This paper explains the research conducted by Minitab statisticians to develop the methods and data checks used in the Assistant in Minitab Statistical Software.

1-Sample t-Test

Overview

The 1-sample t-test is used to estimate mean of your process and to compare the mean with a target value. This test is considered a robust procedure because it is extremely insensitive to the normality assumption when the sample is moderately large. According to most statistical textbooks, the 1-sample t-test and the t-confidence interval for the mean are appropriate for any sample of size 30 or more.

In this paper, we describe the simulations we conducted to evaluate this general rule of a minimum of 30 sample units. Our simulations focused on the impact of nonnormality on the 1-sample t-test. We also wanted to assess the impact of unusual data on the test results.

Based on our research, the Assistant automatically performs the following checks on your data and displays the results in the Report Card:

- Unusual data
- Normality (Is the sample large enough for normality not to be an issue?)
- Sample size

For general information about the methodology for the 1-sample t-test, refer to Arnold (1990), Casella and Berger (1990), Moore and McCabe (1993), and Srivastava (1958).

Note The results in this paper also apply to the Paired t-test in the Assistant because the Paired t-test applies the method for the 1-sample t-test to a sample of paired differences.

Data checks

Unusual data

Unusual data are extremely large or small data values, also known as outliers. Unusual data can have a strong influence on the results of the analysis. When the sample is small, they can affect the chances of finding statistically significant results. Unusual data can indicate problems with data collection, or may be due to unusual behavior of the process you are studying. These data points are often worth investigating and should be corrected when possible.

Objective

We wanted to develop a method to check for data values that are very large or very small relative to the overall sample, which may affect the results of the analysis.

Method

We developed a method to check for unusual data based on the method described by Hoaglin, Iglewicz, and Tukey (1986) to identify outliers in boxplots.

Results

The Assistant identifies a data point as unusual if it is more than 1.5 times the interquartile range beyond the lower or upper quartile of the distribution. The lower and upper quartiles are the 25th and 75th percentiles of the data. The interquartile range is the difference between the two quartiles. This method works well even when there are multiple outliers because it makes it possible to detect each specific outlier.

When checking for unusual data, the Assistant Report Card for the 1-sample t-test displays the following status indicators:

Status	Condition
✓	There are no unusual data points.
⚠	At least one data point is unusual and may affect the test results.

Normality

The 1-sample t-test is derived under the assumption that the population is normally distributed. Fortunately, even when data are not normally distributed, this method works well when the sample size is large enough.

Objective

We wanted to determine the effect of nonnormality on the Type I error and the Type II error of the test to provide guidelines on the sample size and normality.

Method

We conducted simulations to determine the sample size for which the normality assumption can be ignored when performing a 1-sample t-test or calculating a t-confidence interval for the mean of a population.

We designed the first study to assess the effect of nonnormality on the Type I error rate of the test. Specifically, we wanted to deduce the minimum sample size needed for the test to be insensitive to the distribution of the population. We performed the 1-sample t-test on small, moderate, and large samples generated from normal and nonnormal populations. The nonnormal populations included mildly to severely skewed populations, symmetric light- and heavy-tailed populations, and contaminated normal populations. The normal population served as a control population for comparison. For each case, we calculated and compared the simulated significance levels with the target, or nominal, significance level of 0.05. If the test performs well, the simulated significance levels should be close to 0.05. We examined the simulated significance levels across all the different conditions to assess the minimum sample size for which they remain close to the target level regardless of the distribution. See Appendix A for more details.

In the second study, we examined the effect of nonnormality on the Type II error of the test. The design of the simulation is identical to the first study. However, we compared simulated power levels under different conditions to target power levels calculated using the theoretical power function of the 1-sample t-test. See Appendix B for more details.

Results

The effect of nonnormality on both the Type I and Type II error rates of the test is minimal for sample sizes as small as 20. However, when the parent population of the sample is extremely skewed, larger samples may be required. We recommend a sample size of about 40 for those cases. For more details, see Appendix A and Appendix B.

Because the test performs well with relatively small samples, the Assistant does not test the data for normality. Instead, it checks the size of the sample and displays the following status indicators in the Report Card:

Status	Condition
✓	The sample size is at least 20; normality is not an issue.
⚠	The sample size is less than 20; normality may be an issue.

Sample Size

Typically, a hypothesis test is performed to gather evidence to reject the null hypothesis of "no difference". If the samples are too small, the power of the test may not be adequate to detect a difference between the means that actually exists, which results in a Type II error. It is therefore crucial to ensure that the sample size is sufficiently large to detect practically important differences with high probability.

Objective

If the data does not provide sufficient evidence against the null hypothesis, we want to determine if the sample size was large enough for the test to detect practical differences of interest with high probability. Although the objective of sample size planning is to ensure that the sample size is large enough to detect important differences with high probability, they should not be so large that meaningless differences become statistically significant with high probability.

Method

The power and sample size analysis is based upon the theoretical power function of the specific test used to conduct the statistical analysis. As discussed earlier, the power function of the 1-sample t-test is insensitive to the normal assumption when the sample size is at least 20. The power function depends upon the sample size, the difference between the target mean and the population mean, and the variance of the population. For more details, see Appendix B.

Results

When the data does not provide enough evidence against the null hypothesis, the Assistant calculates practical differences that can be detected with an 80% and a 90% probability for the given sample sizes. In addition, if the user provides a particular practical difference of interest, the Assistant calculates the sample size that yields an 80% and a 90% chance of detecting the difference.

There is no general result to report because the results depend on the user's specific sample. However, you can refer to Appendix B for more information about power for the 1-sample t-test.

When checking for power and sample size, the Assistant Report Card for the 1-sample t-test displays the following status indicators:

Status	Condition
✓	The test finds a difference between the mean and target, so power is not an issue. OR Power is sufficient. The test did not find a difference between the mean and target, but the sample is large enough to provide at least a 90% chance of detecting the given difference.
⚠	Power may be sufficient. The test did not find a difference between the mean and target, but the sample is large enough to provide an 80% to 90% chance of detecting the given difference. The sample size required to achieve 90% power is reported.
⚠	Power might not be sufficient. The test did not find a difference between the mean and target, and the sample is large enough to provide a 60% to 80% chance of detecting the given difference. The sample sizes required to achieve 80% power and 90% power are reported.
✖	The power is not sufficient. The test did not find a difference between the mean and target, and the sample is not large enough to provide at least a 60% chance of detecting the given difference. The sample sizes required to achieve 80% power and 90% power are reported.
ℹ	The test did not find a difference between the mean and target. You did not specify a practical difference between the mean and target to detect; therefore, the report indicates the differences that you could detect with 80% and 90% chance, based on your sample sizes, standard deviations, and alpha.

References

Arnold, S.F. (1990). *Mathematical statistics*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Casella, G., & Berger, R. L. (1990). *Statistical inference*. Pacific Grove, CA: Wadsworth, Inc.

Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*, *991-999*.

Moore, D.S. & McCabe, G.P. (1993). *Introduction to the practice of statistics*, 2nd ed. New York, NY: W. H. Freeman and Company.

Neyman, J., Iwaszkiewicz, K. & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation, *Journal of the Royal Statistical Society, Series B, 2, 107-180*.

Pearson, E.S., & Hartley, H.O. (Eds.). (1954). *Biometrika tables for statisticians*, Vol. I. London: Cambridge University Press.

Srivastava, A. B. L. (1958). Effect of non-normality on the power function of t-test, *Biometrika*, *45, 421-429*.

Appendix A: Impact of nonnormality on the significance level (validity of the test)

Under the normal assumption, the 1-sample t-test is a *uniformly most powerful* (UMP) *unbiased* size-α test. That is, the test is as powerful as or more powerful than any other unbiased size-α test about the mean. However, when the parent population of the sample is not normally distributed then the above optimality properties hold true if the sample size is large enough. In other words, for sufficiently large samples, the actual significance level of the 1-sample t-test approximately equals the target level for normal as well as nonnormal data, and the power function of the test is also insensitive to the normal assumption (Srivastava, 1958).

We wanted to determine how large a sample must be to be considered as sufficiently large enough for the t-test to be insensitive to the normal assumption. Many text books recommend that if the sample size $n \geq 30$, then the normal assumption can be ignored for most practical purposes (Arnold, 1990; Casella & Berger, 1990; and Moore & McCabe, 1993). The purpose of the investigation described in these appendices is to conduct simulation studies to evaluate this general rule by examining the impact of different nonnormal distributions on the 1-sample t-test.

Simulation study A

We wanted to examine the impact of nonnormality on the Type I error rate of the test to assess a minimum sample size for which it is stable and remains close to the target error rate regardless of the distribution.

To accomplish this, two-sided t-tests using $\alpha = 0.05$ were performed using random samples of various sizes ($n = 10, 15, 20,\ 25,\ 30, 35, 40, 50, 60, 80, 100$) generated from several distributions with different properties. These distributions include:

- The standard normal distribution (N(0,1))
- Symmetric and heavy-tailed distributions, such as the t-distribution with 5 and 10 degrees of freedom (t(5),t(10))
- The Laplace distribution with location 0 and scale 1 (Lpl))
- Skewed and heavy-tailed distributions represented by the exponential distribution with scale 1 (Exp), the Chi-square distributions with 3, 5, and 10 degrees of freedom (Chi(3), Chi(5), Chi(10))

- Symmetric and lighter-tailed distributions such as the uniform distribution (U(0,1)), the Beta distribution with the two parameters set to 3 (B(3,3))
- A left-skewed and heavy-tailed distribution (B (8,1))

In addition, to assess the direct effect of outliers, we generated samples from contaminated normal distributions defined as:

$$\mathrm{CN}(\mathrm{p}, \sigma) = \mathrm{pN}(0,1) + (1 - \mathrm{p})\mathrm{N}(0, \sigma)$$

where p is defined as the mixing parameter and $1 - \mathrm{p}$ is the proportion of contamination or proportion of outliers. We selected two contaminated normal populations for the study: $\mathrm{CN}(0.9,3)$ (10% of the population members are outliers) and $\mathrm{CN}(.8,3)$ (20% of the population members are outliers). These two distributions are symmetric and have long tails due to the outliers.

For each sample size, 10,000 sample replicates were drawn from each population and a 1-sample t-test with null hypothesis $\mu = \mu_o$ and alternative hypothesis $\mu \neq \mu_o$was performed for each of the 10,000 samples. For each test, we set the hypothesized mean μ_o to the true mean of the parent population of the sample. As a result, for a given sample size, the fraction of the 10,000 sample replicates that yield a rejection of the null hypothesis represents the simulated Type I error rate or significance level of the test. Because the target significance level is 5%, the simulation error is about 0.2%.

The simulation results are shown in Tables 1 and 2 and graphically displayed in Figures 1 and 2.

Results and summary

The results (see Table 1 and Figure 1) show that when samples are generated from symmetric populations, the simulated significance levels of the test are close to the target significance level even when the sample sizes are small. However, the test results are slightly conservative for heavy-tailed symmetric distributions when the samples are small, including small samples that are generated from the contaminated normal distributions. It also appears that outliers reduce the significance level of the test when the samples are small. However, this effect is reversed when small samples are generated from symmetric lighter-tailed parent populations (the Beta (3,3) and uniform distributions). The simulated significance levels are slightly higher for these cases.

Table 1 Simulated significance levels for two-sided 1-sample t-test for samples generated from symmetric populations. The target significance level is $\alpha = 0.05$.

Dist.	N(0,1)	t(5)	t(10)	Lpl	CN(.9,3)	CN(.8,3)	B(3,3)	U(0,1)
N	Symmetric and heavy tails						Symmetric and lighter tails	
10	0.050	0.046	0.048	0.044	0.043	0.039	0.057	0.057
15	0.051	0.050	0.049	0.049	0.043	0.043	0.053	0.054
20	0.047	0.051	0.051	0.047	0.043	0.044	0.051	0.052
25	0.050	0.047	0.050	0.046	0.046	0.046	0.048	0.050
30	0.053	0.050	0.048	0.043	0.049	0.046	0.050	0.048
35	0.052	0.047	0.049	0.050	0.047	0.045	0.051	0.054
40	0.046	0.052	0.054	0.048	0.046	0.049	0.044	0.050
50	0.050	0.049	0.051	0.048	0.047	0.051	0.053	0.050
60	0.052	0.049	0.053	0.050	0.051	0.056	0.054	0.052
80	0.049	0.050	0.051	0.047	0.047	0.052	0.049	0.049
100	0.050	0.052	0.049	0.051	0.052	0.054	0.051	0.054

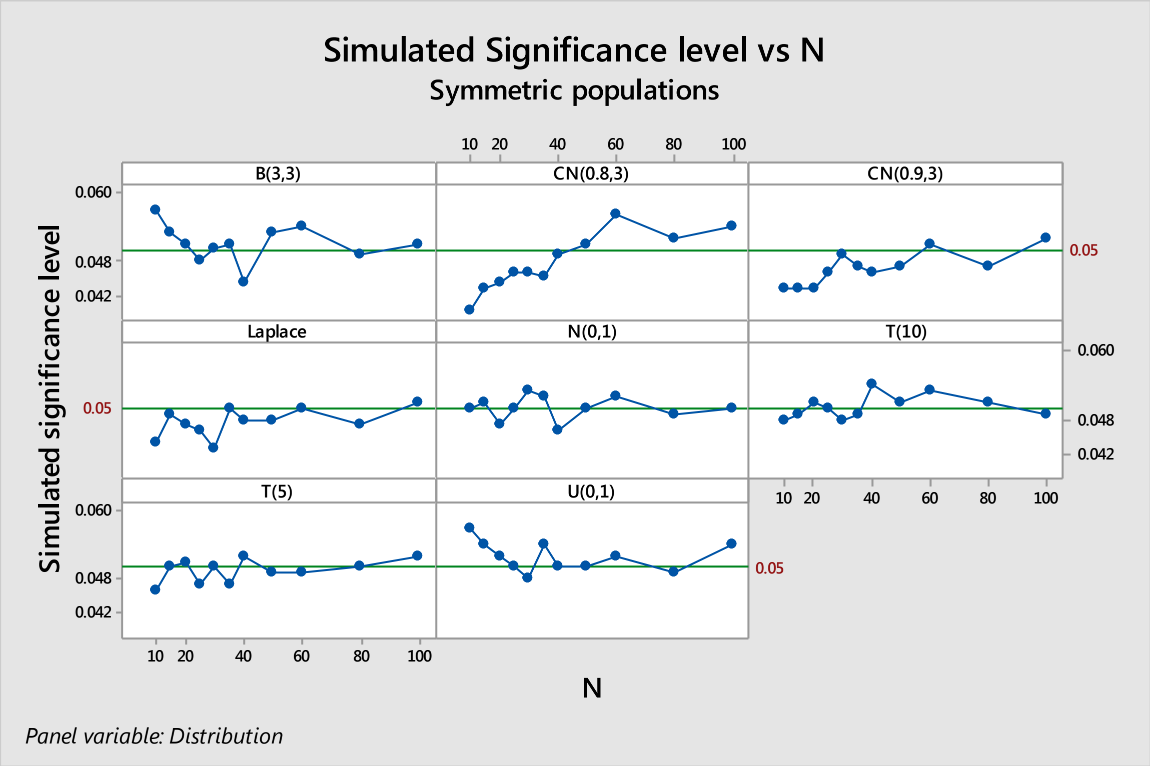

Figure 1 Plot of simulated significance levels for two-sided 1-sample t-tests versus size of samples generated from symmetric populations. The target significance level is $\alpha = 0.05$.

On the other hand, when samples are generated from skewed distributions, the performance of the test depends upon the severity of the skewness. The results in Table 2 and Figure 2 show that the 1-sample t-test is sensitive to skewness in small samples. For severely skewed populations (exponential, Chi(3), and Beta(8,1)), larger samples are required for the simulated significance levels to be near the target significance level. However, for moderately skewed populations (Chi(5) and Chi(10)), a minimum sample size of 20 is sufficient for the simulated significance levels to be close to the target level. With a sample size of 20, the simulated significance level is approximately 0.063 for the Chi-square distribution with 5 degrees of freedom and is about 0.056 for the Chi-square distribution with 10 degrees of freedom.

Table 2 Simulated significance levels for two-sided 1-sample t-test for samples generated from skewed populations. The target significance level is $\alpha = 0.05$.

N	Exp	Chi(3)	B(8,1)	Chi(5)	Chi(10)
	Population Skewness				
	2.0	1.633	-1.423	1.265	0.894
	Simulated Significance Levels				
10	0.101	0.089	0.087	0.069	0.060
15	0.088	0.076	0.072	0.068	0.057
20	0.083	0.073	0.069	0.063	0.056
25	0.075	0.068	0.067	0.067	0.056
30	0.069	0.067	0.066	0.058	0.054
35	0.075	0.067	0.062	0.062	0.056
40	0.067	0.067	0.061	0.059	0.056
50	0.064	0.057	0.062	0.057	0.054
60	0.063	0.056	0.061	0.054	0.055
80	0.059	0.058	0.053	0.052	0.052
100	0.060	0.055	0.055	0.047	0.053

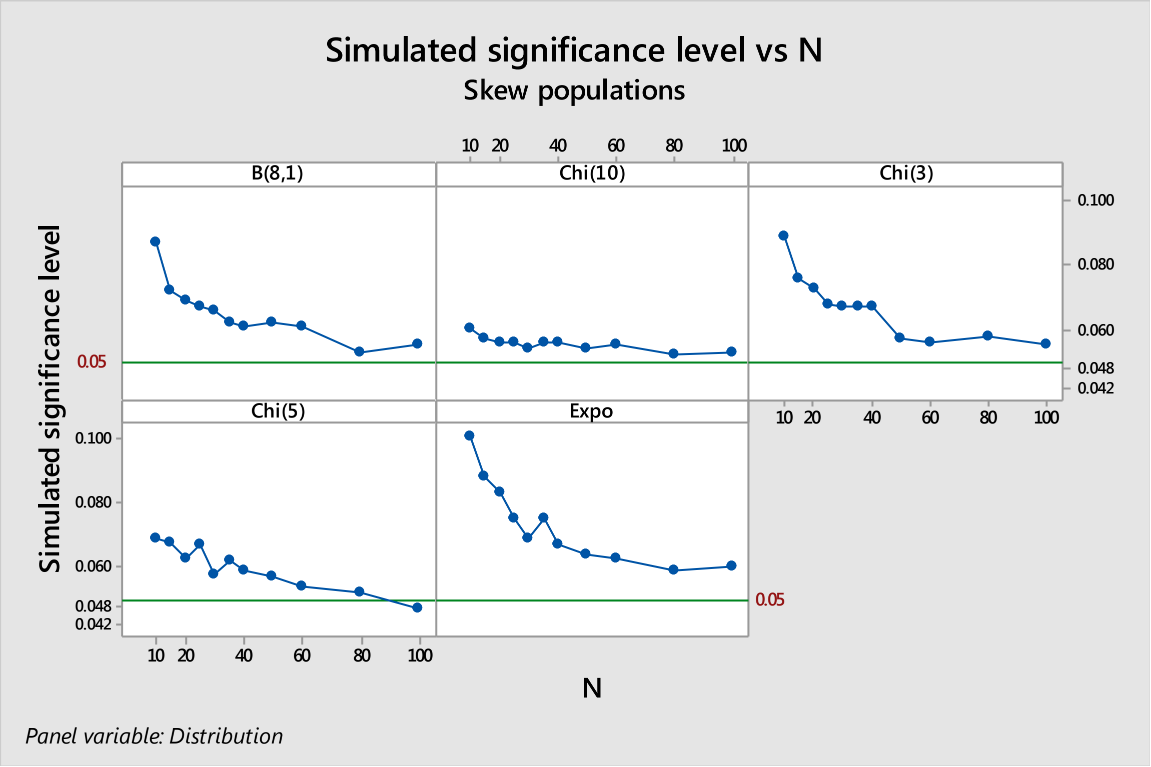

Figure 2 Plot of simulated significance levels for two-sided 1-sample t-tests versus the size of samples generated from skewed populations. The target significance level is $\alpha = 0.05$.

In this investigation, we focused on the hypothesis tests rather than the confidence intervals. However, the results naturally extend to confidence intervals because hypothesis tests and confidence intervals can both be used to determine statistical significance.

Appendix B: Sample size and power of the test

We wanted to examine the sensitivity of the power function to the normal assumption under which it is derived. Note that if β is the Type II error of a test, then $1 - \beta$ is the power of the test. As a result, the planned sample size is determined so that the Type II error rate is small or, equivalently, the power level is reasonably high.

The power functions for t-tests are well known and documented. Pearson and Hartley (1954) and Neyman, Iwaszkiewicz, and Kolodziejczyk (1935) provide charts and tables of the power functions.

For a size α two-sided 1-sample t-test, a mathematical expression of this function may be given in terms of the sample size and the difference δ between the true mean μ and the hypothesized mean μ_o as

$$\pi(n, \delta) = 1 - F_{n-1,\lambda}\left(t_{n-1}^{\alpha/2}\right) + F_{n-1,\lambda}\left(-t_{n-1}^{\alpha/2}\right)$$

where $\mathrm{F}_{\mathrm{d},\lambda}(.)$ is the C.D.F of the non-central t distribution with $\mathrm{d} = \mathrm{n} - 1$ degrees of freedom and non-centrality parameter

$$\lambda = \frac{\delta\sqrt{n}}{\sigma}$$

and where $\mathrm{t}_{\mathrm{d}}^{\alpha}$ denotes the 100α upper percentile point of the t-distribution with d degrees of freedom.

For one-sided alternatives, the power is given as

$$\pi(n, \delta) = 1 - F_{n-1,\lambda}(t_{n-1}^{\alpha})$$

for testing the null hypothesis against $\mu > \mu_o$ and is given as

$$\pi(n, \delta) = F_{n-1,\lambda}(-t_{n-1}^{\alpha})$$

when testing the null hypothesis against $\mu < \mu_o$.

These functions are derived under the supposition that the data is normally distributed and that the significance level of the test is fixed at some value α.

Simulation study B

We designed this simulation to evaluate the effect of nonnormality on the theoretical power function of the 1-sample t-test. To assess the effect of nonnormality, we compared simulated power levels with the target power levels that were calculated using the theoretical power function of the test.

We performed two-sided t-tests at $\alpha = 0.05$ on random samples of various sizes ($n = 10, 15, 20, 25, 30, 35, 40, 50, 60, 80, 100$) generated from the same populations described in the first simulation study (see Appendix A).

For each of the populations, the null hypothesis of the test is $\mu = \mu_o - \delta$ and its alternative hypothesis is $\neq \mu_o - \delta$, where μ_o is set at the population true mean and $\delta = \sigma/2$ (σ is the standard deviation of the parent population). Thus, the difference between the true mean and the hypothesized mean is 0, so the correct decision is to reject the null hypothesis.

For each given sample size, 10,000 sample replicates are drawn from each of the distributions. For each given sample size, the fraction of the 10,000 replicates for which the null hypothesis is rejected represents the simulated power level of the test at the given sample size and difference δ. Note that we chose this particular difference value because it produces power values that are relatively small when the sample sizes are small.

In addition, the corresponding theoretical power values referred to as target power values are calculated at the difference δ and the various sample sizes for comparison with the simulated power values.

The simulation results are in Table 3 and Table 4 and graphically displayed in Figure 3 and Figure 4.

Results and summary

The results confirm that the power of the 1-sample t-test is in general insensitive to the normal assumption when the sample size is large enough. For samples generated from symmetric populations, the results in Table 3 show that the target power and simulated power levels are close even for small samples. The corresponding power curves displayed in Figure 3 are practically indistinguishable. For samples generated from the contaminated normal distributions, the power values are somewhat conservative for small to moderate sample sizes. This may be because the actual significance level of the test for those populations is slightly higher than the fixed target significance level α.

Table 3 Simulated power levels at a difference $\delta = \sigma/2$ for a size $\alpha = 0.05$ two-sided 1-sample t-test when samples are generated from symmetric populations. The simulated power levels are compared to the theoretical target power levels derived under the normality assumption.

n	Target power	N(0,1)	t(5)	t(10)	Lpl	CN(.9,3)	CN(.8,3)	B(3,3)	U(0,1)
		Simulated Power level at $\boldsymbol{\delta = \sigma/2}$ (Symmetric Populations)							
10	0.293	0.299	0.334	0.311	0.357	0.361	0.385	0.280	0.269
15	0.438	0.438	0.480	0.450	0.491	0.512	0.511	0.423	0.421
20	0.565	0.570	0.603	0.578	0.600	0.629	0.623	0.557	0.548
25	0.670	0.674	0.695	0.680	0.691	0.712	0.700	0.665	0.670
30	0.754	0.756	0.770	0.756	0.767	0.768	0.765	0.754	0.750
35	0.820	0.819	0.827	0.815	0.820	0.819	0.812	0.822	0.818
40	0.869	0.870	0.871	0.868	0.862	0.869	0.868	0.875	0.867
50	0.934	0.933	0.929	0.930	0.929	0.923	0.925	0.932	0.940
60	0.968	0.967	0.963	0.965	0.964	0.955	0.955	0.968	0.971
80	0.993	0.993	0.989	0.992	0.991	0.988	0.989	0.994	0.994
100	0.999	0.998	0.996	0.998	0.999	0.998	0.996	0.999	0.999

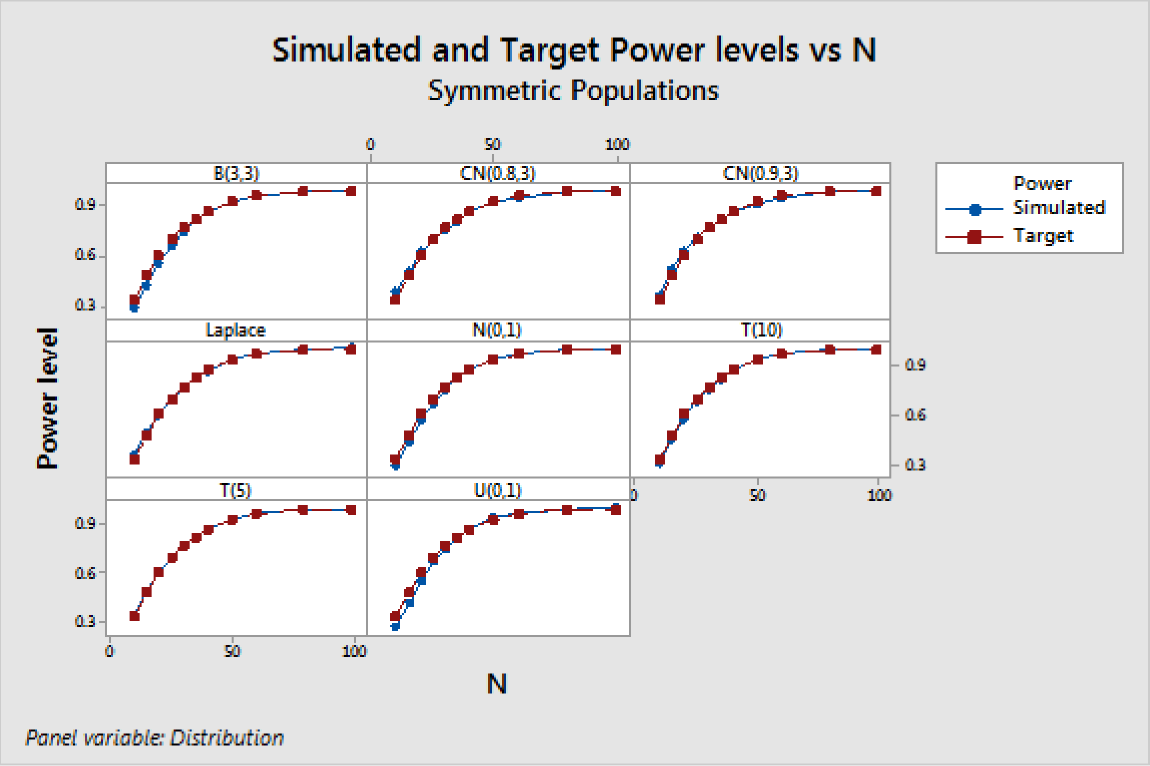

Figure 3 Simulated power curves compared to the theoretical target power curves at $\alpha = 0.05$, two-sided 1-sample t-test when samples are generated from symmetric populations. The power values are evaluated at a difference of $\delta = \sigma/2$.

However, when samples are from skewed populations, the simulated power values are off target for small samples, as shown in Table 4 and Figure 4. For moderately skewed populations, such as the Chi-square distribution with 5 degrees of freedom and the Chi-square distribution with 10 degrees of freedom, when the sample size is at least 20, the target and simulated power levels are close. For example, for $n = 20$, the target power level is 0.565 when the simulated power levels are 0.576 and 0.577 for the Chi-square 5 and Chi-square 10 distributions, respectively. For extremely skewed distributions, larger samples are required for the simulated power levels to approach the target significance level. This may be because the 1-sample t-test does not properly control Type I error when the sample sizes are small and the parent populations are extremely skewed.

Table 4 Simulated power values at a difference $\delta = \sigma/2$ for a size $\alpha = 0.05$ two-sided 1-sample t-test when samples are generated from skewed populations. The simulated power values are compared to the target power values derived under the normality assumption.

N	Target power	Exp		Chi(3)	B(8,1)	Chi(5)	Chi(10)
			Population Skewness				
		2.0		1.633	-1.423	1.265	0.894
			Simulated Power Levels				
10	0.293	0.206		0.212	0.390	0.225	0.238
15	0.438	0.416		0.413	0.484	0.409	0.407
20	0.565	0.604		0.591	0.566	0.576	0.577
25	0.670	0.763		0.734	0.657	0.709	0.695
30	0.754	0.859		0.834	0.729	0.808	0.785
35	0.820	0.917		0.895	0.776	0.874	0.835
40	0.869	0.955		0.935	0.823	0.925	0.905
50	0.934	0.987		0.981	0.900	0.973	0.960
60	0.968	0.997		0.994	0.937	0.991	0.985
80	0.993	1.000		0.999	0.980	0.999	0.997
100	0.999	1.000		1.000	0.994	1.000	1.000

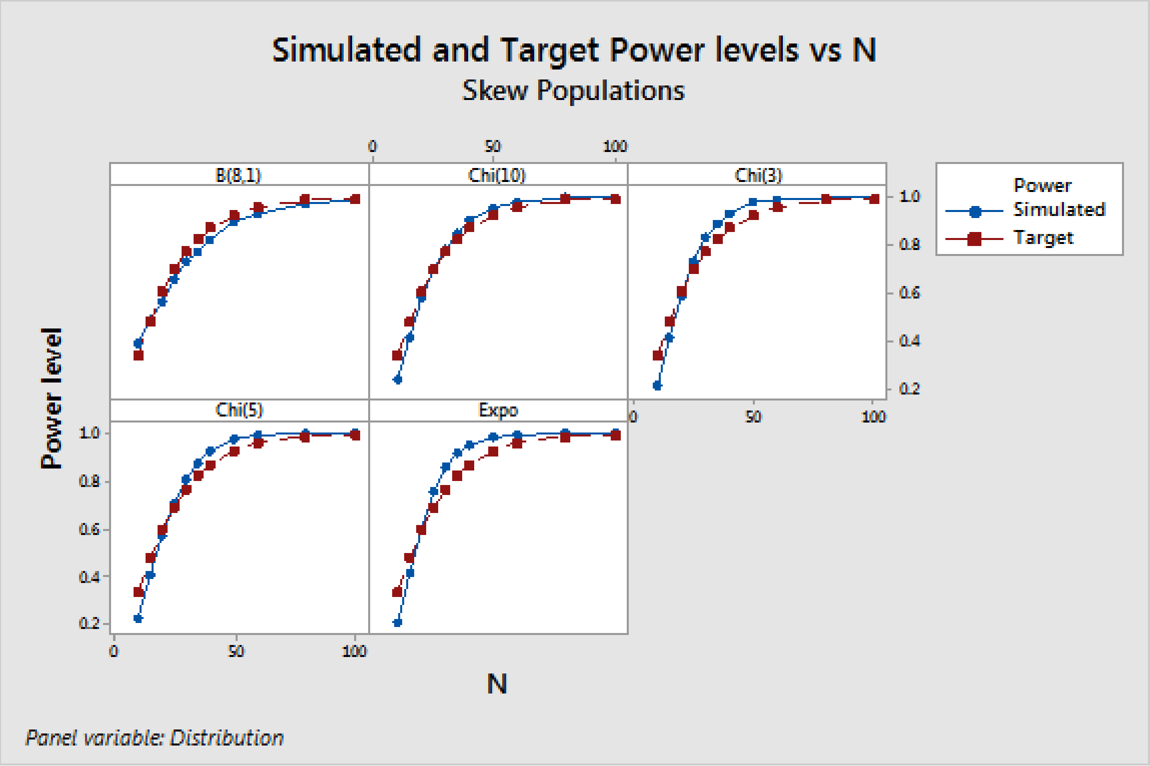

Figure 4 Simulated power curves compared to theoretical target power curves at $\alpha = 0.05$, two-sided 1-sample t-test when samples are generated from symmetric populations. The power values are evaluated at a difference of $\delta = \sigma/2$.

In summary, for moderately skewed distributions, the power function is reliable if the sample size is at least 20, regardless of the parent population from which the sample is drawn. For extremely skewed populations, a larger sample size (about 40) is required for the simulated power to be near the target power.