Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Preliminary Goodness-of-Fit Tests for Normality do not Validate the One-Sample Student *t*

William R. Schucany & H. K. Tony Ng

To cite this article: William R. Schucany & H. K. Tony Ng (2006) Preliminary Goodness-of-Fit Tests for Normality do not Validate the One-Sample Student *t* , Communications in Statistics - Theory and Methods, 35:12, 2275-2286, DOI: 10.1080/03610920600853308

To link to this article: https://doi.org/10.1080/03610920600853308

Published online: 08 Dec 2006.

Communications in Statistics—Theory and Methods, 35: 2275–2286, 2006
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920600853308

Taylor & Francis
Taylor & Francis Group

Statistical Education

Preliminary Goodness-of-Fit Tests for Normality do not Validate the One-Sample Student *t*

WILLIAM R. SCHUCANY AND H. K. TONY NG

Department of Statistical Science, Southern Methodist University,
Dallas, Texas, USA

One of the most basic topics in many introductory statistical methods texts is inference for a population mean, μ. The primary tool for confidence intervals and tests is the Student t sampling distribution. Although the derivation requires independent identically distributed normal random variables with constant variance, σ^2, most authors reassure the readers about some robustness to the normality and constant variance assumptions. Some point out that if one is concerned about assumptions, one may statistically test these prior to reliance on the Student t. Most software packages provide optional test results for both (a) *the Gaussian assumption and* (b) *homogeneity of variance. Many textbooks advise only informal graphical assessments, such as certain scatterplots for independence, others for constant variance, and normal quantile–quantile plots for the adequacy of the Gaussian model. We concur with this recommendation. As convincing evidence against formal tests of* (a)*, such as the Shapiro–Wilk, we offer a simulation study of the tails of the resulting conditional sampling distributions of the Studentized mean. We analyze the results of systematically screening all samples from normal, uniform, exponential, and Cauchy populations. This pretest does not correct the erroneous significance levels and makes matters worse for the exponential. In practice, we conclude that graphical diagnostics are better than a formal pretest. Furthermore, rank or permutation methods are recommended for exact validity in the symmetric case.*

Keywords Adaptive procedure; Monte Carlo simulation; Nonparametric; Permutation; Pretest; Robustness; Shapiro–Wilk.

Mathematics Subject Classification Primary 62F03; Secondary 62A01.

1. Introduction

Is it ever a good idea to test for normality before using a sample $X_1, X_2, \ldots, X_n$ to make inferences about a mean, μ, relying upon the Student t sampling

Received October 7, 2005; Accepted March 17, 2006

Address correspondence to William R. Schucany, Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA; E-mail: schucany@smu.edu

distribution? Most software packages provide optional test results for (a) the Gaussian assumption (e.g., SAS, PROC UNIVARIATE) and (b) homogeneity of variance (e.g., SAS, PROC TTEST; SPSS, The Independent-Samples *t* Test procedure). Good textbooks on statistical methodology, e.g., Ramsey and Shafer (2002, p. 73), argue against formal preliminary tests and favor informal graphical diagnostics. These applied statistics texts advise only graphical assessments, such as certain scatterplots for independence, others for constant variance, and normal quantile–quantile plots for the adequacy of the Gaussian model. We support this recommendation with concrete evidence for the case that formal tests of (a), as recommended by Romeu (2003), are actually flawed. In teaching beginning courses we find it useful to address this basic point directly. For paired data and multiple samples the issues are different.

Suppose this random sample of size n is from a distribution F with population mean, μ. A statistical practice that is possible in many packages is to first use a goodness-of-fit (GOF) test for normality, i.e.,

$$\begin{aligned} H_0^* &: \text{The true } F \text{ is normal} \\ \text{against } H_1^* &: F \text{ is not normal} \end{aligned} \tag{1}$$

at the level of significance α_g before making inference based on the t-statistic. If one does not reject H_0^*, one treats the sample as coming from a normal distribution and uses the Student t-statistic

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}, \quad \text{where } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

to test the hypothesis

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ \text{against } H_1 &: \mu \neq \mu_0 \end{aligned} \tag{2}$$

at the level of significance α_t. In order to assess the adequacy of the preliminary GOF tests for normality, we compare the true Type-I error rate given that the sample passed this pretest, i.e.,

$$\alpha = \Pr(\text{Reject } H_0 \,|\, \text{do not reject } H_0^* \text{ and } H_0 \text{ is true}), \tag{3}$$

to the pre-specified nominal Type-I error rate, α_t.

Thoughtful readers recognize a logical problem here. The nominal level, α_t, requires that H_0^* be true but in (3) one is only given that H_0^* was not rejected. Consequently, hoping for this two-stage procedure to work puts one in the ill-advised position of *accepting* a narrow null hypothesis. Strictly speaking, this null of normality is not simple, but it is one quite special family in the space of all continuous distributions.

On the parallel issue (b) of common variance, preliminary testing of $\sigma_1^2 = \sigma_2^2$ has a more extensive literature. The Behrens–Fisher problem has been troublesome for decades. A fresh approach that does not pretest (b) is given by Sprott and Farewell (1993). Their position agrees well with ours that “a procedure, based on *accepting* the null hypothesis, seems logically flawed.” They give a collection of confidence intervals for a difference between two means that adapt to the statistical evidence on the ratio σ_2^2/σ_1^2.

There is a sizable literature on GOF and much of it pertains to tests for normality or a few other specific families. There is a growing consensus that neither the chi-square (Moore, 1986, p. 91) nor the Kolmogorov–Smirnov can be recommended due to the superiority of many others based on moments, probability plots, or other empirical distribution function (EDF) tests. Tukey (1993, p. 31) stated that "The Kolmogoroff–Smirnov construction ... is logically correct, but practically almost useless." He cited Michael (1983) for a transformation better suited to the sup norm. For this reason we concentrate only on the Shapiro–Wilk statistic, W (Shapiro and Wilk, 1965) and the EDF Anderson–Darling statistic, A^2 (Stephens, 1974) for normality. See D'Agostino and Stephens (1986, Chapter 9) for definitions, theory, and efficiency comparisons in finite samples. Because we find the results using these two pretests to be very similar, we only report those for W, which essentially tests for linearity in normal quantile–quantile plots.

2. A Simulation of the Effects of a Preliminary Test

The algorithm for our Monte Carlo study follows.

Step 1. Simulate a random sample of size n from a distribution F.

Step 2. Use the Shapiro–Wilk statistic, W, to test for normality (1), at the level of significance α_g with the International Mathematical and Statistical Libraries (IMSL) function DSPWLK (Visual Numerics, Inc., 1994).

Step 3. If H_0^* is not rejected in Step 2, treat the sample as coming from a normal distribution and use the t-test of the hypothesis (2) at the level of significance α_t. If H_0^* is rejected in Step 2, then return to Step 1.

In this simulation experiment we consider sample sizes n = 10, 20, 30, and 50; and the following underlying distributions F, (i) uniform (0, 1), $\mu_0 = 0.5$, (ii) standard exponential, $\mu_0 = 1.0$, (iii) Cauchy, median = 0.0, (iv) standard normal, $\mu_0 = 0.0$. There are four fixed levels of significance for the preliminary GOF test α_g = 10%, 5%, 1%, and 0.5% and crossed with four levels of significance for the t-test α_t = 10%, 5%, 1%, and 0.5%. For each combination of n, α_g, α_t, and F, we independently repeat Steps 1 through 3 until H_0^* is not rejected $M = 100{,}000$ times. Hence this inverse sampling yields 100,000 samples that have passed our normality screening. Then the Type-I error rate in (3) for the t-test is estimated by

$$\hat{\alpha} = \frac{\text{number of times } H_0 \text{ (2) is rejected}}{M}. \tag{4}$$

Easterling and Anderson (1978) report a similar Monte Carlo study of $n = 10, 20$, $\alpha_g = 10\%$, and $M = 1000$. They employ the chi-square (12 bins for deciles and 5th and 95th percentiles) to assess the resulting agreement with the entire Student t sampling distributions. We focus on the relevant feature of the quantiles for inference at four conventional levels, e.g., 90% confidence. Our greater precision allows us to reach stronger conclusions. We also gain some valuable insight on the sensitivity to pretest levels down to $\alpha_g = 0.5\%$. Even though the t-test is not justifiable for the Cauchy, it is useful to assess its performance.

3. Results of Screening Out Nonnormal Samples

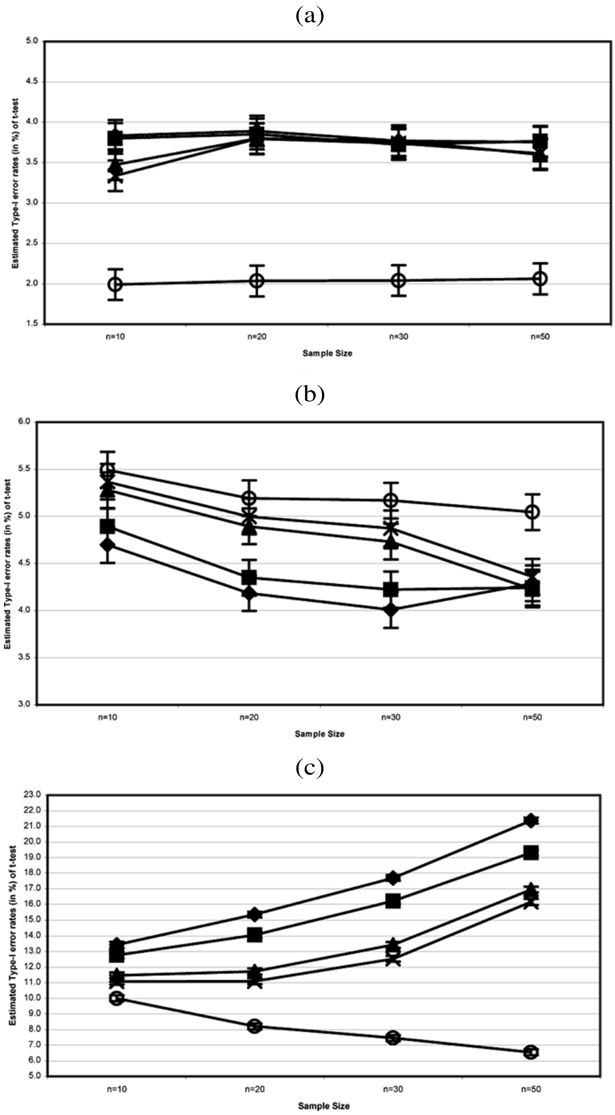
Tables 1–5 in the Appendix summarize the Type-I error rates for four pretest levels α_g and without any pretest, four conventional levels of significance, α_t, and four different underlying distributions in our experiment. These Type-I error rates are estimated from 100,000 replications of the Student t-statistic. The standard errors (SE) at each of the nominal levels are at the head of each respective column. This final power column has a maximum $\mathrm{SE} = 0.12\%$ because of the negative binomial stopping rule to produce 100,000 acceptable samples. The number of simulated samples to produce 100,000 samples passing Shapiro–Wilk pretest can be obtained from this final power column. For example in Table 3, ($\alpha_g = 1\%$), when F is the exponential and $n = 10$, the simulated power of Shapiro–Wilk test is 23.4%, implying $100{,}000/(1 - 0.234) = 130{,}548$ samples to produce 100,000 passing the GOF pretest.

When F is Gaussian, as one might hope, the preliminary GOF test does no harm whenever the data are truly normally distributed. When the underlying distributions are nonnormal, one might hope that the preliminary goodness-of-fit tests screen out the blatantly nonnormal samples and correct the Type-I error rates for the t-test. However, our results show this not to be the case in general. To demonstrate the effect of the pretest on the Type-I error rate for the t-test, Fig. 1 plot the independent estimates (4) for the t-test at $\alpha_t = 5\%$ after passing Shapiro–Wilk tests at $\alpha_g = 10\%, 5\%, 1\%, 0.5\%$, and also with no GOF test. Some interesting interactions are apparent.

From Fig. 1(a), when F is Cauchy, we see that for sample sizes from 10 to 50 the GOF screens yield somewhat better results (simulated Type-I error rates closer to the nominal level) than applying the t-test without any pretest. For example, at $n = 30$ and $\alpha_t = 5\%$, the true Type-I error rate improved from 2.0 to 3.7% ($\mathrm{SE} = 0.07\%$). However, the simulated Type-I error rates are still significantly less than the nominal α_t for every n in our study. The absence of any effect of n is consistent with the theory for the Cauchy. Multiple comparisons are protected at each n. The vertical bars at each plotting symbol are 95% family-wise Tukey-corrected intervals for the 10 pairwise differences among the five rates. For displays of such intervals see Mason et al. (2003, Sec. 6.5).

The genuinely noteworthy effect is that the pretest actually makes matters worse for both the uniform and the exponential. That is, for these two populations the true Type-I error rates are closer to the nominal α_t when one uses the t-test on every sample, rather than selectively using only those samples with acceptable preliminary GOF statistics. From Fig. 1(b), when the underlying distribution is uniform, GOF tests at $\alpha_g = 1\%$ and 0.5% corrected the problem for small sample sizes ($n = 10$ and 20). However, for large sample sizes, the true Type-I error rates for the t-test without the GOF pretest are closer to the nominal α_t. For example, at $n = 50$ and $\alpha_t = 5\%$ the estimated Type-I error rates with preliminary GOF tests are around 4.2 to 4.4%, while the rate without any pretest is 5.0% (from Table 5, $\mathrm{SE} = 0.07\%$). This can be explained by the rapid convergence to normality of the mean of samples from the symmetric short-tailed uniform distribution. The degradation of the conditional test is due to increasing power and corresponding greater selectivity for samples with relatively long tails.

From Fig. 1(c) when the underlying distribution is exponential, preliminary GOF tests not only do not help bring the conditional Type-I error rates closer to the nominal level, they actually makes matters worse than no pretest. This is especially


(a)

(b)

(c)

Figure 1. Conditional Type-I error rates (in %) of t-test for $\alpha_t = 5\%$ at different $\alpha_g(-\blacklozenge-$ $\alpha_g = 10\%$, $-\blacksquare-$ $\alpha_g = 5\%$, $-\blacktriangle-$ $\alpha_g = 1\%$, $-\times-\alpha_g = 0.5\%$, $-\circ-$ without GOF test); the underlying distribution is (a) Cauchy Distribution (b) Uniform Distribution, and (c) Exponential Distribution .

so when the sample sizes are large. This is our "smoking gun" evidence. For example, at $n = 50$ and $\alpha_t = 5\%$ the estimated Type-I error rate with preliminary goodness-of-fit tests at $\alpha_g = 0.5\%$ is 16.2% (from Table 4) while the estimated rate without is 6.5% (from Table 5, SE = 0.07%). Somewhat surprisingly, when the sample size increases, the conditional Type-I error rates increase, getting further away from the nominal rate, α_t. It appears to be quite detrimental to use the GOF test of normality before applying the t-test, because they yield actual Type-I error

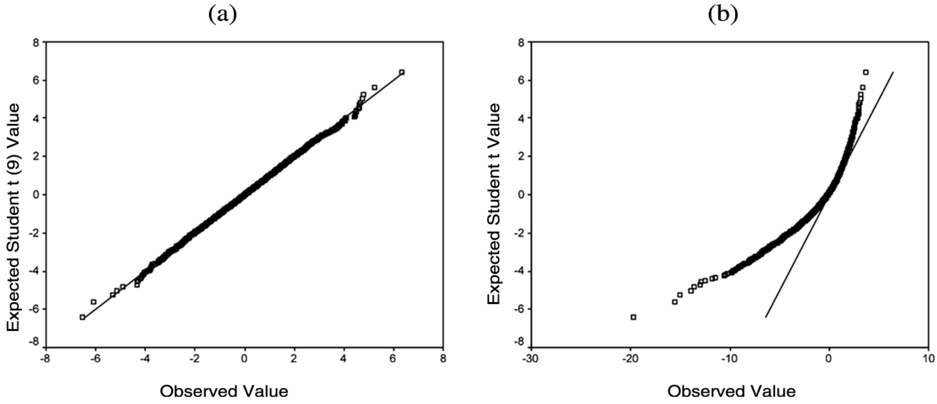


Figure 2. Student $t(9)$ Q–Q plot of 10,000 simulated samples of size $n = 10$ passing the Anderson–Darling GOF pretest at $\alpha_g = 10\%$ when the underlying distribution is (a) Normal Distribution and (b) Exponential Distribution.

rates that are much greater than the nominal level. The reason is subtle. What is being detected is a changing selection effect on the conditional mean value. Even though we are likely to detect these skewed distributions in large samples, for small $n = 10$ the $\alpha_g = 5\%$ level tests let about 56% of the samples be treated as normal.

Graphical evidence that even Anderson–Darling pretests do no harm to the normal may be seen in Fig. 2(a). Using A^2 at $\alpha_g = 10\%$ for $n = 10$ selects about 90% of normal samples with relatively light tails. Consequently, one might suspect a distortion of some features of the subsequent $t(9)$ sampling distribution. There is no evidence of any such selection bias in Q–Q plots of 10,000 Student t with nine degrees of freedom. On the other hand in Fig. 2(b), the disparity is quite obvious, when we subject the exponential to the same process.

4. Other Valid Inference Procedures

In practice, reliable inference on means, when there is convincing evidence of nonnormality, takes two or more non-Student paths. The issues are different for skewed and symmetric alternatives. Our symmetric nonnormals, the Cauchy and the uniform, represent relatively heavier- and lighter-tailed symmetric alternatives, respectively. For both of these, one might consider using the Wilcoxon signed-rank or a permutation distribution (Ernst, 2004) either (i) on the rejected samples or (ii) on all samples from the outset. We offer a portion of a larger simulation study that strongly suggests that (ii) is the recommended strategy.

Table 6 summarizes two analyses of $M = 100{,}000$ samples of size $n = 10$ for only one nominal level $\alpha_t = 5\%$. The empirical levels in the table for both distributions are subject to a simulation SE = .07%. In contrast to the approach in Sec. 3, we assess the Type-I error rates incurred by the adaptive procedure with the t-test on "accepted" samples and distribution-free procedures on the "rejected" ones. The two-sided Wilcoxon-signed rank is exact of size 4.88% (see Lehmann, 1998, p. 123ff). The exact conditional permutation distribution has $2^{10} = 1{,}024$ points. Thus, its true level can be closer to 5%.

What is evident in Table 6 is that the validity of the t-test is not repaired for the Cauchy and the uniform by resorting to either of these distribution-free tests on the significantly nonnormal samples. As we know, it does control Type-I error rates exactly to use either Wilcoxon or permutation tests on every sample with no pretesting. However, the net result of the two stages in (i) can be significantly different than the nominal level. For instance, the Shapiro–Wilk at $\alpha_g = 1\%$ yields a combined level for the t/Wilcoxon of 4.52% (t/Permutation of 4.58%) for the Cauchy, which is detected about 48% of the time. These are significantly less than 5% with exact binomial two-sided p-values $<10^{-9}$. The corresponding results for the uniform, t/Wilcoxon of 5.42%, (t/Permutation of 5.48%) are significantly greater with p-values $<10^{-9}$.

The effect of α_g is in the anticipated direction. That is, larger α_g yield greater power, which improves the approximation to α_t. However, the important conclusion is that none of these two-stage procedures have the desired validity, which is present in the appropriate rank or permutation test with no pretest (option (ii)). On the other hand, more refined adaptive tests can produce satisfactory increases in power for paired samples, see Freidlin et al. (2003).

We analyzed these same procedures at $n = 20, 30$, and 50 as well as for other levels $\alpha_t = .10, .01$, and .005. The patterns were consistent with the ones for $n = 10$ and $\alpha_t = 5\%$; although the effect of α_g was less pronounced at $n = 50$ for the Cauchy, where the corresponding powers were .996, .993, .987, and .984.

The proper handling of skewed alternatives is much more difficult. If conditions are right for the lognormal, then the log transformation is ideal. However, other right-skewed families, such as exponential, gamma, Weibull, Gumbel, inverse Gaussian, and Pareto, are not perfectly symmetrized by the logs, reciprocals, or square roots, which tend to be used in practice.

The fundamental problem comes from trying to make inference on the mean of any skewed distribution. When these are paired differences, symmetry is a natural result. When there are two independent samples, permutation tests permit inference on shift parameters without requiring normality or even symmetry. Therefore, the one-sample location problem for skewed families is not as well posed as when μ is the center of symmetry. All of the other inferences about comparisons are reasonably well handled by permutation tests. We recommend that they be used in parallel with t-tools and not only whenever the sample appears to be nonnormal.

5. Conclusions and Discussions

There are some apparently counterintuitive results in Sec. 3. If one does any preliminary test of H_0^*, it appears best to use a very small level, e.g., $\alpha_g = 0.5\%$. This may seem to refute many analysts' experience with GOF tests, which have notoriously low power in small to moderate sample sizes. Hence, it may seem that one would argue for $\alpha_g = 10\%$ to have any chance of screening out the Cauchy or the exponential. Even though our intuition may be right about the power of these tests, it fails us whenever we imagine that the Student t approximation is conditionally better, when one rejects more of the exponential samples. This can be seen in Fig. 1(c) that the exponential is best with no preliminary tests. The pretest's selection effect on the *conditional* mean depends on both α_g and n.

Based on the results here, we recommend that one use formal GOF tests of normality before applying the t-test with great caution, especially when the underlying population distribution is suspected to be skewed. This precaution holds for any other tests like F-tests, that are sensitive to nonnormality. When the sample size is reasonably large, say $n \geq 50$, the central limit theorem supports the Student t approximation and any pretest is ill-advised. If one wishes to use the GOF test of normality, we suggest fixing the level of significance at a very low level, say $\alpha_g = 0.1\%$. Diagnostic plots and tests should guide us to transform clearly nonnormal data. When nonnormality is blatant in such graphics, it may be a surrogate for $p < .001$ in a test. Obviously, the same conclusion holds for confidence intervals. The inaccuracies that we demonstrate in each tail of the conditional sampling distributions have precisely the same effect on the coverage probabilities of one- or two-sided intervals for μ. Easterling (1976) described a simultaneous inferential procedure for both the distributional model and its parameters.

In practice for the one-sample problem our recommendation is to use t-tests and intervals along with permutation methods. As a standard part of the diagnostics, the next stage should include normal quantile-quantile or log density (Hazelton, 2003) plots. Whenever these (or GOF tests) have strong evidence of obvious nonnormality, one should rely on the tests and intervals from rank-based and permutation distributions or look for satisfactory transformations for improved inference on μ. This seems to be a better strategy over many data analyses, than one of screening all samples with a pretest for normality. We find this point to be helpful for beginning students.

Appendix

Table 1
Simulated Type-I error rates (in %) of the t-test after acceptable Shapiro–Wilk tests of normality with $\alpha_g = 10\%$

Underlying distribution	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02	Power of Shapiro–Wilk test
Uniform	10	8.8	4.7	1.3	0.8	16.2
	20	8.5	4.2	0.9	0.5	36.0
	30	8.2	4.0	0.8	0.4	61.8
	50	8.8	4.3	0.9	0.5	95.1
Exponential	10	19.6	13.4	6.5	5.0	56.5
	20	22.9	15.4	7.2	5.6	90.3
	30	25.7	17.7	8.2	6.2	98.6
	50	29.8	21.4	10.3	7.5	99.9
Cauchy	10	8.8	3.8	0.5	0.2	66.3
	20	8.5	3.9	0.6	0.2	89.4
	30	8.3	3.8	0.6	0.3	96.6
	50	8.0	3.8	0.6	0.3	99.6
Normal	10	10.1	5.1	1.0	0.5	9.9
	20	10.0	5.0	1.0	0.5	9.8
	30	10.1	5.1	1.1	0.5	9.7
	50	10.1	5.0	1.0	0.5	10.0

Table 2
Simulated Type-I error rates (in %) of the t-test after acceptable Shapiro–Wilk tests of normality with $\alpha_g = 5\%$

Underlying distribution	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02	Power of Shapiro–Wilk test
Uniform	10	9.2	4.9	1.4	0.8	7.5
	20	8.8	4.4	0.9	0.5	20.4
	30	8.5	4.2	0.9	0.4	42.8
	50	8.8	4.2	0.8	0.4	88.1
Exponential	10	18.5	12.7	6.1	4.6	44.3
	20	21.2	14.0	6.5	4.9	83.5
	30	23.9	16.2	7.3	5.5	96.8
	50	27.2	19.3	8.9	6.2	99.9
Cauchy	10	8.8	3.8	0.4	0.2	60.1
	20	8.5	3.9	0.5	0.2	86.4
	30	8.3	3.7	0.5	0.3	95.3
	50	8.3	3.8	0.7	0.3	99.4
Normal	10	10.1	5.1	1.0	0.5	5.0
	20	10.0	5.0	1.0	0.5	5.0
	30	10.0	5.1	1.1	0.5	4.9
	50	10.1	5.0	1.0	0.5	5.1

Table 3
Simulated Type-I error rates (in %) of the t-test after acceptable Shapiro–Wilk tests of normality with $\alpha_g = 1\%$

Underlying distribution	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02	Power of Shapiro–Wilk test
Uniform	10	10.0	5.3	1.4	0.8	1.0
	20	9.6	4.9	1.1	0.6	3.5
	30	9.3	4.7	1.0	0.5	12.2
	50	8.8	4.2	0.8	0.5	59.6
Exponential	10	16.7	11.5	5.4	4.0	23.4
	20	18.2	11.7	5.2	3.9	64.0
	30	20.6	13.4	5.7	4.2	88.5
	50	24.7	17.0	7.7	5.4	99.7
Cauchy	10	8.7	3.5	0.3	0.1	48.1
	20	8.5	3.8	0.5	0.2	79.5
	30	8.4	3.8	0.6	0.3	92.0
	50	8.0	3.6	0.5	0.2	98.8
Normal	10	10.2	5.1	1.1	0.5	1.0
	20	10.0	5.0	1.0	0.5	1.1
	30	10.0	5.1	1.0	0.5	0.9
	50	10.1	5.0	1.0	0.5	1.1

Table 4
Simulated Type-I error rates (in %) of the t-test after acceptable Shapiro–Wilk tests of normality with $\alpha_g = 0.5\%$

Underlying distribution	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02	Power of Shapiro–Wilk test
Uniform	10	10.1	5.4	1.5	0.8	0.4
	20	9.8	5.0	1.1	0.6	1.4
	30	9.6	4.9	1.0	0.5	6.0
	50	9.0	4.4	0.8	0.5	45.4
Exponential	10	16.2	11.1	5.1	3.8	17.4
	20	17.3	11.1	4.9	3.7	55.1
	30	19.5	12.5	5.3	3.9	83.1
	50	23.9	16.2	7.1	5.2	99.3
Cauchy	10	8.6	3.3	0.3	0.1	43.6
	20	8.6	3.8	0.5	0.2	76.5
	30	8.4	3.7	0.6	0.2	90.4
	50	8.0	3.6	0.6	0.3	98.5
Normal	10	10.2	5.1	1.1	0.5	0.5
	20	10.0	5.0	1.0	0.5	0.6
	30	10.0	5.1	1.0	0.5	0.5
	50	10.1	5.0	1.0	0.5	0.6

Table 5
Simulated Type-I error rates (in %) of the t-test without pretest

Underlying distribution	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02
Uniform	10	10.3	5.5	1.5	0.9
	20	10.0	5.2	1.2	0.7
	30	10.0	5.2	1.1	0.6
	50	10.1	5.0	1.0	0.5
Exponential	10	14.8	10.0	4.6	3.4
	20	13.0	8.2	3.5	2.5
	30	12.2	7.5	2.9	2.1
	50	11.6	6.5	2.3	1.5
Cauchy	10	5.8	2.0	0.2	0.1
	20	6.1	2.0	0.2	0.0
	30	6.1	2.0	0.2	0.1
	50	6.2	2.1	0.2	0.1
Normal	10	10.1	5.1	1.0	0.5
	20	10.0	5.0	1.0	0.5
	30	10.0	5.1	1.0	0.5
	50	10.0	5.0	1.0	0.5

Table 6
Simulated Type-I error rates (in %) for several inference procedures with 5% nominal levels at $n = 10$

Procedure	Cauchy		Uniform	
	Without preliminary test			
Student t	1.99		5.50	
Wilcoxon**	4.96		4.96	
Permutation	5.01		5.07	
	With preliminary test			
$\alpha_g = 10\%$	Power = 66.4%		Power = 16.1%	
	t/Wilcoxon	t/Permutation	t/Wilcoxon	t/Permutation
H_0^* is not rejected	1.30	1.30	3.94	3.94
H_0^* is rejected	3.56	3.62	1.29	1.39
Combined	4.86	4.92	5.23	5.33
$\alpha_g = 5\%$	Power = 60.1%		Power = 7.5%	
	t/Wilcoxon	t/Permutation	t/Wilcoxon	t/Permutation
H_0^* is not rejected	1.54	1.54	4.53	4.53
H_0^* is rejected	3.26	3.32	0.78	0.86
Combined	4.80	4.86	5.31	5.39
$\alpha_g = 1\%$	Power = 48.1%		Power = 1.0%	
	t/Wilcoxon	t/Permutation	t/Wilcoxon	t/Permutation
H_0^* is not rejected	1.82	1.82	5.21	5.21
H_0^* is rejected	2.70	2.76	0.21	0.26
Combined	4.52	4.58	5.42	5.47
$\alpha_g = .5\%$	Power = 43.6%		Power = .39%	
	t/Wilcoxon	t/Permutation	t/Wilcoxon	t/Permutation
H_0^* is not rejected	1.90	1.90	5.34	5.34
H_0^* is rejected	2.45	2.51	0.10	0.14
Combined	4.35	4.42	5.44	5.48

**-exact type-I error rate = 4.88%

Acknowledgment

We would like to thank Ian Harris and Lynne Stokes for their comments on a preliminary version of the article. We greatly appreciate the issues raised by Pat Carmack, Mike Ernst, and Rob Easterling.

References

D'Agostino, R. B., Stephens, M. A. (Eds.) (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Easterling, R. G. (1976). Goodness of fit and parameter estimation. *Technometrics* 18:1–9.

Easterling, R. G., Anderson, H. E. (1978). The effect of preliminary normality goodness of fit tests on subsequent inference. *J. Statist. Computat. Simula.* 8:1–11.

Ernst, M. D. (2004). Permutation methods: a basis for exact inference. *Statist. Sci.* 19:676–685.

Freidlin, B., Miao, W., Gastwirth, J. L. (2003). On the use of the Shapiro-Wilk test in two-stage adaptive inference for paired data from moderate to very heavy tailed distributions. *Biometrical J.* 45:887–900.

Hazelton, M. L. (2003). A graphical tool for assessing normality. *Amer. Statistician* 57:285–288. Comment by Jones and reply, *The American Statistician* 58:176–177.

Lehmann, E. L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*. 2nd ed. New York: McGraw-Hill.

Mason, R. L., Gunst, R. F., Hess, J. L. (2003). *Statistical Design and Analysis of Experiments With Applications to Engineering and Science*. 2nd ed. New York: Wiley.

Michael, J. R. (1983). The stabilized probability plot. *Biometrika* 70:11–17.

Moore, D. S. (1986). Tests of chi-squared type. In: D'Agostino, R. B., Stephens, M. A., eds. *Goodness-of-Fit Techniques*. New York: Marcel Dekker, pp. 63–96.

Ramsey, F. L., Shafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. 2nd ed. Duxbury Press.

Romeu, J. L. (2003). Anderson–Darling: a goodness of fit test for small samples assumptions. *Reliabil. Anal. Center, Selected Topics in Assurance Related Technol.* 10(5).

Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika* 52:591–611.

Sprott, D. A., Farewell, V. T. (1993). The difference between two normal means. *Ameri. Statist.* 47:126–128.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.* 69:730–737.

Tukey, J. W. (1993). Graphic comparisons of several linked aspects: alternatives and suggested principles (with discussion). *J. Computat. Graphical Statist.* 2:1–49.

Visual Numerics, Inc. (1994). *IMSL/LIBRARY: FORTRAN Subroutines for Statistical Applications.* Texas: Houston.