The Power to See: A New Graphical Test of Normality

Sivan Aldor-Noiman[*] Lawrence D. Brown[*]
Robert A. Stine [*]

Department of Statistics
The Wharton School
University of Pennsylvania
PA, USA

Abstract

Many statistical procedures assume the underlying data generating process involves Gaussian errors. Among the well-known procedures are ANOVA, multiple regression, linear discriminant analysis and many more. There are a few popular procedures that are commonly used to test for normality such as the Kolmogorov-Smirnov test and the Shapiro-Wilk test. Excluding the Kolmogorov-Smirnov testing procedure, these methods do not have a graphical representation. As such these testing methods offer very little insight as to how the observed process deviates from the normality assumption. In this paper we discuss a simple new graphical procedure which provides confidence bands for a normal quantile-quantile plot. These bands define a test of normality and are much narrower in the tails than those related to the Kolmogorov-Smirnov test. Correspondingly the new procedure has much greater power to detect deviations from normality in the tails.

Key words: normality test, confidence bands, graphical presentation, power analysis, quantile-quantile plot

[*]Research supported in part by NSF Grant DMS-1007657

1 Introduction

To motivate our procedure we first look at two different data sets that were previously explored. In both examples the researchers were interested in testing the normality assumption.

The first data series contains monthly log returns of IBM stock from March 1967 to December 2008. This is part of a data set that was examined in Tsay (2010) and we obtained it from its corresponding website. In Figure 1 we show the time plot for the data and the quantile-quantile plot of the data with both our proposed Tail-Sensitive (TS) 95% confidence bands and the Kolmogorov-Smirnov (KS) confidence bands. It is apparent that some of the points fall outside the TS confidence bands but inside the KS confidence bands. It is apparent that five points in the left tail fall outside the TS confidence bands but are well inside the KS confidence bands. Therefore, according to the TS confidence bands the data do not follow the normal distribution and the log returns for this stock have a 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 second data series contains measurements from experiments that test the effectiveness of body armors. The data was collected as part of a National Academies report requested by the US Army Unknown (2012). The Army wanted to investigate the difference between two methods of assessing how deep a bullet penetrates ceramic body armor begin tested for approval for use. 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

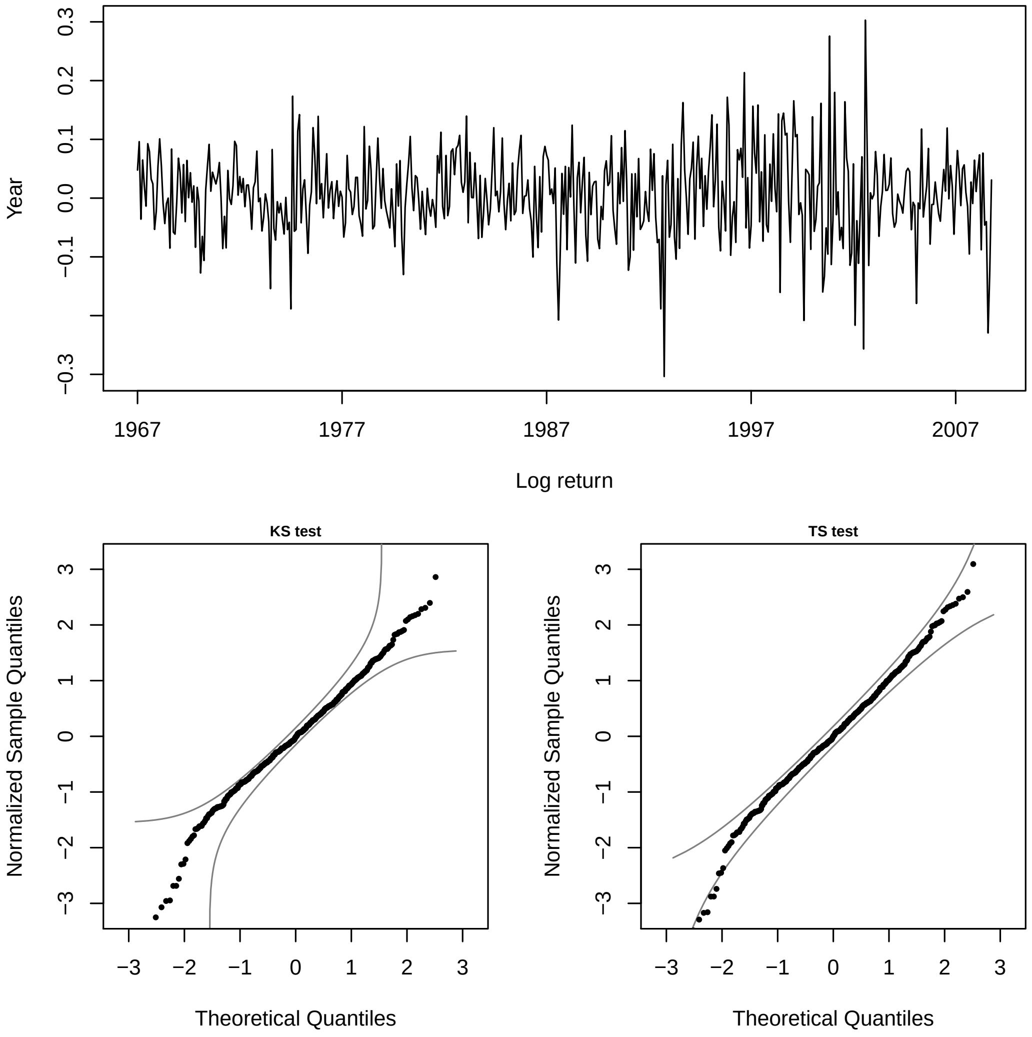

Figure 1: The monthly log returns for IBM stock from March 1967 to December 2008. The upper plot shows the time plot of the data against the time index. The lower plots shows the 95% Kolmogorov-Smirnov confidence bands and the corresponding TS confidence bands, respectively.

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 involves using a manually controlled digital caliper. A more recently-adopted method is to measure 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 2 shows the quantile-quantile 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 proposed Tail-Sensitive (TS) 95% confidence bands and the Kolmogorov-Smirnov (KS) confidence bands. Based on the KS bands, observations from both methods are consistent with the normality assumption. However, based on the TS confidence bands there is a suspicious outlier on the right tail of the caliper-based measurements. On the laser-based measurements we see two points on the right tail that fall outside the bands and several suspicious data points on the left tail that lie on the boarder of the bands. These points indicate that the data deviate from the normality assumption. Our confidence bands procedure indicates that if the Army adopts the laser based method it should not rely on the normality assumption to establish its safety standards.

In the next section we list a few common procedures that are used to test the normality assumption. We later compare these testing procedures performances with our TS procedure.

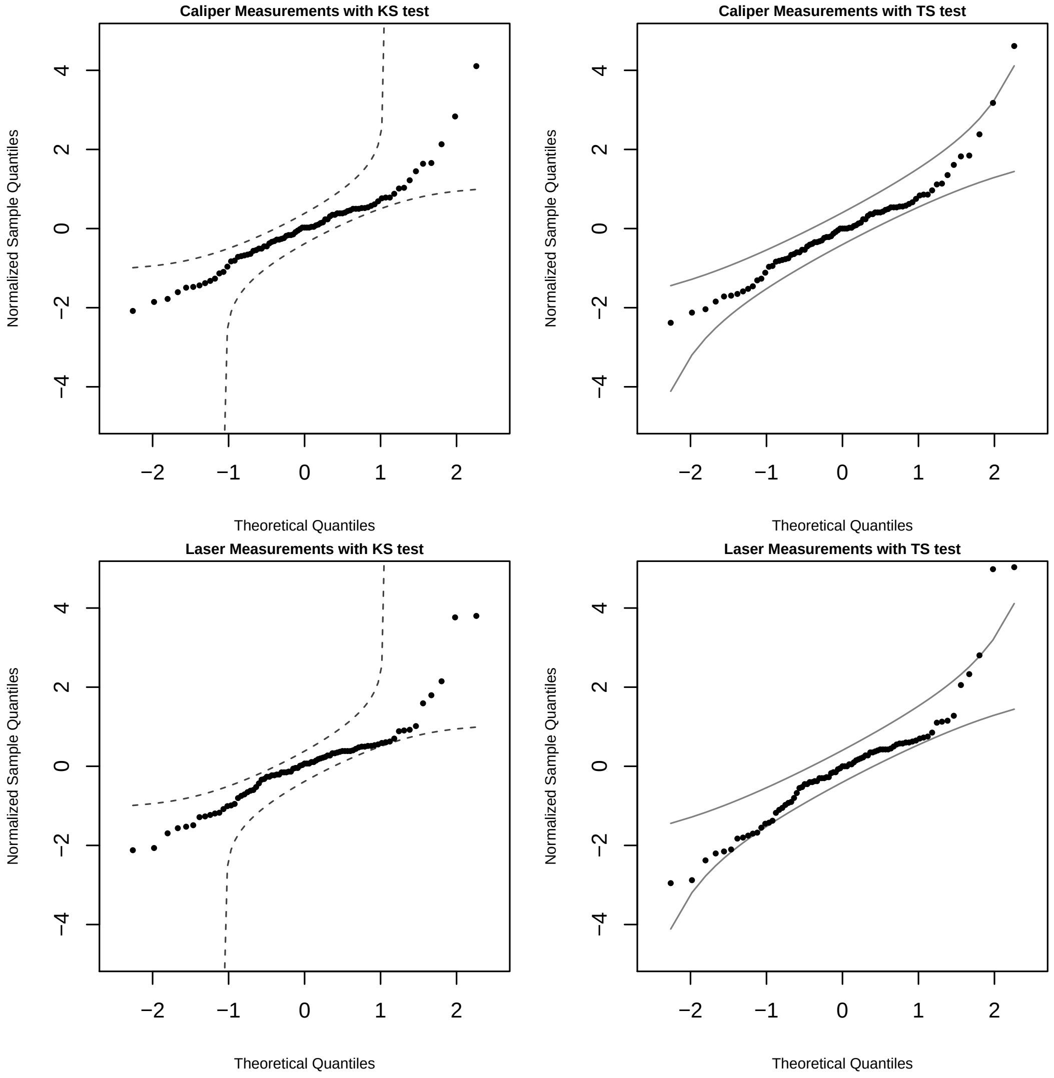

Figure 2: Measurements of bullet impressions on a ceramic armor. The upper plots shows the quantile-quantile plot of measurements taken using the digital caliper. The lower plots shows the quantile-quantile plot of measurements taken using the laser-based device. Both plots show the the proposed 95% TS confidence bands and the corresponding Kolmogorov-Smirnov confidence bands.

1.1 Common testing procedures

Several statistics have been proposed to test the assumption of normality with fixed mean, μ, and variance, σ^2. The hypotheses in question can be written as follows:

$$\begin{aligned} H_0 : X_i &\overset{\text{iid}}{\sim} N(\mu, \sigma^2) \text{ for } i = 1, \ldots, n \\ H_1 : X_i &\overset{\text{iid}}{\sim} F \text{ for } i = 1, \ldots, n \end{aligned} \tag{1}$$

where F is a general symbol for any arbitrary continuous CDF different from those in H_0. Until Section 2.3 we 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 not assumed known, and must be estimated from the data. For the case in which 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. Common testing procedures for the case with μ, σ^2 known rely on some function of the deviation between the sample cumulative distribution function (CDF), $F_n(\cdot)$, and the normal null cumulative distribution $F_0 = \Phi(\cdot)$. We proceed by reviewing a few of the more common testing procedures.

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

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

Since the CDF is a continuous function we can rewrite ω_n as follows:

$$\begin{aligned}\omega_n &= n\int_{-\infty}^{\infty}(\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{[t>X_j]} - F_0(t))^2 dF_0(t)\\ &= n\int_{0}^{1}(\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{[t>F_0(X_j)]} - t)^2 dt\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function. Faraway and Csorgo (1996) studied the asymptotic distribution of ω_w.

In the late 40s, the popular Kolmogorov-Smirnov test was developed (Feller, 1948). The test statistic, denoted by B_n, is the maximum difference between the empirical cumulative distribution function and the hypothesized (normal) cumulative distribution function. Formally, the statistic can be written as:

$$B_n = \sqrt{n} \sup_{-\infty<t<\infty} |F_n(t) - F_0(t)| \tag{3}$$

The distribution of B_n was described and tabulated in Kolmogoroff (1941) and Smirnov (1948). This testing procedure also has a visual representation using the corresponding confidence bands. This probably contributes to its popularity among practitioners who use commercial software such as SAS, STATA and JMP. An example of the visual representation is shown in Figure 1. The KS bands are initially constructed on the uniform scale using the tables from Smirnov (1948) and then translated to the desired null distribution using its CDF. For further details on how these bands are constructed the reader is referred to (DasGupta, 2011). In Lilliefors (1967), the author investigated how to adjust the critical values of B_n when the null hypothesis is the normal distribution with unknown mean and standard deviation.

A few years later, Anderson and Darling (1954) suggested the following test statistic

$$A_n = n \int_{-\infty}^{\infty} \frac{(F_n(t) - F_0(t))^2}{F_0(t) \cdot (1 - F_0(t))} dF_0(t). \tag{4}$$

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). Similar to the Kolmogorov-Smirnov test, the Anderson-Darling statistic, A_n, behavior was also examined for the case where the parameters are unknown and tables to compute the adjusted p-values were reported in Stephens (1974). We can view the Anderson-Darling statistic as a weighted version of the Cramèr-Von-Mises where the weight function is $[F_0(t) \cdot (1 - F_0(t))]^{-1}$. By using this weight function Anderson and Darling place more emphasis on the deviation at the tails of the distribution.

The A_n, ω_n and B_n tests can be used for any specified null distribution, not just the normal one. 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 following form

$$\begin{aligned} W_n &= \frac{b^2}{S^2} = \frac{(\hat{\sigma} \cdot a)^2}{S^2} \text{ where} \\ \hat{\sigma} &= \frac{\boldsymbol{m}^t \boldsymbol{V}^{-1} y}{\boldsymbol{m}^t \boldsymbol{V}^{-1} \boldsymbol{m}} \text{ and} \\ S^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} \end{aligned} \tag{5}$$

where $y = [y_{(1)}, \ldots, y_{(n)}]$ 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 [1]. b is, up to a normalizing constant, the estimated slope of the generalized linear 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.

Various studies, for example Stephens (1974), Razali and Wah (2011), show 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 has versus the other tests is its visual presentation. By visually inspecting the deviations on this plot, the researcher may be able to better understand and possibly correct the non-normality in the data by a simple transformation or might be able to assume some different underlying process.

In the next section we describe our new TS procedure. For most alternatives

[1] In later computations we use the approximation suggested in Wilk and Gnanadesikan (1968) to evaluate $\boldsymbol{m}$ and $\boldsymbol{V}$

this procedure is approximately as powerful as the Shapiro-Wilk test and can be depicted in a simple quantile-quantile plot like KS. We also 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 New testing method

In this section we describe in detail our graphical method and the corresponding test. We begin by constructing the $1-\alpha$ confidence bands for the normal quantiles. Once these bands are calculated all that remains is to examine if the quantile plot of the sample order statistics falls within these bands (or outside) to determine if one retains or rejects the normality assumption.

2.1 $1-\alpha$ Simultaneous confidence 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 observations to have mean 0 and standard deviation of 1 (i.e. use $z_i = \frac{x_i - \mu_0}{\sigma_0}$). We then construct the following TS confidence bands and confirm if the quantile-quantile plot of the normalized sample z_i, falls entirely inside the confidence bands.

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

1. Build individual $1-\gamma$ confidence intervals for each of the quantiles of the uniform distribution.

2. Adjust the confidence level, γ, to account for multiplicity in order to form $1-\alpha$ coverage simultaneous confidence bands. Establishing the exact γ is very difficult and therefore we use simulated data to determine it. In principle, this can be done by sequentially choosing values of γ until the desired accuracy has been achieved. With this scheme, for each choice of γ one needs to simulate from the uniform distribution in order to determine with desired accuracy the value of simultaneous coverage achieved when using that value of $1-\gamma$ as the individual coverage value. Then successive values of γ are chosen until the coverage bands have simultaneous coverage $1-\alpha$ to within the desired numerical accuracy. This scheme achieves simultaneous confidence bands that are accurate to within any desired statistical accuracy. But it is computationally inefficient because it requires repeated large simulations in order to reach the final value of γ. Therefore we use an alternative, more efficient, method to establish the desired γ. This alternative method is equivalent to the initial idea but circumvents the need for the incremental search. We will detail this method in Section 2.1.

Step 1 establishes the shape of the confidence bands while Step 2 ensures that the bands correspond to a size α test. As a final step, we apply the inverse normal CDF to obtain bounds for the normal QQ plot. Here are the details for the computational algorithm:

Step 1: Individual confidence intervals Assume $Y_1, \ldots, Y_n$ are n independent identically distributed standard uniform random variables. Sort them from the smallest to largest to obtain the order statistics, $Y_{(i)}$. From elementary probability theory, $Y_{(i)}$ follows a Beta distribution with shape parameters $\alpha = i$ and $\beta = n - i + 1$ which we denote by $G_{(i,n+1-i)}$.

This last fact allows us to construct a $1 - a$ level confidence interval for the i^{th} order statistic. Choose the Beta quantiles such that $P(L_i(\gamma) \leq Y_{(i)} \leq U_i(\gamma)) \geq 1 - a$. A simple choice is the equal-tail quantiles, i.e. $L_i(\gamma) = G^{-1}_{(i,n+1-i)}(\frac{\gamma}{2})$ and $U_i(\gamma) = G^{-1}_{(i,n+1-i)}(1 - \frac{\gamma}{2})$.

Step 2: Confidence bands The set of individual confidence intervals allows us to make inference on each order statistic, $Y_{(i)}$, separately. These confidence bounds, however, do not ensure a simultaneous $1 - \alpha$ confidence band. Therefore, we need to modify these individual confidence limits in order to achieve bands with an expected coverage of $1 - \alpha$. We propose to chose γ so that $P(L_i(\gamma) \leq Y_{(i)} \leq U_i(\gamma), \forall i) \approx 1 - \alpha$. To this end we simulate data sets from the uniform distribution and find the smallest two-sided p-value for each of the simulated data sets, i.e. $C^m = min_{1 \leq i \leq n} p_i^m$. We then proceed by finding the $\alpha \cdot 100\%$-percentile over $[C^1, C^2, \ldots, C^M]$ and adjust the confidence intervals in Step 1 according to this value. The procedure guarantees that only $\alpha\%$ of the data sets lie outside the confidence bands and therefore meets the requirement for an α size test. The following algorithm describes the method:

1. Simulate M samples each having n observations from the standard uniform distribution. In all our examples we use $M = 5,000$. Define $Y_{(i)}^m$ as the i^{th} order statistic in the m^{th} simulated sample for $i = 1, \ldots, n$ and $m = 1, \ldots, M$.

2. For each $i = 1, \ldots, n$ and $m = 1, \ldots, M$ calculate $A_i^m = G_{(i,n+1-i)}^{-1}(Y_{(i)}^m)$.

3. For each $i = 1, \ldots, n$ and $m = 1, \ldots, M$ calculate $p_i^m = 2 \cdot min(A_i^m, 1 - A_i^m)$ under the null Beta distribution. p_i^m indicates the significance level of $Y_{(i)}^m$ relative to the corresponding Beta distribution.

4. For each simulated sample find the smallest significance value associated with it, i.e. for $m = 1, \ldots, M$ find $C^m = min_{1 \leq i \leq n} p_i^m$.

5. Among $[C^1, C^2, \ldots, C^M]$, find the $\alpha \cdot 100\%$-percentile. Denote this value by C_α. Adjust the individual confidence interval constructed in Step 1 using $\gamma = C_\alpha$.

6. Finally, for testing normality, we use the inverse normal cumulative distribution function, $\Phi^{-1}(\cdot)$ to transform the simultaneous uniform confidence bands to the corresponding standard normal confidence bands.

It can easily be verified that this algorithm yields bounds that achieve $P(L_i(\gamma) \leq Y_{(i)} \leq U_i(\gamma), \forall i) \approx 1 - \alpha$ except for simulation error. Indeed, by construction, among the M simulated samples $[(1 - \alpha) \cdot M]$ have $L_i(\gamma) \leq Y_{(i)}^m \leq U_i(\gamma), \forall i$.

The R function that creates the TS confidence bands is available online at http://www-stat.wharton.upenn.edu/ sivana/QConBands.r. It takes about 2 seconds to produce results based on $M = 1000$ simulations for sample size

$n = 100$ and $\alpha = 0.05$. Throughout the code we make use of the quicksort algorithm and the standard sampling method in R to simulate from the Uniform distribution. One can think of better sampling and/or sorting algorithms to try and improve the computational efficiency of the proposed method. We leave these improvements for future research. Next we investigate the shape of the resulting TS confidence bands and compare them with the KS confidence bands.

2.1.1 The TS confidence bands

To illustrate the results of the TS procedure we first examine the confidence bands on the uniform scale (before the inverse normal transformation is applied). Figure 3 demonstrate the simultaneous 95% confidence bands for a sample size of $n = 100$. The bands are football shaped (narrower at the extremes) which is to be expected since we set quantiles from a Beta distribution. $Y_{(1)}$ and $Y_{(n)}$ have a variance of $\frac{n}{(1+n)^2 \cdot (2+n)}$ while the median, $Y_{(n/2)}$ has the higher variance of $\frac{1}{4 \cdot (1+n)}$. Also, the distributions of $Y_{(1)}$ and $Y_{(n)}$ are highly skewed to the right and left respectively while that of $Y_{(n/2)}$ is symmetric unimodal distribution. Therefore the resulting confidence bands are not symmetric. The plot also shows the 95% Kolmogorov-Smirnov confidence bands which form two parallel lines around the 45 degree line.

Figure 3 reveals that the Kolmogorov-Smirnov bands are especially wide at the tails of the distribution. In fact, the Kolmogorov-Smirnov confidence bands need to be truncated to be between the values of 0 and 1 (since the standard uniform distribution cannot exceed these values). The TS bands never reach beyond these boundaries. The consequence of this difference is

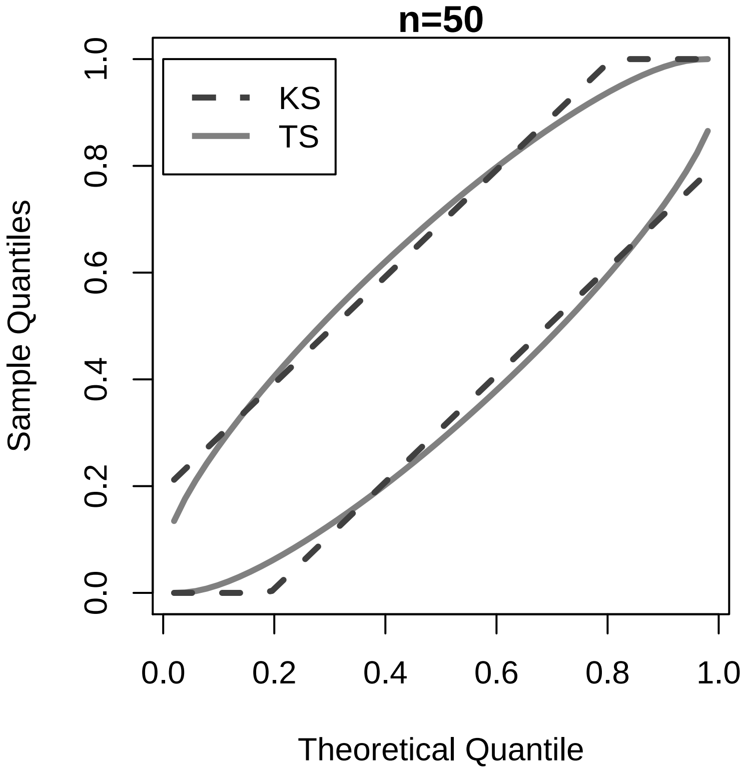

Figure 3: The 95% TS confidence bands versus the corresponding KS confidence bands (dash).

that the Kolmogorov-Smirnov bands generally produce a less powerful test compared to the TS bands. The difference 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.

2.2 Comparison of the normal TS and the KS confidence bands

The best way to clarify the strength and weakness of the KS and TS confidence bands is by looking at some plots. Figure 4 shows the 95% TS confidence bands versus the KS confidence bands for $n = 50, 100, 1000$. As this figure reveals, compared to the KS test, the suggested TS confidence bands are considerably tighter at the tails of the distribution. By contrast, Figure 5 zooms in on axes values in the central region between $[-1, 1]$. These two figures show that even though we have tightened the bands at the ends we do not sacrifice much of their width in the center of the distribution.

To further understand the difference between the two test procedures we look at the locations where the tests falsely reject the null hypothesis. By locations we are referring to the quantiles and the frequency in which they lie outside the confidence bands. To examine this we simulate $T = 50,000$ random samples from the standard normal distribution and record 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 6 shows the histogram of the locations where the test is rejected for each of the two testing procedures for sample sizes $n = 100$ and $n = 1000$. The histograms that correspond to the KS reveal a unimodal symmetric shape while the normal TS histograms resemble the uniform distribution. 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 while the TS test rejects whether the deviations are at the tails or center of the null distribution.

The results in Figure 6 also suggest why the TS procedure performs better than the KS test against common non-normal alternatives. Typically, when suitably scaled and centered, such alternatives have nearly normal behavior near their center but deviate from normality in the tails. Figure 6 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

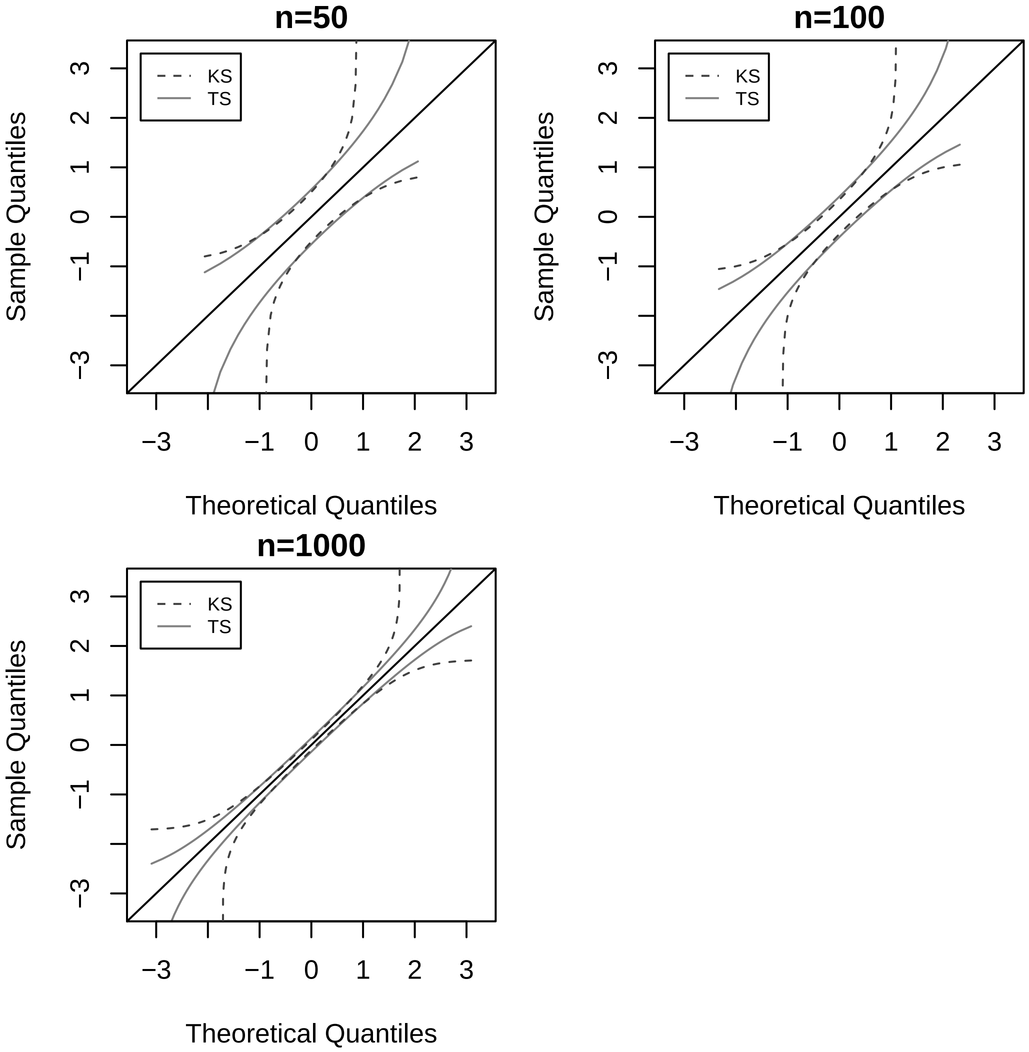

Figure 4: The 95% TS confidence bands versus the corresponding KS confidence bands (dash).

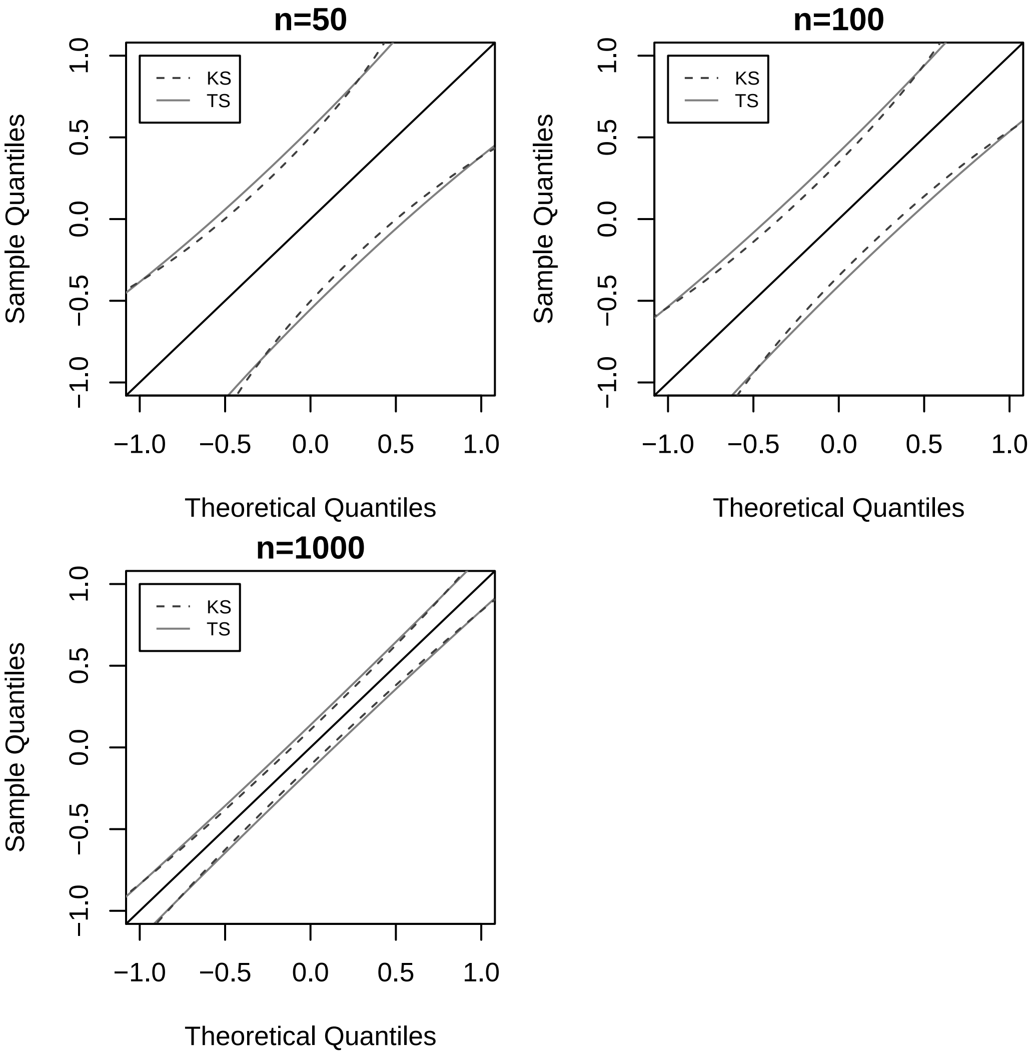

Figure 5: Zoomed-in plots of the 95% TS confidence bands versus the corresponding KS confidence bands (dash). These plots focus on the center of the domain and show the two tests are nearly identical over that range.

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 is only relevant when the null distribution is fully specified. However, in many applications the researcher only knows the underlying distribution's family but not its population parameters. This uncertainty in the parameters needs to be reflected in the confidence bands since not knowing the parameters adds another source of variability to the problem.

We will now demonstrate how our procedure can be modified to handle a situation when the parameters are not pre-specified. We will use the normal distribution as an example for our null distribution but as we previously mentioned, this procedure can be easily modified to handle other families of continuous distributions.

2.3.1 Confidence bands in the case of unknown parameters

To test whether n observations, $x_1, \ldots, x_n$, are normally distributed with unknown parameters we first estimate the population mean and standard deviation using the pair of estimators $(\tilde{\mu}, \tilde{\sigma})$, respectively. We discuss desirable choices for $(\tilde{\mu}$ and $\tilde{\sigma})$ in Section 4. We proceed by normalizing the sample by letting $z_i = \frac{x_i - \tilde{\mu}}{\tilde{\sigma}}$. Then we create the relevant confidence bands using a modified version of the procedure previously described and apply these to the quantile-quantile plot of the normalized sample z_i. Finally, if one or more of

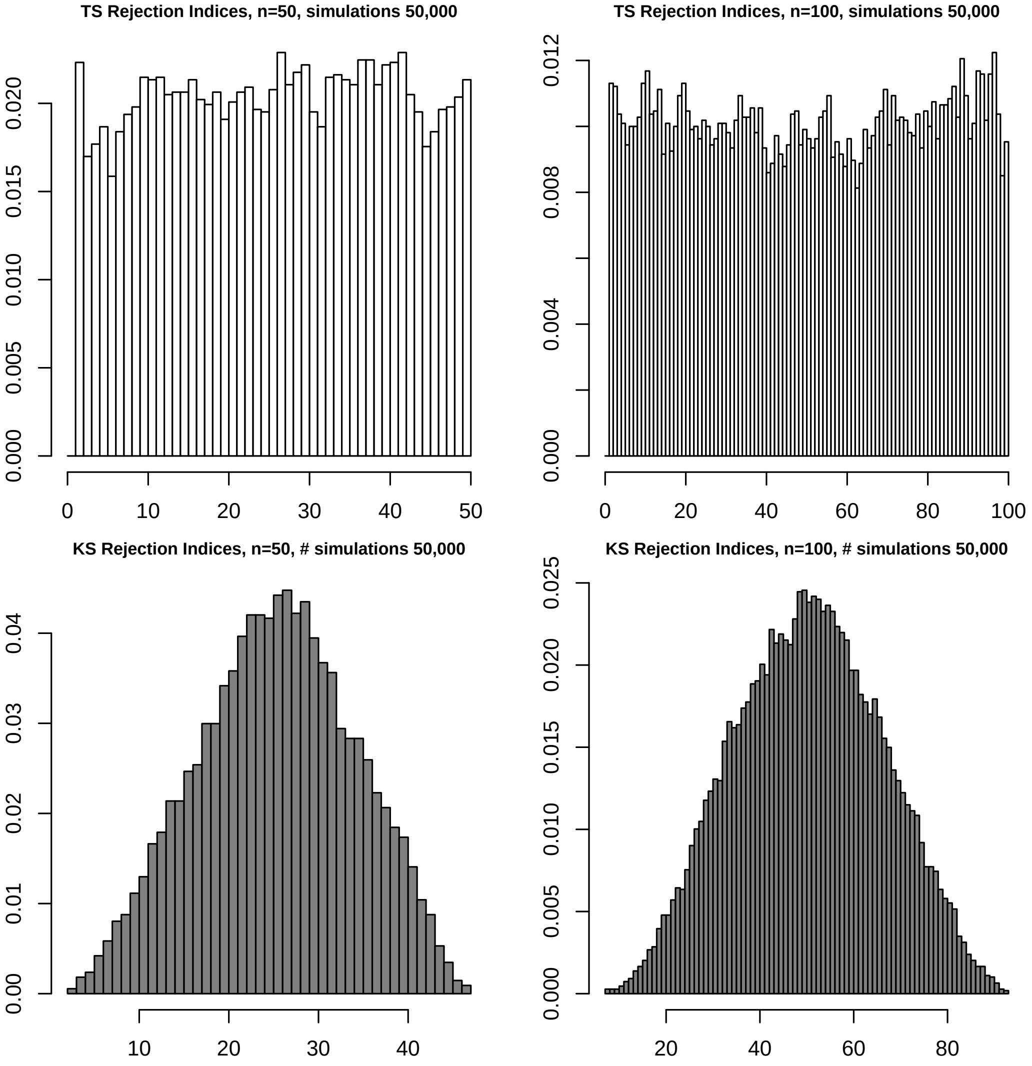

Figure 6: The histograms for the locations where the KS and TS tests falsely reject the null hypothesis. The left panel histograms (blue) correspond to a sample size of $n = 100$ and the right panel histograms (black) correspond to a sample size of $n = 1000$.

the z_i's lie outside the TS confidence bands we reject the null and conclude that the observed values are not normally distributed. The steps required to construct the confidence bands are similar to the ones described in Section 2.1 with the exception of the first step. We replace the first step with the following three:

1. Simulate M samples each having n observations from the standard normal distribution.

2. Normalize each of the samples using estimates for the mean, $\hat{\mu}$, and the standard deviation, $\hat{\sigma}$, i.e. $z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$. For example, we may use the maximum likelihood estimators $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n}}$. Let $Z_{(i)}^m$ be the i^{th} order statistic in the m^{th} simulated normalized sample for $i = 1, \ldots, n$ and $m = 1, \ldots, M$.

3. Convert the order statistic to the uniform scale using the normal distribution CDF. In other words, let $Y_{(i)}^m = \Phi(Z_{(i)}^m)$

These three steps require simulating from the standard normal distribution and normalizing these simulated samples. They are necessary steps to maintain the desired α-level significance value. In the case where the parameter of the normal distribution were known, we could simply simulate from the standard uniform distribution which allowed us to construct the entire test on the uniform scale and later translate the confidence bands back to the desired normal scale using the appropriate CDF. However, in the case of unknown parameters, we need to account for the uncertainty in parameters when we calculate the confidence bands and these two steps allow us to do so.

3 Power Analysis

We use simulations studies 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 is generated according to the alternative distribution.

We study the power of the TS testing procedure under two scenarios: (i) the mean μ and the standard deviation σ are pre-specified and known. (ii) the mean and the standard deviations are unknown. In the first scenario, we employ the confidence bands described in 2.1 and for the second we use 2.3 to construct the appropriate confidence bands.

To study the power of the TS test procedure we set the significance level to 5% and $n = 100$. We choose alternative distributions that were previously studied in similar power studies in Wilk and Gnanadesikan (1968) and Rogers and Tukey (1972). Table 1 lists the alternative distributions most of which are either skewed or heavy tailed.

Although we only present the results for sample size $n = 100$ we have conducted similar studies with sample size ranging between $n = 20$ and $n = 1000$ and the general pattern of results holds throughout the different sample sizes.

3.0.2 Known mean and standard deviation

The procedure detailed in 2.1 assumes the parameters of the normal distribution are known. In this section, we use the theoretical mean and standard deviations of each of the alternative distributions listed in 1 to normalize

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

Table 1: Power analysis at nominal level $\alpha = 0.05$ for $n = 100$. The boldface numbers in each row show the largest power against the given alternative distribution. The SE range between 0.0000 and 0.0168.

the sample. We compare the performance of the TS testing procedure to the Kolmogorov-Smirnov (KS) and the Anderson-Darling (AD) tests, both of whom originally require the mean and standard deviation to be known. We set all three procedures to a $\alpha = 0.05$ significance level and it can be shown that all of them approximately achieve the desired size.

Table 1 and Figure 7 summarize the power analysis results. As can be seen, our procedure generally outperforms both the KS and AD tests. The improvement is most apparent in the heavy tailed distributions such as $\chi^2(30)$, $t(2)$, Laplace and the first Normal mixture. We also see an advantage using this 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

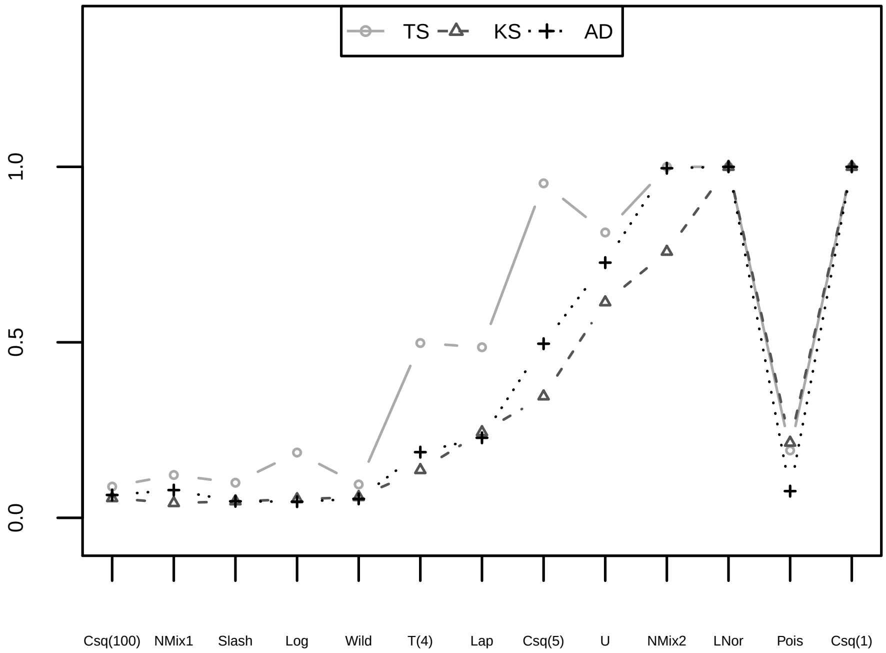

Figure 7: The power for each of the test procedures at each alternative distribution. The x-axis distributions are ordered by the Kullback-Leibler distance between the alternative distribution and the appropriately scaled and centered normal distribution, i.e. KL(F_1,$N(\mu_1, \sigma_1^2)$) where $\mu_1 = E_{F_1}(X)$ and $\sigma_1^2 = E_{F_1}(X - \mu_1)^2$.

and the Wild [2], Slash [3] and one of the Normal mixtures [4] distributions that are studied. Morgenthaler and Tukey (1991) referred to these three distributions as the corner distributions and used them to model extreme behavior in data. These are all distributions that are symmetric but heavier tailed compared to a normal distribution. However, they are not as heavy tailed as a Cauchy distribution and as such they may be harder to distinguish from the normal distribution. We can see that in the Normal mixture distributions the test can distinguish easily when the mixture probability is 10% and not as well when that probability is low. We experimented with different values for p and the standard deviations σ_1 and σ_2; it seems that the power goes down significantly

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

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

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

when the mixture probability p goes down and is far less sensitive to small changes in σ_1 and σ_2 (for a fixed p).

4 Unknown parameters power analysis

As we previously discussed, the procedure detailed in 2.1 assumes the parameters of the normal distribution are known. However, most applications do not have known parameters values and therefore require the use of estimated values instead.

The key issue is to chose wisely the parameter estimates that will allow us to best tell apart scenarios where the underlying distribution is normal from other distributions. In other words we are interested in testing the following hypotheses:

$$\begin{aligned} H_0 : X_i &\sim N(\mu, \sigma^2) \text{ for } i = 1, \ldots, n \qquad (6) \\ H_1 : X_i &\sim F_1 \text{ for } i = 1, \ldots, n \end{aligned}$$

where μ and σ are unknown and F_1 denotes a continuous distribution different from the normal distribution. Particularly, we would like our procedure to effectively distinguish between the normal distribution and similar symmetric distributions that are heavier tailed. Pictorially this means that if the data follows a normal distribution then we simply need to adjust the intercept (location) and slope (scale) of the quantile-quantile line such that all of its points fall within the bands $(1 - \alpha)100\%$ of the time. However, if the data does not follow a normal distribution it will be more difficult (probabilistically speaking) to find a pair of location and scale estimators that will produce a quantile-quantile line that is entirely contained in the confidence bands.

There are a few standard suggestions as to how one might go about estimating the mean and standard deviation of the normal distribution. An obvious choice is to use the standard maximum likelihood based estimators (MLE):

$$\begin{aligned}\breve{\mu} &= \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \\ \breve{\sigma} &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\end{aligned}$$

which are simply the sample mean and the sample standard deviation. Although this choice of estimators seems the most reasonable under the null distribution it does not guarantee the most powerful testing procedure against the alternatives that we are interested in detecting. Therefore we also explore more robust estimators for the location and scale that allow us to both maintain the appropriate significance level but will be more powerful in detecting heavier tailed distributions.

A more robust alternative is to use the median, $m(x)$ and the median absolute deviation (MAD)

$$mad(x) = m(|x - m(x)|) \cdot 1.4826$$

Lloyd (1952) suggested using the following generalized least-squares estimators based on the order statistics

$$\begin{aligned}\tilde{\mu} &= \bar{x} \\ \tilde{\sigma} &= \frac{m^t V^{-1} y}{m^t V^{-1} m}\end{aligned}$$

In this case, the estimator of the standard deviation σ is the robust scale estimator that Shapiro and Wilk (1965) used to construct the numerator for their statistic, as referred to in (5).

In Croux and Rousseeuw (1993) the authors proposed an alternative robust estimator for the scale. Their estimator, denoted by Q_n, is a robust estimator but is both a more efficient estimator than the MAD and it does not rest on an underlying assumption of symmetry like the MAD. Q_n is defined as

$$\tilde{\sigma}_{RC} = 2.219144 \cdot \{|x_i - x_j|; i < j\}_{(0.25)}$$

where $\{\cdot\}_{(0.25)}$ denotes the 0.25 quantile of the pair distances $\{|x_i - x_j|; i < j\}$. We pair the Q_n estimate with the median, $m(x)$, as the location estimate as recommended by these authors.

All of the options listed above are estimators of location and scale and we would like to advise the researcher which one to use based on power analysis. We compare the power of our method under four different alternatives:

1. Using the biased-adjusted maximum likelihood estimators (MLE)

2. Using the median and MAD (MM)

3. Using the generalized least squares estimators suggested by Lloyd (GLS)

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

Table 2: Power analysis with unknown parameters. The bold-faced number in each row indicates the highest power against the listed alternative distribution. The SE based on simulation range between 0.0000 and 0.0162.

4. Using Q_n and the median as suggested by Rousseeuw and Croux (Q_n)

We compare our method performance using the above listed options with the following more common testing procedures:

5. the Shapiro-Wilk (SW)

6. the Lilliefors test (LI)

7. the Cramèr-Von-Mises (CVM)

8. an adjusted (for unknown parameters) Anderson-Darling test (AAD)

The outcomes of the power analysis are listed in Table 2 and presented in Figure 8. These results indicate that one should use the pair of median and Q_n to estimate the location and scale parameters. In general our procedure with this choice of location and scale estimator performs at least as well as the Shapiro-Wilk testing procedure (if not better). It is interesting to notice that

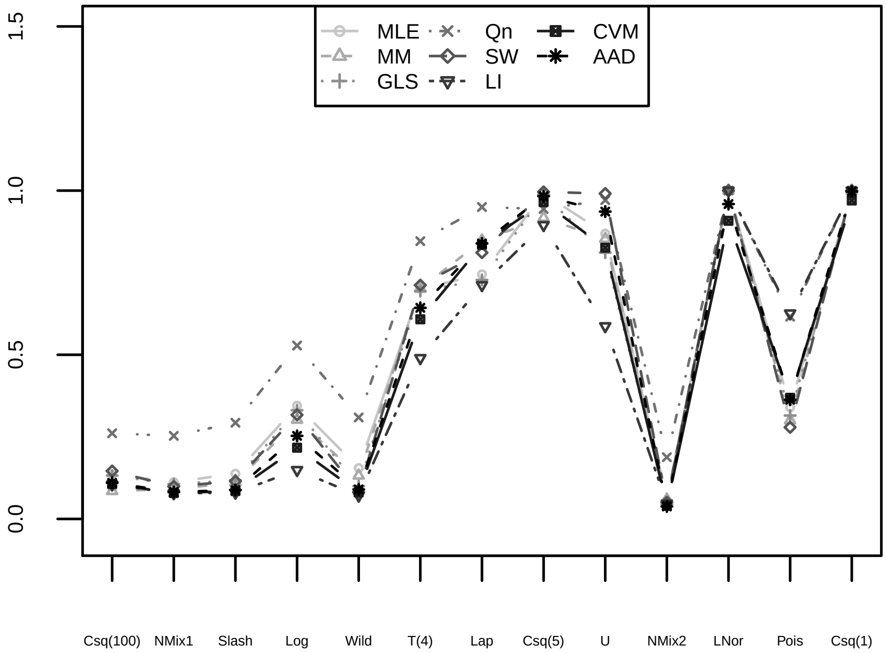

Figure 8: The power for each of the test procedures at each alternative distribution. The x-axis distributions are ordered by the Kullback-Leibler distance between the alternative distribution and the appropriately scaled and centered normal distribution, i.e. KL(F_1,$N(\mu_1, \sigma_1^2)$) where $\mu_1 = E_{F_1}(X)$ and $\sigma_1^2 = E_{F_1}(X - \mu_1)^2$.

the our method performs relatively well in Tukey's three corner distributions. It seems that even if we knew the true location and scale parameters the power would still be lower than if we use the Q_n and median estimators. One explanation as to why this is the case 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.

5 Discussion

The TS procedure introduces an attractive alternative to the commonly used KS testing procedure. It offers a visual method in combination with the classical normal quantile-quantile plot. The confidence bands for this procedure also yield a test whether an observed sample follows a normal distribution. Most testing procedures can distinguish well between the normal distribution and non-symmetric or symmetric very heavy tailed distributions. However, they under-perform when asked to tell apart a normal distribution from a mild heavy tailed symmetric distribution. The TS procedure performs reasonable well even for such alternatives.

We explore the performance of this procedure both when the parameters of the normal distribution are fully specified and when they are not a-priori available. The proposed procedure can be modified to handle other fully specified null continuous distributions. Future research may explore the power of this procedure for distributions other than the normal distribution.

Whether or not we know the parameters of the normal distribution, our procedure 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 $\frac{C_\alpha}{\sqrt{n}}$ to calculate the margin of error around the 45 degree line as n grows. Our motivation for exploring such an equation is because the KS confidence bands exhibit such a limiting behavior and its C_α values have been previously listed in Smirnov (1948). This simple asymptotic behavior is part of its appeal. After careful consideration that is detailed in the supplementary material, 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. However, there are applications that require a relaxation of these assumptions. One such example in 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 will not strictly apply. One can use the studentized residuals to adjust for the hetroscedasticity issue however the TS procedure will still need to be modified to account for the dependence between these residuals. We leave this modification to be further studied in future research.

References

Anderson, T. W. and Darling, D. A. (1954), A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769. ISSN 01621459. doi: 10.2307/2281537.

Chibisov, D. M. (1964), Some theorems on the limiting behavior of empirical distribution functions. *Selected Transcripts in Mathematical Statistics and Probability*, (6):147–156.

Croux, C. and Rousseeuw P. J. (1993), Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273. doi: 10.2307/2291267.

Darling, D. A. (1957), The Kolmogorov-Smirnov, Cramer-Von Mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838. ISSN 0003-4851. doi: 10.1214/aoms/1177706788.

Einmahl, H. J. and Mason, D. M. (1985), Bounds for weighted multivariate empirical distribution functions. *Z. Wahrscheinlichkeitstheorie*, (70):563–571.

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(1):221–234.

Feller, W. (1948), On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, 19(2):177–189. ISSN 0003-4851. doi: 10.1214/aoms/1177730243.

Kolmogoroff, A. (1941), Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, 12(4):461–463. ISSN 0003-4851. doi: 10.1214/aoms/1177731684.

Lilliefors, H. W. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

Lloyd, E. H. (1952), Least-squares estimation of location and scale parameters using order statistics. *Biometrika*, 39(1-2):88 –95. doi: 10.1093/biomet/39.1-2.88.

Morgenthaler, S. and Tukey, J. W. (1991), *Configural Polysampling: A Route to Practical Robustness*. Wiley-Interscience. ISBN 0471523720.

O'Reilly, N. E. (1974), On the weak convergence of empirical processes in sup-norm metrics. *Annals of probability*, 2(4):642–651.

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(1):21–33.

Rogers, W. H. and Tukey, J. W. (1972), Understanding some long-tailed symmetrical distributions. *Statistica Neerlandica*, 26(3):211–226. ISSN 0039-0402. doi: 10.1111/j.1467-9574.1972.tb00191.x.

Shapiro, S. S. and Wilk, M. B. (1965), An analysis of variance test for normality (complete samples). *Biometrika*, 3(52).

Smirnov, N. (1948), Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281. ISSN 0003-4851. doi: 10.1214/aoms/1177730256.

Stephens, M. A. (1974), EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737. ISSN 01621459.

Tsay, R. S. (2010), *Analysis of financial time series.* John Wiley & Sons, Inc, 3 edition.

Unknown (2012) *Testing of Body Armor Materials for Use by the U.S. Army - Phase III report.* The data cited in the paper is available in Appendix M.

Wilk, M. B. and Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17. ISSN 00063444. doi: 10.2307/2334448.

DasGupta, A. (2011) Probability for statistics and machine learning : fundamentals and advanced topics. Springer. ISBN 9781441996336.