Figure 8 summarize the power analysis. The TS test generally outperforms both the KS and AD tests. The advantage is most apparent for heavy-tailed distributions such as $\chi^2(30)$, $t(2)$, Laplace, and the first Normal mixture. We also see an advantage using our test when the distributions are skewed such as $\chi^2(5)$. Interestingly enough, all three tests have a hard time distinguishing between the normal distribution and the Wild,[1] Slash,[2] and one of the Normal mixtures[3] distributions that are studied. Morgenthaler and Tukey (1991) described these three distributions as the corner distributions and used them to model extreme behavior in data. These distributions are symmetric but heavier-tailed compared to a normal distribution. They are not as heavy tailed as a Cauchy distribution, however, and as such they are harder to distinguish from the normal distribution. Our test (and AD) can easily distinguish normal mixtures when the mixture probability is 10%, but not very well when that probability is low. We experimented with different values for p and the standard deviations σ_1 and σ_2; the power goes down when the mixture probability p goes down and is not sensitive to small changes in σ_1 and σ_2 (for a fixed p).

3.2 Unknown Parameters

Most real applications require adjustments for estimation. The key issue is to chose wisely the parameter estimates that provide the most power to detect deviations from normality. An obvious choice is to use the maximum-likelihood-based estimators under H_0:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n},$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}.$$

Although this choice of estimators seems reasonable under H_0, it does not guarantee the most powerful test against alternatives, particularly those with fat tails. Therefore, we also explore more robust estimators for the location and scale that allow us not only to maintain the appropriate significance level but also to obtain more power in detecting heavier tailed distributions.

The literature contains a variety of suggestions. A more robust pair of estimators combines the median $m(X)$ with the median absolute deviation (MAD)

$$\text{mad}(X) = m(|X - m(X)|) \cdot 1.4826$$

Lloyd (1952) suggested using the pair $\bar{X}$ with the scale estimate $\tilde{\sigma}$ from the Shapiro–Wilk test (2). Croux and Rousseeuw (1993) proposed an alternative robust estimator for the scale. Their estimator, denoted by Q_n, is more efficient than the MAD and

[1] $f(x) = \frac{x}{(b-a)\cdot\sqrt{2\cdot\pi}} \cdot (e^{\frac{x\cdot a^2}{2}} - e^{\frac{x\cdot b^2}{2}})$.

[2] $f(x) = (1-p)\cdot\phi(x) + p\cdot\frac{1}{2\cdot\sqrt{a}}\cdot\mathbf{1}_{x\in[-\sqrt{a},\sqrt{a}]}$.

[3] $f(x) = (1-p)\cdot\frac{1}{\sqrt{2\pi\sigma_1^2}}\cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + p\cdot\frac{1}{\sqrt{2\pi\sigma_2^2}}\cdot e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$.

Table 2. Power analysis with unknown parameters

Alternative Distribution	MLE	MM	GLS	Q_n	SW	LI	CVM	AAD
Log Normal	1.000	1.000	1.000	1.000	1.000	1.000	0.908	0.959
$\chi^2(1)$	1.000	1.000	1.000	1.000	1.000	1.000	0.970	0.997
$\chi^2(5)$	0.991	0.917	0.987	0.994	**0.996**	0.895	0.965	0.983
$\chi^2(100)$	0.136	0.085	0.132	**0.261**	0.146	0.105	0.106	0.110
T(4)	0.701	0.701	0.694	**0.846**	0.712	0.490	0.608	0.643
Logistic	0.345	0.302	0.331	**0.528**	0.317	0.149	0.217	0.253
Poisson($\lambda = 15$)	0.340	0.300	0.315	**0.616**	0.279	0.626	0.369	0.363
Uniform(1,18)	0.869	0.853	0.811	0.971	**0.991**	0.587	0.826	0.936
Laplace ($\mu = 0, b = 50$)	0.745	0.848	0.728	**0.950**	0.811	0.712	0.835	0.839
Norm Mix 1	0.112	0.091	0.106	**0.253**	0.099	0.078	0.079	0.082
Norm Mix 2	0.056	0.059	0.044	**0.188**	0.051	0.047	0.040	0.038
Slash	0.138	0.111	0.118	**0.293**	0.115	0.080	0.085	0.088
Wild $a = 12, p = 0.1$	0.115	0.132	0.100	**0.309**	0.082	0.071	0.081	0.090

Note: The boldfaced number in each row indicates the highest power against the alternative distribution. Simulation standard errors are less than 0.0162.

does not require an underlying assumption of symmetry. Q_n is defined as

$$Q_n = 2.219144 \cdot \{|X_i - X_j|; i < j\}_{(0.25)},$$

where $\{\cdot\}_{(0.25)}$ denotes the 0.25 quantile of the pairwise distances $\{|X_i - X_j|;\ i < j\}$. We pair Q_n with the median as recommended by these authors.

To show the influence of these choices on the power of tests for normality, we compare the power of the TS test under four alternatives:

1. Using the biased-adjusted maximum likelihood estimators.
2. Using the median and MAD (MM).
3. Using the generalized least squares estimators suggested by Lloyd (GLS).
4. Using Q_n and the median as suggested by Rousseeuw and Croux (Q_n).

We compare the performance of the TS test using the aforementioned options with the following more common testing procedures:

5. Shapiro–Wilk (SW).
6. Lilliefors test (LI).
7. Cramèr–von Mises (CVM).
8. Anderson–Darling test adjusted for unknown parameters (AAD).

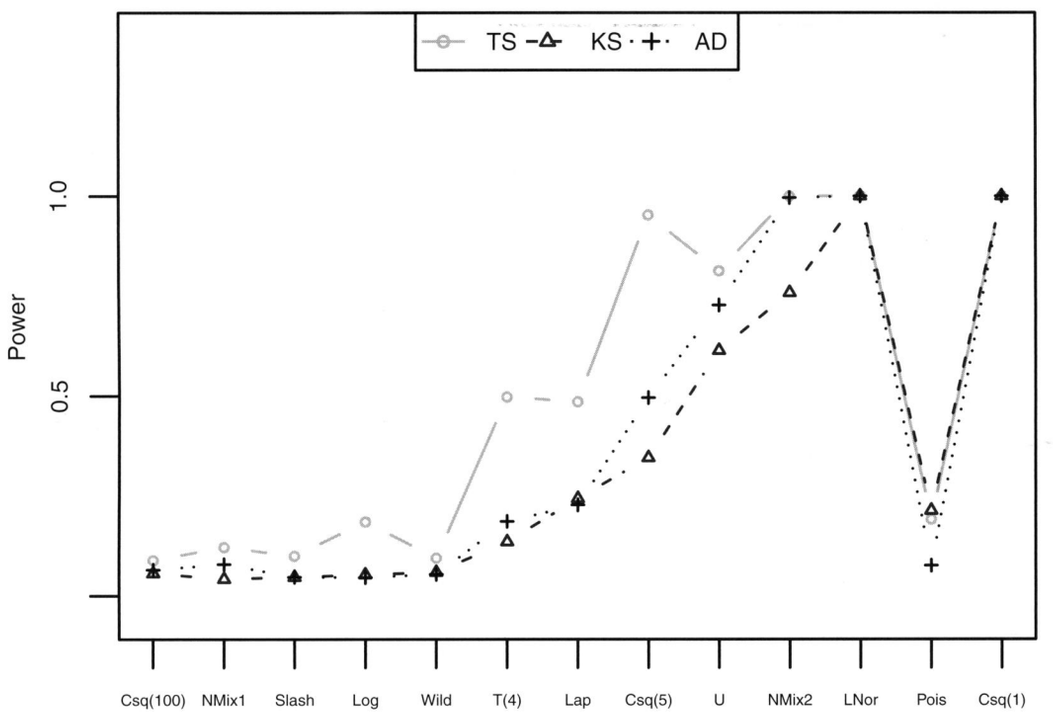

Figure 8. The power for each test procedure and each alternative distribution. The distributions are ordered by the Kullback–Leibler divergence between the alternative and the appropriately scaled and centered normal distribution, $\mathrm{KL}(F_1, N(\mu_1, \sigma_1^2))$ where $\mu_1 = E_{F_1}(X)$ and $\sigma_1^2 = E_{F_1}(X - \mu_1)^2$.

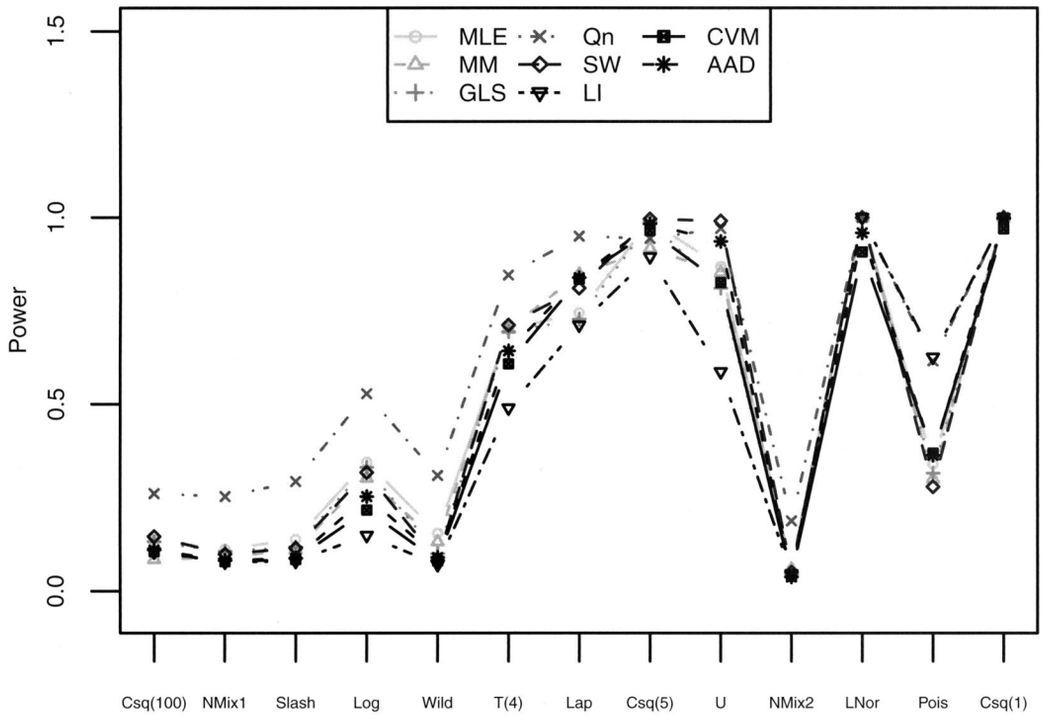

Figure 9. The power for each of the test procedures at each alternative distribution. The distributions are ordered by the Kullback–Leibler between the alternative distribution and the appropriately scaled and centered normal distribution, that is, $\mathrm{KL}(F_1, N(\mu_1, \sigma_1^2))$ where $\mu_1 = E_{F_1}(X)$ and $\sigma_1^2 = E_{F_1}(X - \mu_1)^2$.

The outcomes of the power analysis are listed in Table 2 and shown in Figure 9. These results indicate that for the TS test, one should use the median and Q_n to estimate the location and scale parameters. In general, the TS test with this choice of estimators performs at least as well as the Shapiro–Wilk test (and often much better). It is interesting to notice that the TS test performs relatively well in Tukey's three corner distributions. Its power is higher for these distributions when using estimated rather than known parameters, similar to the effects of adjusting the KS test for estimated parameters. One explanation for the increase in power is that the standard deviation is not the appropriate scaling factor for these distributions. The standard deviation in these situations is too sensitive to the heavy tails of the distributions.

4. DISCUSSION

The TS procedure introduces an attractive alternative to the commonly used KS test. It combines a visual presentation in the classical normal Q–Q plot with higher power against common fat-tailed distributions. Most testing procedures can distinguish well between the normal distribution and nonsymmetric or symmetric very heavy-tailed distributions. These tests underperform, however, when asked to distinguish a normal distribution from a mildly heavy-tailed symmetric distribution. The TS procedure has more power than the commonly used tests for such alternatives.

Whether or not we know the parameters of the normal distribution, the TS test requires a separate calculation for each pair of significance level α and sample size n. A natural question is whether for a given significance level α there exists a closed-form equation of the form $\frac{C_\alpha}{\sqrt{n}}$ as underlies KS confidence bands. This simple asymptotic behavior is part of its appeal. After careful consideration that is detailed in the supplementary material (available online), we conclude that our procedure does not have a limiting behavior similar to the KS test. Instead, our procedure's margin of error grows at a rate of $O(\frac{\log(\log(n))}{\sqrt{n}})$ as n increases for a given significance level α. This rate, of course, is very slow and almost behaves like a constant for large values of n.

Finally, the proposed TS testing procedure is designed to handle independent identically distributed samples. There are applications, however, that require a relaxation of these assumptions. One such example is the linear regression where the quantile–quantile plot is often used to determine whether the sample residuals follow a normal distribution. Since the sample residuals in an ordinary least squares regression are neither independent nor homoscedastic, our procedure does not strictly apply. One can use the studentized residuals to adjust for hetroscedasticity, but the TS procedure will still need to be modified to account for the dependence between these residuals. We leave this modification to be studied in future research.

SUPPLEMENTARY MATERIALS

The online supplementary material for this article discusses whether the tail sensitive bands described in this article have a simple limiting asymptotic form comparable to those available for the KS procedure. The argument shows that such a limiting form does not exist.

[Received February 2012. Revised September 2013.]

REFERENCES

Anderson, T. W., and Darling, D. A. (1954), "A Test of Goodness of Fit," *Journal of the American Statistical Association*, 49, 765–769. [252]

Atkinson, A. C. (1981), "Two Graphical Displays for Outlying and Influential Observations in Regression," *Biometrika*, 68, 13–20. [253]

Beran, R. (1987), "Prepivoting to Reduce Level Error of Confidence Sets," *Biometrika*, 74, 457–468. [253]

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco, CA: Holden-Day. [250]

Birnbaum, Z. W. (1952), "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size," *Journal of the American Statistical Association*, 47, 425–441. [252]

Buja, A., and Rolke, W. (2006), "Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference With Applications to Function Estimation and Functional Data," unpublished, available at *http://stat.wharton.upenn.edu/~buja/PAPERS/paper-sim.pdf*. [253,254]

Croux, C., and Rousseeuw, P. J. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283. [257]

Darling, D. A. (1957), "The Kolmogorov–Smirnov, Cramèr–Von Mises Tests," *The Annals of Mathematical Statistics*, 28, 823–838. [252]

DasGupta, A. (2011), *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, Berlin: Springer. [252]

Faraway, J. J., and Csorgo, S. (1996), "The Exact and Asymptotic Distributions of Cramer–Von Mises Statistics," *Journal of the Royal Statistical Society*, Series B, 58, 221–234. [252]

Kolmogorov, A. (1933), "Sulla Determinazione Empirica di Una Legge di Distributione," *Giornale dell' Istituto Italiano degli Attuari*, 4, 83–91. [252]

Lilliefors, H. W. (1967), "On the Kolmogorov–Smirnov Test for Normality With Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399–402. [252]

Lloyd, E. H. (1952), "Least-Squares Estimation of Location and Scale Parameters Using Order Statistics," *Biometrika*, 39, 88–95. [257]

Massey, F. J. (1951), "The Kolmogorov–Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, 46, 68–78. [252]

Morgenthaler, S., and Tukey, J. W. (1991), *Configural Polysampling: A Route to Practical Robustness*, New York: Wiley-Interscience. [257]

Razali, N. M., and Wah, Y. B. (2011), "Power Comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling Tests," *Journal of Statistical Modeling and Analytics*, 2, 21–33. [253]

Rogers, W. H., and Tukey, J. W. (1972), "Understanding Some Long-Tailed Symmetrical Distributions," *Statistica Neerlandica*, 26, 211–226. [257]

Rosenkrantz, W. (2000), "Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit," *The American Statistician*, 54, 185–190. [250]

Shapiro, S. S., and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611. [252]

Smirnov, N. (1939), "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples," *Bulletin de l'Université de Moscou, Série de internationale (Mathématiques)*, 2, 3–16. [252]

——— (1948), "Table for Estimating the Goodness of Fit of Empirical Distributions," *The Annals of Mathematical Statistics*, 19, 279–281. [252]

Stephens, M. A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730–737. [252,253]

Testing of Body Armor Materials for Use by the U.S. Army—Phase III (2012), "Board on Army Science and Technology, National Research Council," *National Academies Press*. The data cited in this article appear in Appendix M of this report. [249]

Tsay, R. S. (2010), *Analysis of Financial Time Series* (3rd ed.), New York: Wiley. [249]

Wilk, M. B., and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17. [253,257]