Utility of Normality Test Methods

Oklahoma State University

Benedict Kongyir

September 2, 2024

The assumption of normality is foundational in statistical analyses, particularly when using parametric tests such as the *t-test*, Analysis of Variance (*ANOVA*), and regression analysis, which inherently assume that the data follows a normal distribution [Razali and Wah, 2011, Ghasemi and Zahediasl, 2012, Thode, 2002]. Deviations from normality can lead to inaccurate conclusions, making it critical to test for normality before applying these methods [Yap and Sim, 2011]. Various normality test methods, such as the Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov, and Lilliefors tests, have been developed, each with distinct advantages and limitations depending on the data context [Razali and Wah, 2011, Steinskog et al., 2007].

The effectiveness of these tests is typically evaluated based on their Type I error rates, which represent the probability of incorrectly rejecting the null hypothesis of normality, and their statistical power, the probability of correctly detecting deviations from normality. Despite their widespread use, questions persist regarding their real-world applicability, particularly their impact on the power of subsequent parametric tests such as *t-tests* and *ANOVA* [Ghasemi and Zahediasl, 2012, Al-Omar and Degawa, 2016].
In the following section, we assess/compare several available normality test methods by their ability to detect departures from normality for samples drawn from different probability distributions such as Uniform, Exponential, Chi-squared, etc for various sample sizes. Generally, they are more powerful for asymmetric distributions such as exponential, Chi-Squared, and LogNormal distributions. Their ability to detect departures from normality also improves as the sample size increases.

It is also important to note that none of them is uniformly most powerful. For some distributions, the Shapiro-Wilk test method is the most powerful, but for others, Shapiro-Franca is the most powerful, and Cramer Vone Miss. It is interesting to note how the performance of these normality test procedures is similar for both the one-sample and two-sample cases. For samples drawn from Uniform, Exponential, Chi-Squared, Gamma, Weibull, LogNormal, and Pareto distributions, the Shapiro-Wilk test method is the most powerful. For samples drawn t, and Laplace distributions, Shapiro-Francia is the most powerful. Cramer Vone Misses is most powerful for samples taken from Contaminated distribution.

Comparison of Power of Normality Test Methods

One Sample

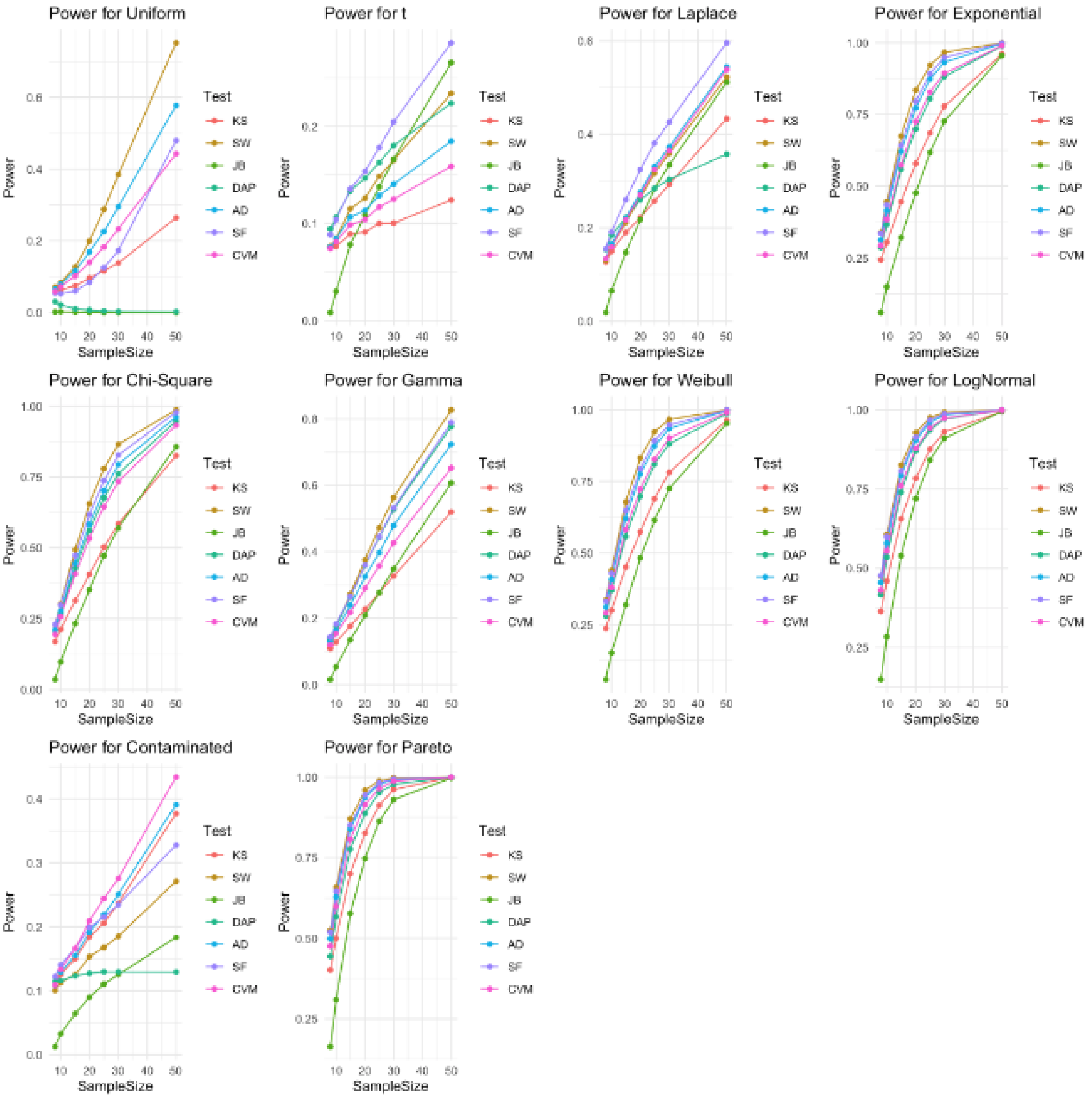

Figure 1: Showing power comparison of different normality test methods for different distributions

Two Samples

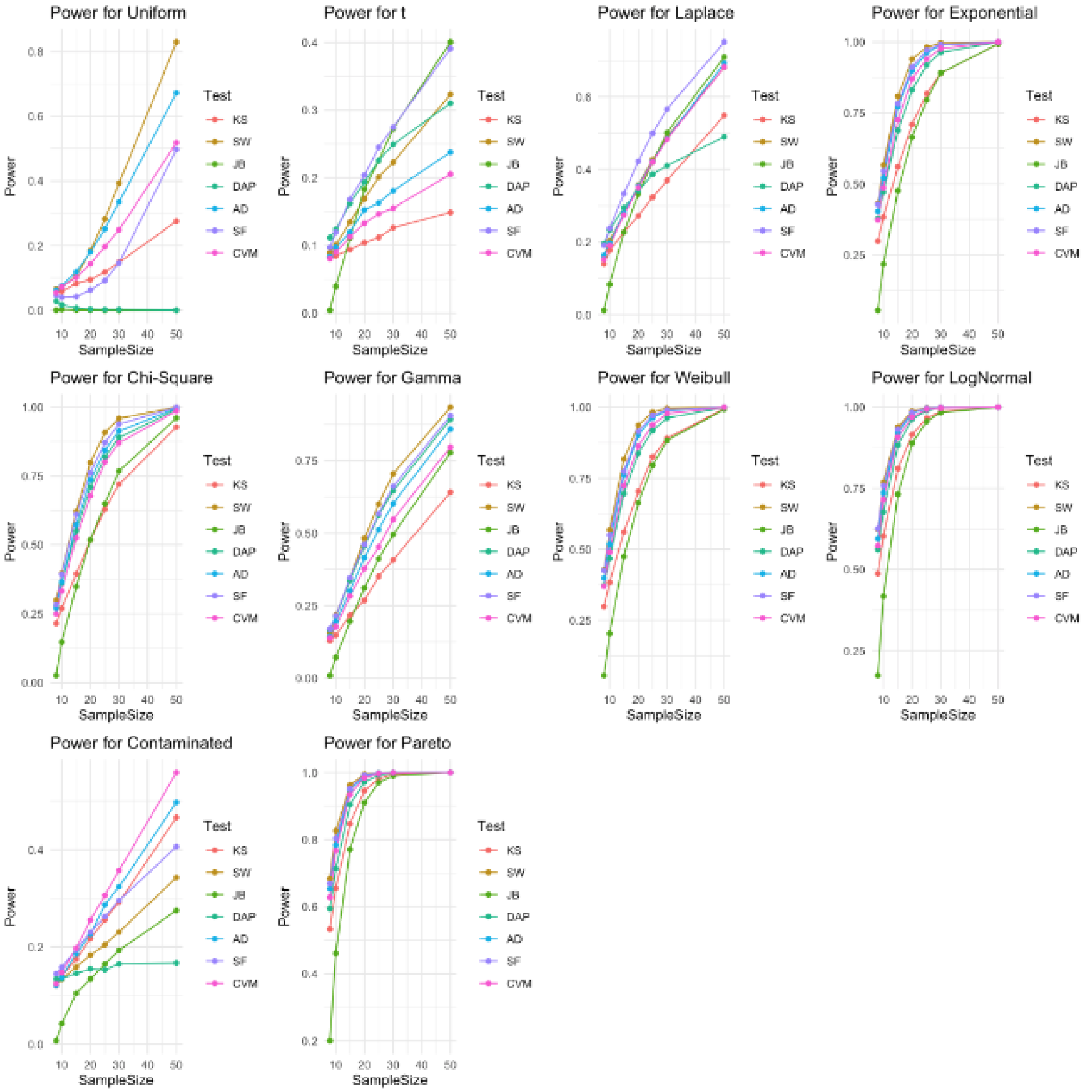

Figure 2: Showing power comparison of different normality test methods for different distributions

The Effects of Non-Normality in Parametric Test Methods

As stated above, deviation from the assumption of normality impacts the probability of Type I error rate, and the power of the main, downstream test. The effects usually vary based on the test type, distribution type, and sample size. In this section, we assess the probability of Type I error rates for the one-sample t-test and the two-sample t-test methods for samples drawn from different distributions. First, the probability of Type I error rates is calculated for the one-sample and two-sample t-tests without checking for the assumption of normality. Second, the probability of Type I error rate is calculated for only samples for which the normality Shapiro-Wilk(SW) test is nonsignificant. The purpose is to ascertain how useful the SW normality test procedure is in terms of controlling Type I errors. We present the setup for the two-stage procedure for the one-sample test. The two-sample test setup follows similarly.

Effect of Non-Normality on t-tests

One-Sample t-test

The one sample test statistic is computed as:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}},$$

where $\bar{X}$ is the sample mean, μ_0 is the hypothesized population mean, s is the sample standard deviation, and n is the sample size.

Two-Sample t-test

The two sample test statistic for the two-sample t-test is given by:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes of the two groups.

Robustness of One-Sample and Two-Sample t-test to non-normality

When the underlying population is not normally distributed, particularly in the presence of skewness, the t-distribution assumption for the test statistic becomes invalid. Skewness affects the sample mean, $\bar{X}$, leading to biased estimates. Since the t-distribution is symmetric, applying it to a skewed distribution results in incorrect critical values. This typically leads to an inflated Type I error rate, making the one-sample t-test less robust under skewed conditions Huber and Ronchetti [2011]. The two-sample t-test however is less sensitive to deviations from normality, especially when sample sizes are equal. According to the Central Limit Theorem (CLT), the distribution of the test statistic approaches normality even if the underlying distributions are skewed, provided the sample sizes are sufficiently large Feller [1991]. When sample sizes are equal, the effect of skewness is averaged out, leading to a more accurate control of the Type I error rate. For one sample case, the distribution of the test statistic deviates significantly from the t distribution, but for two sample cases, the test statistic is approximately t.

Simulation Setup

Let N represent the total number of simulations(iterations), and α the significance level. In this study, $N = 1000$ and $\alpha = 0.05$.

The distributions examined are denoted as: Distributions: $\mathcal{D} = \{$Standard Normal, Uniform, t, Contaminated Normal, Laplace, Exponential, Chi-Square, Gamma, Weibull, LogNormal$\}$

Sample sizes are represented by the vector $\mathbf{n} = \{8, 10, 15, 20, 25, 30, 50\}$.

For each distribution $d \in \mathcal{D}$ and each sample size $n \in \mathbf{n}$, a sample $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is generated from the distribution d.

The Shapiro-Wilk test is applied to each generated sample $\mathbf{x}$ to assess normality. For each valid sample (i.e., those where the Shapiro-Wilk test did not reject normality), a t-test is performed and the conditional Type I error is calculated as:

$$\text{Type I Error} = \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ is true}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{p_i < \alpha\}$$

Where p_i is the p-value from the t-test on the i-th sample, and $\mathbf{1}\{\cdot\}$ is an indicator function that equals 1 if the condition inside is true, and 0 otherwise.

One Sample Unconditional Probability of Type I error rates

n	Normal	Uniform	t	Laplace	Exponential	Chi.Square	Gamma	Weibull	LogNormal	Contaminated	Pareto
8	0.0524	0.0552	0.0475	0.0385	0.1056	0.0925	0.0721	0.1025	0.1691	0.0427	0.1858
10	0.0518	0.0546	0.0472	0.0445	0.0974	0.086	0.0691	0.0955	0.1624	0.0461	0.1801
15	0.0493	0.0509	0.0466	0.0418	0.0856	0.0734	0.0638	0.0843	0.1423	0.0468	0.1633
20	0.0475	0.0534	0.0468	0.0470	0.0781	0.0719	0.0610	0.0799	0.1299	0.0497	0.1543
25	0.0448	0.0459	0.0487	0.0424	0.0761	0.0654	0.0647	0.0748	0.1215	0.0489	0.1459
30	0.0501	0.0497	0.0491	0.0454	0.0713	0.0633	0.0576	0.0679	0.1183	0.0513	0.1329
50	0.0490	0.0478	0.0476	0.0454	0.0628	0.0571	0.0595	0.0583	0.0984	0.0486	0.1103

Table 1: Showing One Sample Probability of Type I error rates

In Table 1 above, Unconditional Type I error rates are controlled for symmetric distributions(Normal, t, Uniform, Laplace, and Contaminated), but inflated for all asymmetric (skewed) distributions(Exponential, Chi-Squared, Gamma, Weibull, LogNormal, and Pareto) highlighting the effects of nonnormality on the one sample t-test.

Interestingly pretesting for normality did not help either. Conditional Type I error rates are rather more inflated for skewed distributions than unconditional Type I error rates as shown in Table 2 below.

One Sample Conditional Type I error rates

n	Normal	Uniform	t	Contaminated	Laplace	Exponential	Chi.Square	Gamma	Weibull	LogNormal
8	0.040	0.064	0.045	0.036	0.059	0.119	0.115	0.080	0.129	0.277
10	0.044	0.056	0.049	0.053	0.057	0.116	0.111	0.077	0.115	0.284
15	0.049	0.040	0.045	0.050	0.038	0.119	0.105	0.091	0.126	0.351
20	0.048	0.038	0.056	0.051	0.043	0.130	0.105	0.077	0.156	0.421
25	0.048	0.029	0.045	0.045	0.036	0.164	0.108	0.082	0.157	0.499
30	0.050	0.041	0.049	0.042	0.042	0.158	0.119	0.083	0.164	0.537
50	0.047	0.033	0.045	0.047	0.053	0.203	0.128	0.087	0.183	0.727

Table 2: Showing One Sample Conditional Probability of Type I error rates

Two Sample Unconditional Type I error rates

n	Normal	Uniform	t	Contaminated	Laplace	Exponential	Chi.Square	Gamma	Weibull	LogNormal	Pareto
8	0.0496	0.0494	0.0473	0.0467	0.0419	0.0358	0.0386	0.0426	0.0342	0.0270	0.0241
10	0.0490	0.0483	0.0441	0.0462	0.0421	0.0363	0.0409	0.0457	0.0346	0.0294	0.0235
15	0.0484	0.0461	0.0478	0.0502	0.0442	0.0433	0.0461	0.0480	0.0384	0.0310	0.0269
20	0.0504	0.0481	0.0487	0.0482	0.0445	0.0425	0.0476	0.0480	0.0440	0.0351	0.0344
25	0.0513	0.0492	0.0503	0.0461	0.0431	0.0480	0.0460	0.0487	0.0453	0.0376	0.0356
30	0.0530	0.0503	0.0503	0.0501	0.0481	0.0500	0.0473	0.0506	0.0464	0.0408	0.0340
50	0.0479	0.0478	0.0494	0.0473	0.0469	0.0459	0.0514	0.0527	0.0476	0.0415	0.0387

Table 3: Showing Two Sample Probability of Type I error rates

From Table 3 above, probability of Type I error rates is controlled for all distributions for all sample sizes.

We see from Figure 3 below that the probability of Type I error rate is controlled for all symmetric distributions such as Normal, Uniform, t, Laplace, and Contaminated distribution for almost all sample sizes, except for sample sizes less than 10.

Type I error rate is inflated for all asymmetric distributions, especially for LogNormal and Pareto distributions. Unconditional Type I error rates decrease as sample size increases, but conditional Type I error rate for the two-stage procedure increases as sample size increases. This phenomenon was noted in Rochon et al. [2012] due to distortion of the test statistic as a result of the selection mechanism. Based on the results presented so far, it is clear that there are no incentives for performing normality checks using existing normality test methods, rather, there are far more disincentives for doing that as indicated by the pronounced inflation of Type I error rates.

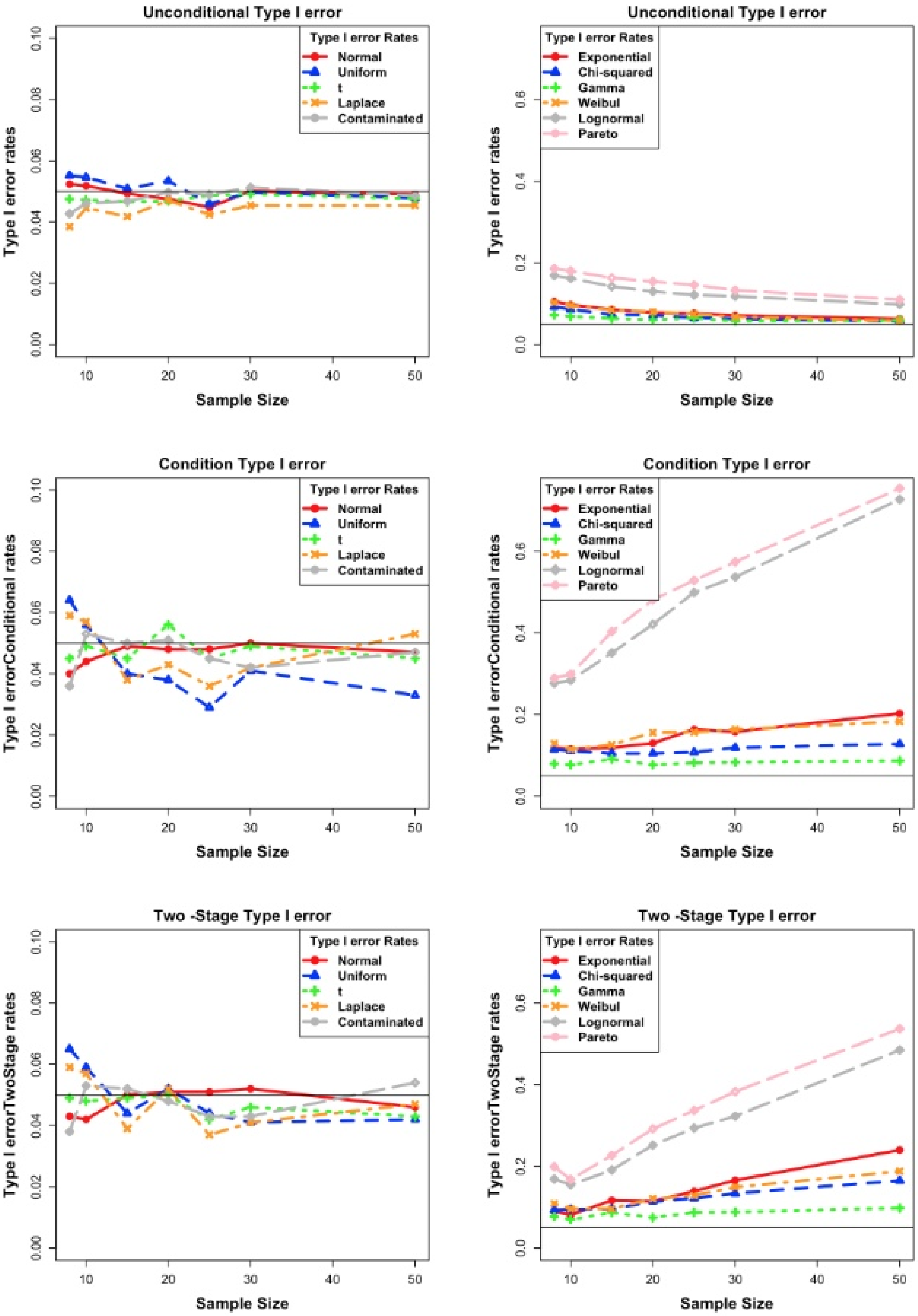

Figure 3: Showing Probability of Type I for different sample sizes for different distributions

Two Sample Case

Figure 4 below indicates that the two-sample t-test procedure is robust to departures from normality. Probability of both unconditional and two-stage procedure Type I error rate is controlled for almost all distributions at virtually all sample sizes.

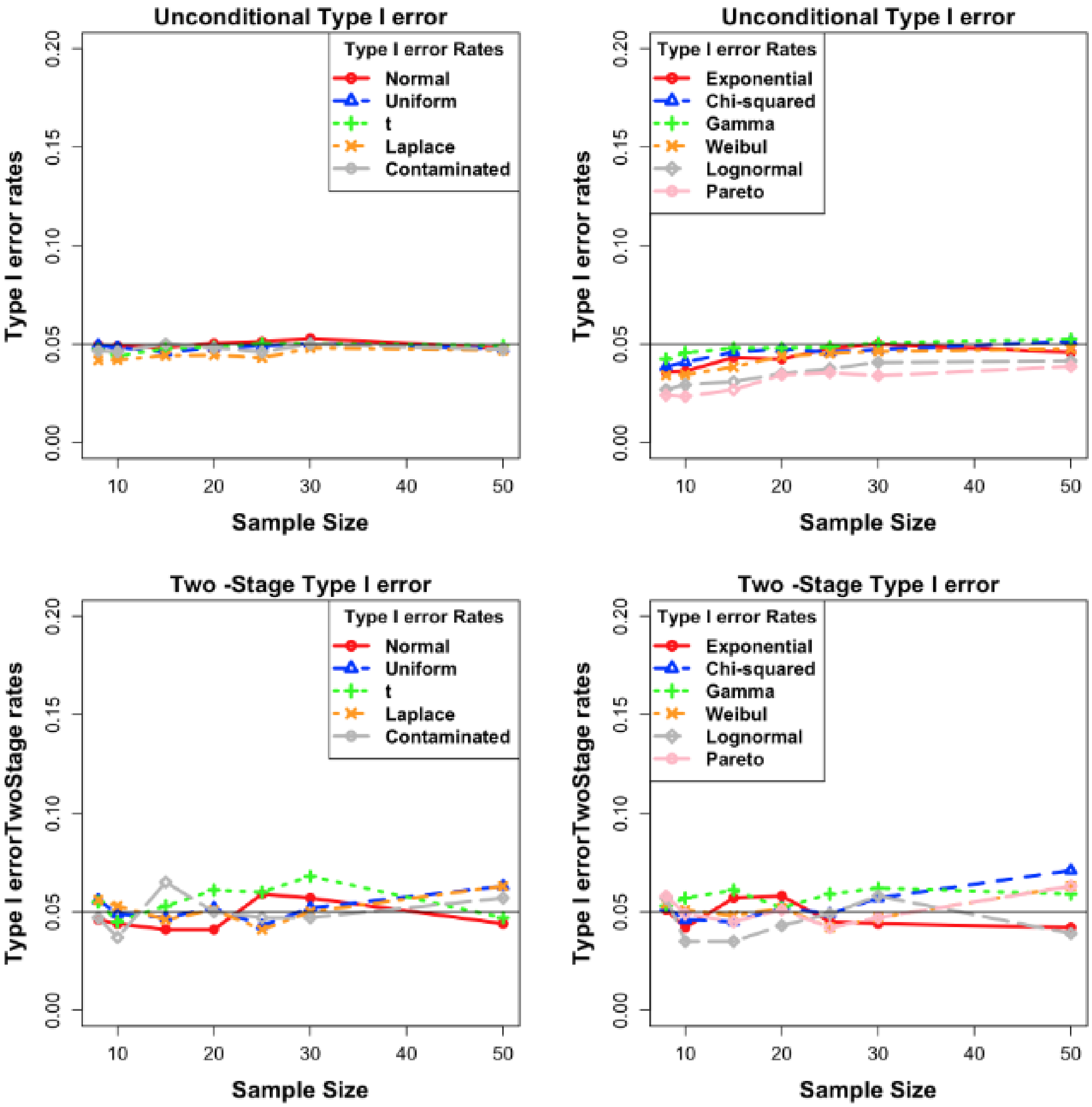

Figure 4: Showing Probability of Type I for different sample sizes for different distributions

Theoretical versus Empirical Test Statistic Distribution

0.0.1 One Sample Case

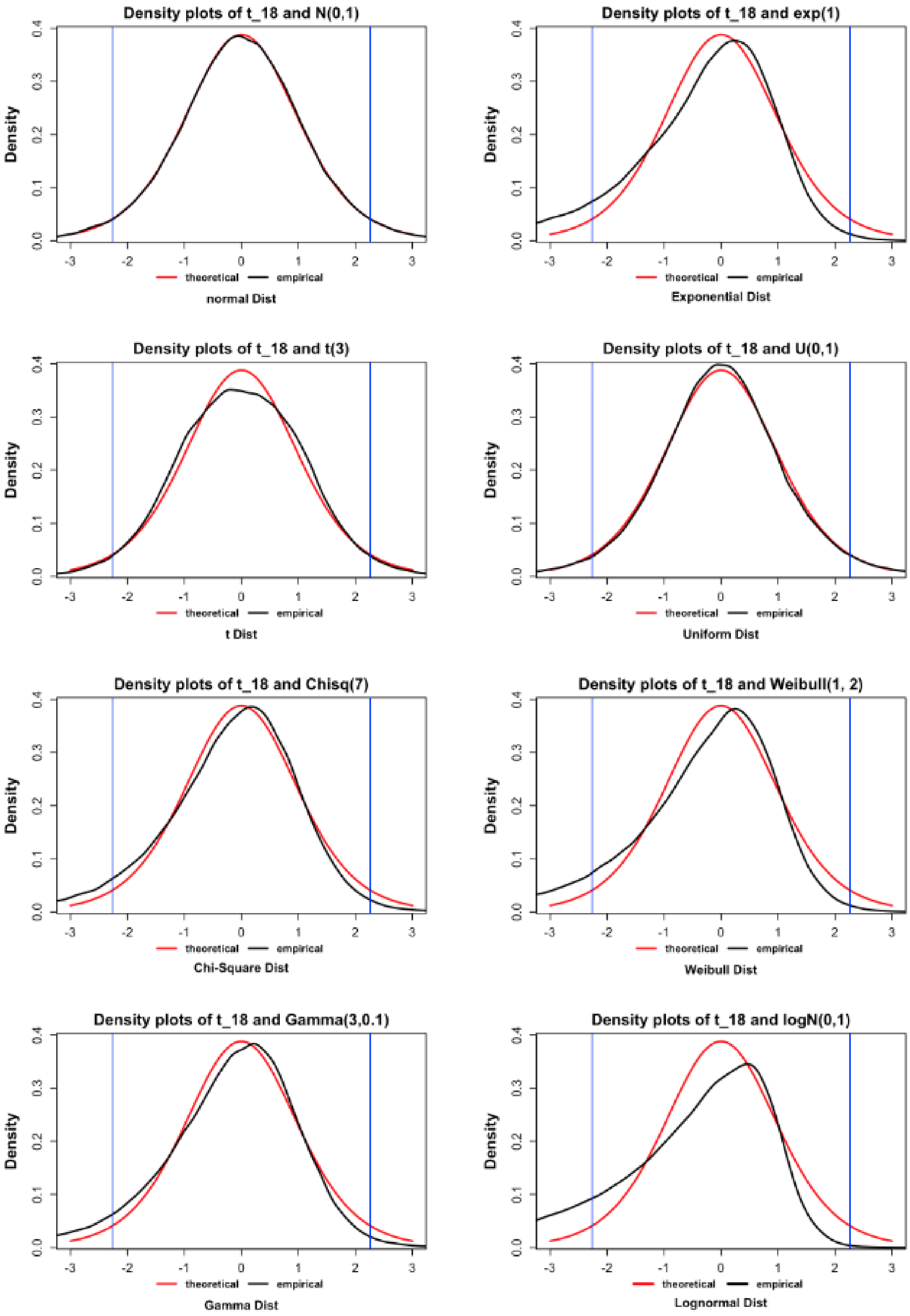

Figure 5: Showing the density plots of the theoretical versus empirical test statistics for different distributions

0.0.2 Two Sample Case

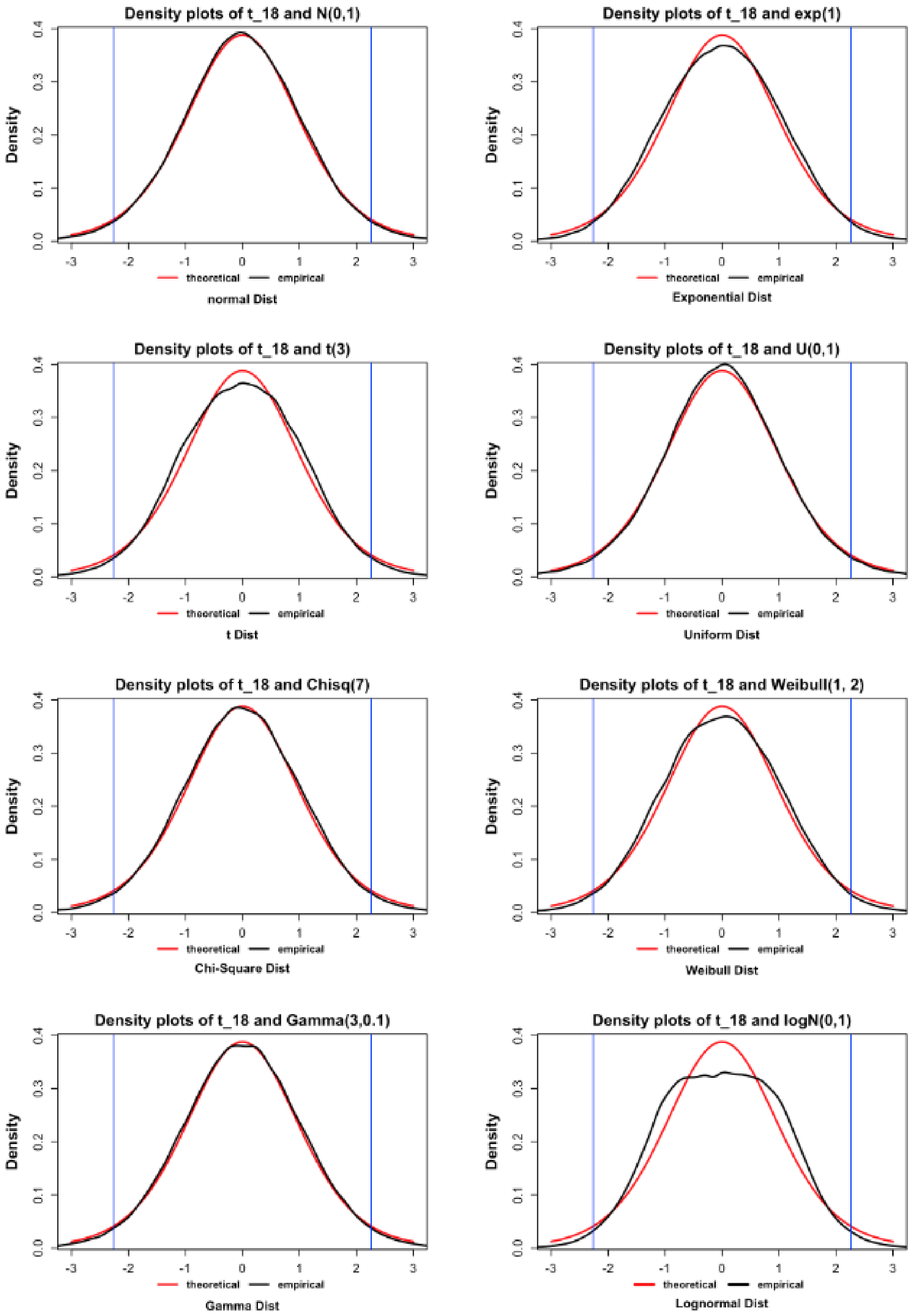

Figure 6: Showing the density plots of the theoretical versus empirical test statistics for different distributions

Probability of Type I error estimated from 95% Quantile

	Probability of Type I error	
distribution	One sample	Two sample
Normal	0.04992	0.05042
t	0.03901	0.04316
Uniform	0.05369	0.05149
Exponetial	0.09853	0.04309
Chi-squared	0.06451	0.04714
Weilbull	0.09844	0.04191
Gamma	0.06776	0.04814
Lognormal	0.16088	0.03451

Table 4: Showing Estimated Probability of Type I error rates from the density plots for one sample and two samples test

Observations

The distribution of the test statistic deviates significantly from the t-distribution for samples drawn from an asymmetric distribution like Chi-Squared, Weibull, Gamma, and Lognormal for one-sample test. The empirical density plots of the test statistic tend to have fatter tails than the theoretical t-density plot.

The opposite can be seen in the two-sample case. There are no significant deviations of the density of the empirical test statistic from the density of the theoretical test statistic. Rather, the empirical density tends to have a slightly slimmer tail than the theoretical test statistic.

The result is consistent with the findings above, further highlighting why the existing preliminary normality test methods are not very useful.

Expected Power loss

In this section, we seek to ascertain the usefulness of the normality test by calculating the expected power loss and the expected Type I error rates on the downstream test.

Simulation Setup for Power loss

The simulation study aims to compare the power of the traditional two-sample t-test and the permutation test under different distributional assumptions. The following steps describe the methodology employed in the simulation. We assume the permutation test is the right test with no distributional assumptions.

Data Generation

Let X_i and Y_i denote two independent samples drawn from the same distribution F. The null hypothesis H_0 is that the two samples come from the same distribution F, i.e.,

$$H_0 : X_i \sim F \quad \text{and} \quad Y_i \sim F$$

For each distribution $d \in \mathcal{D} = \{$Standard Normal, Uniform, t, Contaminated Normal, Laplace, Exponential, Chi-Square, Gamma, Weibull, LogNormal, Pareto$\}$, and for the sample size $n \in \mathbf{n} = \{5, 10, 15, 20, 25, 30\}$, random samples $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ are generated from the distribution d with effect size Δ set to 0.5.

t test

The traditional t-test is performed on the pair of samples if the Shapiro-Wilk normality test is nonsignificant where the t-test statistic is defined as:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means, and s_X^2 and s_Y^2 are the sample variances.

Permutation Test

For the same samples X and Y, we perform a permutation test. The test statistics is the difference in means scaled by the standard error as below:

$$T_{obs} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

We then perform B permutations to generate the distribution of the test statistic under the null hypothesis. For each permutation b, we shuffle the combined sample $Z = \{X, Y\}$ and split it into Z_X and Z_Y. The permuted test statistic T_{perm}^b is calculated as:

$$T_{perm}^b = \frac{\bar{X^b} - \bar{Y^b}}{\sqrt{\frac{(S_1^b)^2}{n_1} + \frac{(S_2^b)^2}{n_2}}}$$

The p-value for the permutation test is the proportion of permuted test statistics that are at least as extreme as the observed test statistic:

$$p_{perm} = \frac{1}{B} \sum_{b=1}^{B} I(|T_{perm}^b| \geq |T_{obs}|)$$

Power Calculation

The power of each test is calculated as the proportion of simulations where the p-value is less than the significance level α :

$$\text{Power}_{\ t} = \frac{1}{N}\sum_{i=1}^{N} I(p^i_{test} < \alpha)$$

$$\text{Power}_{\ perm} = \frac{1}{N}\sum_{i=1}^{N} I(p^i_{perm} < \alpha)$$

Expected Power Loss

The expected power loss due to non-normality is computed as:

$$\text{Expected Power Loss} = (\text{Power}_{\text{perm}} - \text{Power}_t) \times \text{Prob}_{\text{SW not significant}}$$

where $\text{Power}_{\text{perm}}$ and Power_t are the powers of the permutation test and t-test, respectively, and $\text{Prob}_{\text{SW not significant}}$ is the probability that the Shapiro-Wilk test does not reject the null hypothesis of normality.

One Sample Case

The setup for one sample permutation is slightly different. Given a sample $X = \{x_1, x_2, \ldots, x_n\}$, we calculate the observed test statistic T_{obs}:

$$T_{\text{obs}} = \frac{\bar{x}\cdot\sqrt{n}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $\bar{x}$ is the sample mean, n is the sample size, and the denominator is the sample standard deviation.

For each permutation, generate a permuted dataset $X^{(b)}$ by randomly assigning signs to the data points in X:

$$X^{(b)} = \{x_1 \cdot s_1, x_2 \cdot s_2, \ldots, x_n \cdot s_n\}$$

where s_b is a randomly selected sign ($+1$ or -1) for the bth observation, and $b = 1, 2, \ldots, P$, with P being the number of permutations.

Calculate the test statistic for each permuted sample $X^{(i)}$:

$$T^{(b)}_{\text{perm}} = \frac{\bar{x}^{(b)}\cdot\sqrt{n}}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x^{(b)}_j - \bar{x}^{(b)})^2}}$$

The p-value for the permutation test is the proportion of permuted test statistics that are at least as extreme as the observed test statistic:

$$p_{perm} = \frac{1}{B}\sum_{b=1}^{B} I(|T_{perm}^{b}|\geq |T_{obs}|)$$

For the one-sample case, only symmetric distributions are considered.

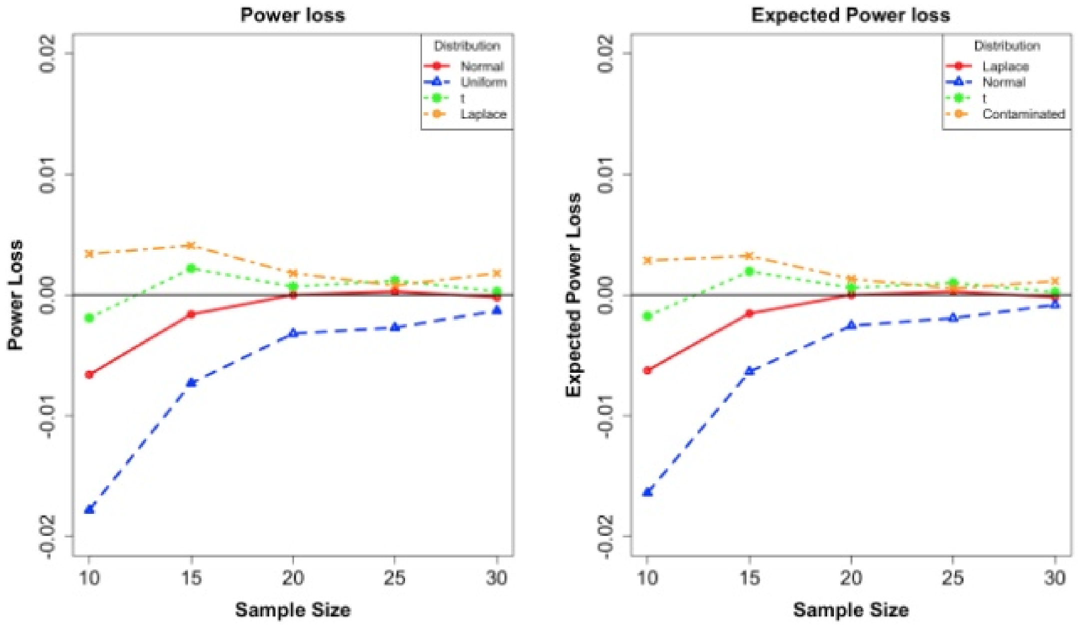

Figure 7: Showing power loss, and expected power loss for one sample test

For the one-sample case, there is little or no power loss and expected power loss. Power loss and expected power loss decrease to zero as the sample size increases.

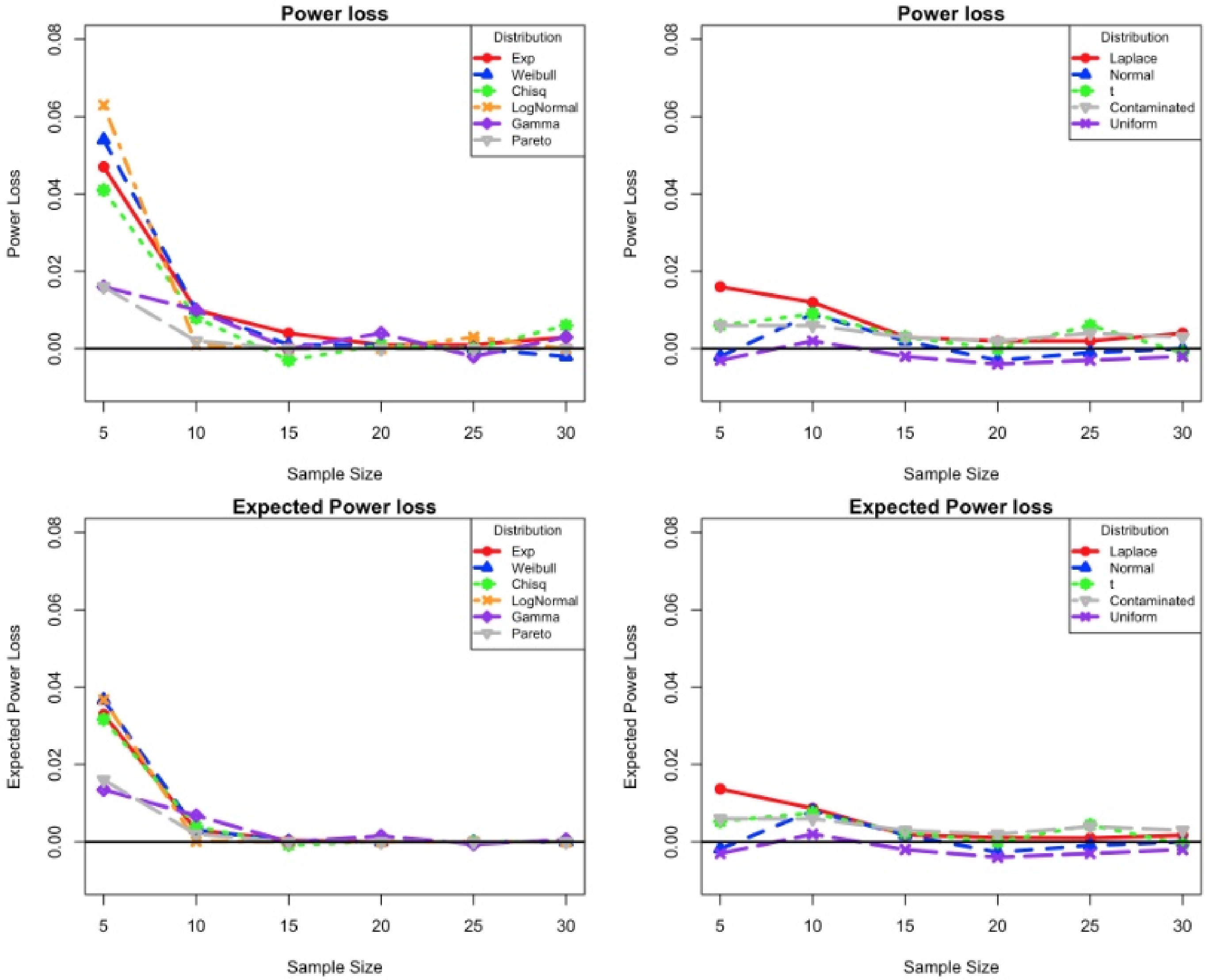

Figure 8: Showing power loss, and expected power loss for two sample tests

Power loss and expected power loss are high for smaller sample sizes, especially for ($n < 20$), but decrease to zero afterward.

General Observations

The above observation again supports our claim that the normality test is not very useful, especially in helping detect departures from normality, improving the power of downstream tests. They are powerful at larger sample sizes, 30 and above which is not very helpful, by CLT.

Area Under Curve

This section seeks to find a unified measure of the overall power and overall Type I error rate for four downstream test methods. The purpose is to provide a tool for comparing the performance of different downstream test methods in terms of their overall power and overall Type I error rates. We shall do that by estimating the area under the power curve and the area under the Type I error rates using the Trapezium Rule in basic calculus. The following downstream methods are considered: the traditional t-test, Wilcoxon, the two-stage method, and the permutation test.

Simulation Setup

Let X_i and Y_i denote two independent samples drawn from the same distribution F. The null hypothesis H_0 is that the two samples come from the same distribution F, i.e.,

$$H_0 : X_i \sim F \quad \text{and} \quad Y_i \sim F$$

For each distribution $d \in \mathcal{D} = \{\text{Standard Normal, Exponential, Chi-Square, LogNormal}\}$, and for the sample size $n \in \mathbf{n} = \{5, 10, 15, 20, 25, 30\}$, random samples $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ are generated from the distribution d with effect size Δ set to 0.5.

t test

The traditional t-test is performed on the pair of samples if the Shapiro-Wilk normality test is nonsignificant where the t-test statistic is defined as:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means, and s_X^2 and s_Y^2 are the sample variances.

Wilcoxon Rank-Sum Test

The Wilcoxon Rank-Sum test is applied to compare two independent samples X and Y, where Y is adjusted by adding an effect size d. The test evaluates the null hypothesis that the samples come from the same distribution.

Given two independent samples $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$, the Wilcoxon Rank-Sum test statistic W is computed as follows:

$$W = \sum_{i=1}^{n} R(X_i) + R(Y_i + d),$$

where $R(\cdot)$ represents the rank of the observations in the combined sample $Z = \{X, Y + d\}$, and the p-value is calculated as

$$p_{wilcox} = P(W \leq W_{\text{obs}}),$$

where W_{obs} is the observed value of the test statistic W.

Two-Stage Procedure:

For each sample X and Y, we perform the Shapiro-Wilk test for normality. We calculate and obtain the p-values $p_{SW}(X)$ and $p_{SW}(Y)$. If the Shapiro-Wilk test is nonsignificant for both samples, X and Y, we perform a t-test, otherwise, we perform the Mann-Whitney U test. The p-value p_{test} for the two-stage procedure is obtained as:

$$p_{t/Wilcox} = \begin{cases} \text{t-test}(X, Y) & \text{if } p_{\text{SW}}(X) > \alpha \text{ and } p_{\text{SW}}(Y) > \alpha \\ \text{Wilcoxon}(X, Y) & \text{otherwise} \end{cases}$$

Permutation Test

For the same samples X and Y, we perform a permutation test. The test statistics is the difference in means scaled by the standard error as below:

$$T_{obs} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

We then perform B permutations to generate the distribution of the test statistic under the null hypothesis. For each permutation b, we shuffle the combined sample $Z = \{X, Y\}$ and split it into Z_X and Z_Y. The permuted test statistic T_{perm}^b is calculated as:

$$T_{perm}^b = \frac{\bar{X^b} - \bar{Y^b}}{\sqrt{\frac{(S_1^b)^2}{n_1} + \frac{(S_2^b)^2}{n_2}}}$$

The p-value for the permutation test is the proportion of permuted test statistics that are at least as extreme as the observed test statistic:

$$p_{perm} = \frac{1}{B} \sum_{b=1}^{B} I(|T_{perm}^b| \geq |T_{obs}|)$$

Power Calculation

The power of each test is calculated as the proportion of simulations where the p-value is less than the significance level α :

$$\text{Power }_t = \frac{1}{N} \sum_{i=1}^{N} I(p_{test}^i < \alpha)$$

$$\text{Power }_{Wilcox} = \frac{1}{N} \sum_{i=1}^{N} I(p_{Wilcox}^i < \alpha)$$

$$\text{Power }_{t/Wilcox} = \frac{1}{N} \sum_{i=1}^{N} I(p_{t/wilcox}^i < \alpha)$$

$$\text{Power }_{perm} = \frac{1}{N} \sum_{i=1}^{N} I(p_{perm}^i < \alpha)$$

Area Under the Power Curve

In general, an area under a curve on finite support can be approximated using the Trapezoid rule as follows:

Let $\{x_i\}$ be a partition of $[n_0, n_k]$ such that $n_0 = x_0 < x_1 < \ldots < x_P = n_k$ where n_0 and n_k are the smallest and latest sample sizes respectively. Let Δx_i be the length of the i^{th} subinterval(i.e $\Delta x_i = x_i - x_{i-1}$), then

$$\text{Area } = \int_{n_0}^{n_k} f(x)\,dx \approx \sum_{i=1}^{k} \frac{f(x_{i-1}) + f(x_i)}{2} \Delta x_i.$$

Thus the area under each power curve was calculated using the trapezoid rule:

$$\text{AUC} = \frac{1}{\max(n) - \min(n)} \sum_{i=1}^{k-1} \left((n_{i+1} - n_i) \left(\frac{P_i + P_{i+1}}{2} \right) \right)$$

where $\frac{1}{\max(n)-\min(n)}$ is a scaling factor, n_i are the sample sizes and P_i are the corresponding power values.

Power One Sample Case

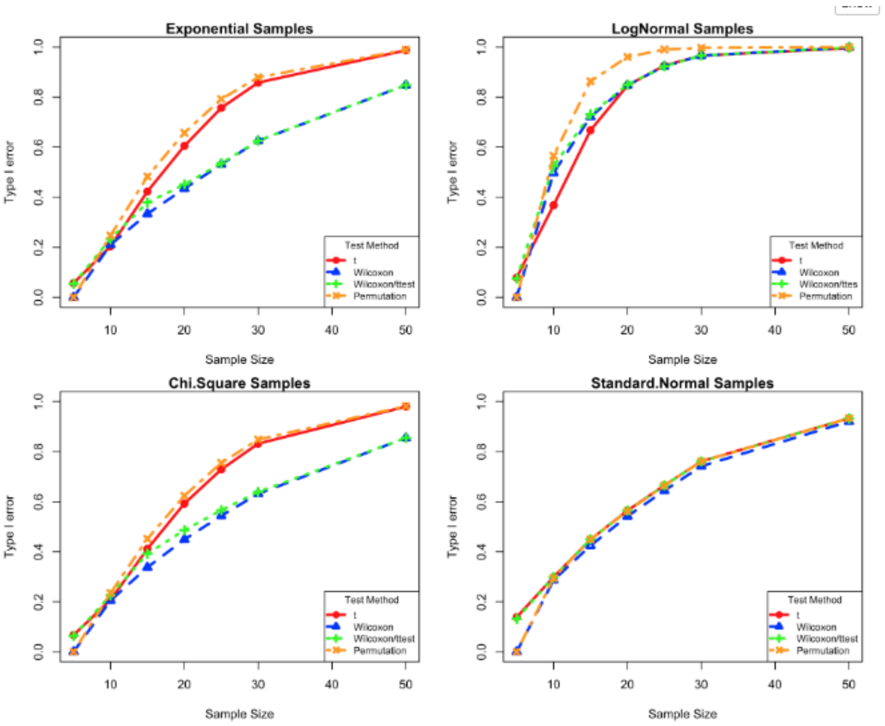

Figure 9: Showing power comparison for one sample test of t-test, Wilcoxon, t/Wilcoxon, and Permutation test

Test Method	Normal	Exponential	Chi-Squared	LogNormal
t	0.6461	0.6819	0.6682	0.8060
Wilcoxon	0.6213	0.5301	0.5360	0.8223
t/Wilcoxon	0.6454	0.5432	0.5565	0.8309
Permutation	0.6375	0.7062	0.6833	0.8747

Table 5: Area Under Power Curve for Two Sample Case

Power Two Sample Case

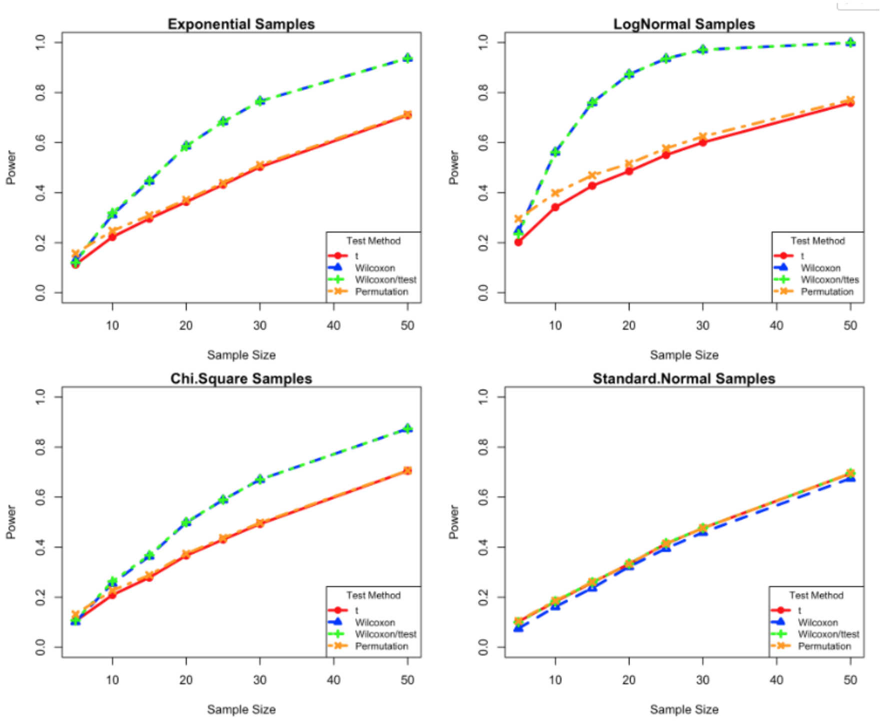

Figure 10: Showing power comparison for two sample test of t-test, Wilcoxon, t/Wilcoxon, and Permutation test

Test Method	Normal	Exponential	Chi-Squared	LogNormal
t	0.4242	0.4495	0.4422	0.5473
Wilcoxon	0.4057	0.6531	0.5755	0.8532
t/Wilcoxon	0.4250	0.6538	0.5774	0.8525
Permutation	0.4252	0.4608	0.4498	0.5790

Table 6: Area Under Power Curve for Two Sample Case

Type I error rates One Sample Case

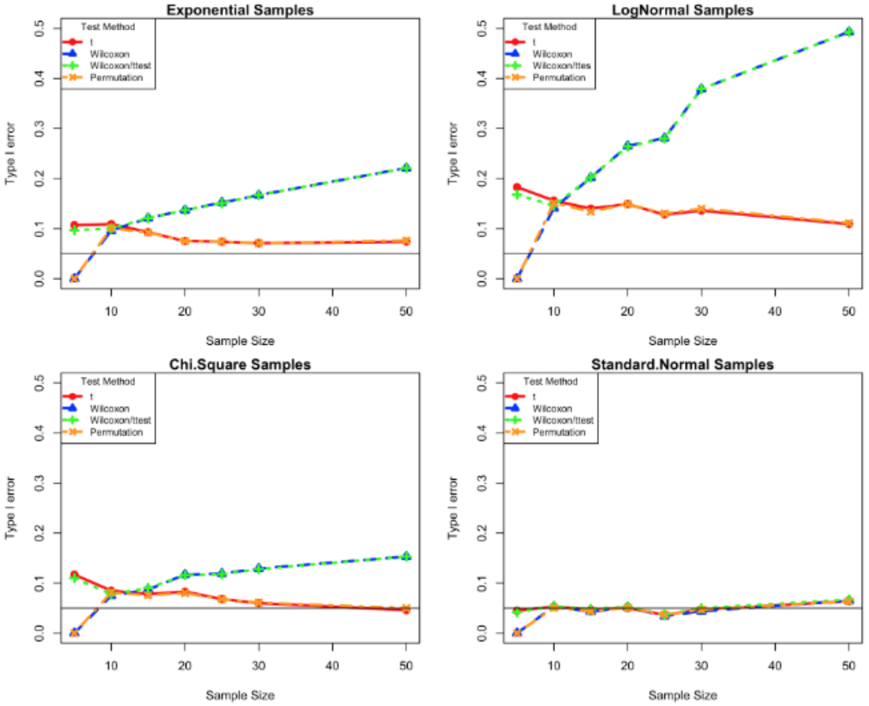

Figure 11: Showing Type I error rates comparison for one sample test of t-test, Wilcoxon, t/Wilcoxon, and Permutation test

Test Method	Normal	Exponential	Chi-Squared	LogNormal
t	0.0005	0.0280	0.0165	0.0773
Wilcoxon	-0.0028	0.0916	0.0577	0.2373
t/Wilcoxon	0.0022	0.0967	0.0633	0.2460
Permutation	-0.0022	0.0222	0.0108	0.0686

Table 7: Area Under Type I error rate Curve for One Sample Case

Type I error rates Two Sample Case

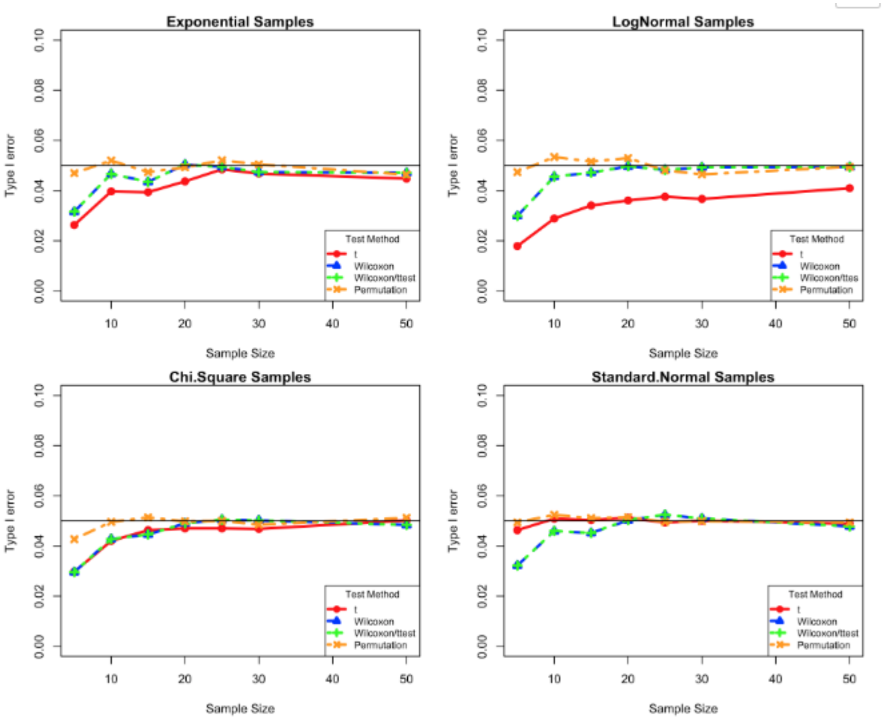

Figure 12: Showing Type I error rates comparison for two sample test of t-test, Wilcoxon, t/Wilcoxon, and Permutation test

Test Method	Normal	Exponential	Chi-Squared	LogNormal
t	-0.0002	-0.0059	-0.0035	-0.0131
Wilcoxon	-0.0017	-0.0031	-0.0026	-0.0022
t/Wilcoxon	-0.0017	-0.0031	-0.0026	-0.0022
Permutation	0.0003	-0.0007	-0.0004	-0.0005

Table 8: Area Under Type I error rate Curve for Two Sample Case

Proposed Method

Develop a machine learning-based approach that can classify whether a dataset follows a normal distribution, potentially improving detection accuracy and robustness.

Instead of directly testing the raw data, we extract features that capture the characteristics of the distribution: skewness, Excess kurtosis, Anderson-Darling Statistic, Jarque-Bera Statistic, Tail Index, Entropy, and Gini Coefficient and use them as predictors to train a machine algorithm to predict sample dataset as either normally distributed or not.

Mathematical Formulation of Methods

This section presents the mathematical formulations of the methods used in the analysis. The features are derived from random samples generated from different statistical distributions, which are then used for classification using machine learning models.

Random Sample Generation

Random samples were generated from both normal and non-normal distributions. The normal samples were generated from the standard normal distribution and a normal distribution with a mean of 25 and a variance of 12. The non-normal samples were generated from the exponential, chi-square, lognormal, and gamma distributions. The probability density functions (PDFs) for these distributions are as follows:

Feature Calculation

For each generated sample, the following features were computed:

- **Skewness:** Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable around its mean. It is defined as:

$$\text{Skewness} = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$$

where μ is the mean, and σ is the standard deviation of the distribution.

- **Kurtosis:** Kurtosis is a measure of the ”tailedness” of the probability distribution. It is given by:

$$\text{Kurtosis} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$$

where μ is the mean, and σ is the standard deviation.

- **Jarque-Bera Statistic:** The Jarque-Bera test statistic is used to test whether a sample has the skewness and kurtosis matching a normal distribution. It is defined as:

$$\text{JB} = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

where S is the sample skewness, K is the sample kurtosis, and n is the sample size.

- **Anderson-Darling Statistic:** The Anderson-Darling test is used to test if a sample comes from a specified distribution. The test statistic is given by:

$$\text{AD} = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln(F(X_i)) + \ln(1-F(X_{n+1-i}))\right]$$

where $F(X_i)$ is the cumulative distribution function of the sample data.

- **Zero-Crossing Rate:** The zero-crossing rate is the rate at which the signal changes sign. For a discrete signal, it can be calculated as:

$$\text{ZCR} = \frac{1}{n-1}\sum_{i=1}^{n-1} \mathbb{I}[(X_i > 0) \neq (X_{i+1} > 0)]$$

  where $\mathbb{I}[\cdot]$ is the indicator function.

- **Gini Coefficient:** The Gini coefficient is a measure of statistical dispersion intended to represent the income inequality or distribution. It is defined as:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|}{2n^2\overline{X}}$$

  where $\overline{X}$ is the mean of the data.

Machine Learning Models

The features were used as input variables for three machine learning models:

- **Logistic Regression:** Logistic regression models the probability of the binary dependent variable as a function of the independent variables. The model is given by:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  where $P(Y = 1)$ is the probability of the outcome being 1, and $X_1, \ldots, X_p$ are the predictors.

- **Random Forest:** Random forest is an ensemble learning method for classification that constructs multiple decision trees during training and outputs the mode of the classes. The prediction is given by:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_B(x)\}$$

  where $T_b(x)$ is the prediction of the b-th tree, and B is the total number of trees.

- **Artificial Neural Network (ANN):** The ANN model used in this study is a feed-forward neural network. The prediction is obtained by:

$$\hat{y} = g\left(\sum_{j=1}^{p} w_j x_j + b\right)$$

  where $g(\cdot)$ is an activation function (e.g., sigmoid, ReLU), w_j are the weights, and b is the bias term.

Evaluation Metrics

The models were evaluated using accuracy, confusion matrices, and ROC curves. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, and the area under the curve (AUC) is used as a summary measure of performance.

- **True Positive Rate (TPR):**

  $$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  where TP is the number of true positives and FN is the number of false negatives.

- **False Positive Rate (FPR):**

  $$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

  where FP is the number of false positives and TN is the number of true negatives.

The ROC curve is generated by plotting TPR against FPR, and the area under the ROC curve (AUC) provides a single measure of the model's ability to discriminate between the classes.

References

Khaled A Al-Omar and Elhadi Degawa. An evaluation of normality tests for statistical analysis. *International Journal of Information and Education Technology*, 6(6):429–432, 2016.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons, 1991.

Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486–489, 2012.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

Nik AK Razali and Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1): 21–33, 2011.

Justine Rochon, Matthias Gondan, and Meinhard Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC medical research methodology*, 12:1–11, 2012.

Dag J Steinskog, Dag Tjøstheim, and Nils Gunnar Kvamstø. A cautionary note on the use of the kolmogorov-smirnov test for normality. *Monthly Weather Review*, 135(3):1151–1157, 2007.

Henry C Thode. *Testing for normality*. CRC Press, 2002.

Bee Wah Yap and Chen Yen Sim. Comparative study of various tests for normality. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, 2011.