Normality tests

Shengping Yang PhD, Gilbert Berdine MD

I have completed a clinical trial with one primary outcome and various secondary outcomes. The goal is to compare the randomly assigned two treatments to see if the outcomes differ. I am planning to perform a two-sample t-test for each of the outcomes, and one of the assumptions for a t-test is that data are normally distributed. I am wondering what the appropriate approaches are to check the normality assumption.

The normal distribution is the most important distribution in statistical data analysis. Most statistical tests are developed based on the assumption that the outcome variable/model residual is normally distributed. There are both graphical and numerical methods for evaluating data normality (we will focus on univariate normality in this article).

1. Graphical evaluation

(a) Histogram

A histogram is an approximate representation of a numerical data distribution. It is a common practice to create a histogram to visually exam the data distribution and potential outliers before performing a formal statistical test. A bell-shaped histogram is often an indication that data distribution is approximately normal if sample size is sufficiently large; otherwise, severe skewness (a measure of symmetry in a distribution), and/or higher kurtosis (a measure of the "tailedness" of the distribution of random variables), as well as outliers are indications of violation of normality. As an example, we randomly generated 200 values from a normal distribution, and the histogram (Figure 1; left panel) is approximately bell-shaped. As a comparison, a histogram for a distribution skewed to the right is also presented in Figure 1 (right panel). Note that, if the sample size is small, e.g., less than 10, then such an assessment might not be informative.

Corresponding author: **Shengping Yang**
Contact Information: Shengping.Yang@pbrc.edu
DOI: 10.12746/swrccc.v9i37.805

(b) The normal Q-Q plot

A Q-Q plot is a scatterplot created by plotting two sets of quantiles—one is the data quantile, and the other is the theoretical distribution quantile-against one another. If the two sets of quantiles came from the same distribution, then the points on the plot roughly form a straight line. For example, a normal Q-Q plot represents the correlation between the data and normal quantiles and measures how well the data are modeled by a normal distribution. Therefore, if the data are sampled from a normal distribution, then the data quantile should be highly positively correlated with the theoretical normal distribution quantile, and the plotted points should fall approximately on a straight line.

Very often a "fat pencil test" can be performed to make an informal evaluation on whether the data point line is straight. Imagine placing a "fat pencil" on top of the data line—if the "fat pencil" covers all the points on the plot, then we could conclude that the data are approximately normally distributed. Otherwise, if there are points that are visible beyond the edges of the "fat pencil", then the data are probably not normally distributed. Figure 2 is a normal Q-Q plot for 200 randomly generated values from a normal distribution; it is easy to see that there would be no visible points if a "fat pencil" is place on the top of the data line. Note that the "fat pencil test" is fast and intuitive, but it is not a substitute for a formal statistical test.

2. Statistical testing of normality

Several tests can be used for testing data normality; the Shapiro-Wilk *W* test is considered the most powerful one in most situations.

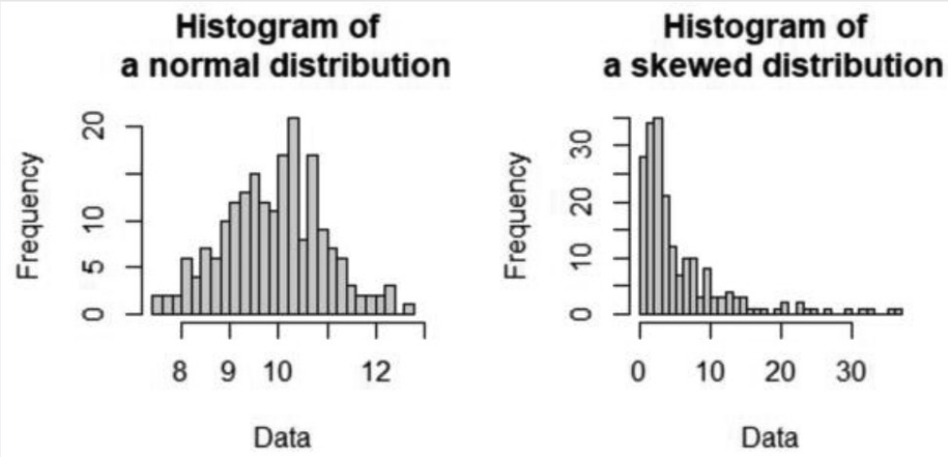

Figure 1. An example histogram.

(a) The Shapiro-Wilk W *test*

The Shapiro-Wilk test tests whether the outcome data, a random sample from the entire population, came from a normally distributed population. In other words, the Shapiro-Wilk test evaluates how likely it is that the values in the sample are observed, if the outcome variable is normally distributed in the entire population. In fact, the test statistic *W* is roughly a measure of the straightness of the normal Q-Q plot. More specifically, *W* is the ratio of two estimates of the variance of a normal distribution given data, that is,

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

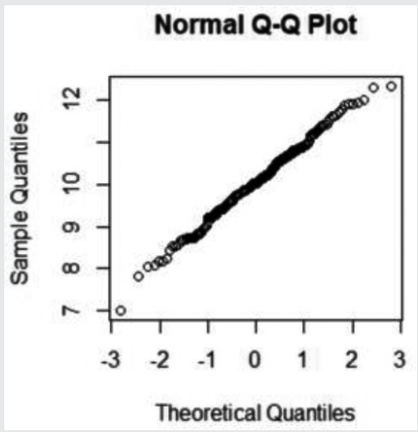

Figure 2. An example Q-Q plot.

where $x_{(i)}$ is the i^{th} order statistic of the sample, $\bar{x}$ is the sample mean, $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C}$, where $C = \|V^{-1}m\|, m = (m_1, \ldots, m_n)^T$ is the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and *V* is the covariance matrix of those normal order statistics.

The null and alternative hypotheses are:

H_0: The population is normally distributed.
H_a: The population is not normally distributed.

If the p value from the test is less than a prespecified significance level, then the null hypothesis is rejected, and we can conclude that there is evidence that the data are not normally distributed. Otherwise, the null hypothesis cannot be rejected. Like many other tests, sample size is a concern when performing a Shapiro-Wilk test. On the one hand, if the sample size is small, then the test has low power. On the other hand, if the sample size is large, then the Shapiro-Wilk test might detect a deviation from normality that might be too small and not meaningful (we will explain why a small deviation from normality is acceptable for most statistical tests that require normality, if sample size is large, later in the article). Therefore, it is preferable to take both the numerical and the graphical (e.g., a normal Q-Q plot) evaluations into account before reaching a conclusion. Note that although the Shapiro-Wilk test is a powerful test, it does not work well if there are many identical values in the data.

(b) Other tests

i. The Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov test is a widely used non-parametric test for comparing two samples and can also be used to quantify the distance between an empirical distribution function of the sample and the cumulative distribution function of a reference distribution. In general, the Kolmogorov–Smirnov test is less powerful for testing normality than the Shapiro–Wilk test.

ii. The Anderson-Darling test

The Anderson-Darling test is a modification of the Kolmogorov-Smirnov test and gives more weight to the tails than does the Kolmogorov–Smirnov test. Therefore, this test is more powerful than the Kolmogorov-Smirnov test and can be as powerful as the Shapiro-Wilk test under certain situations.

Other tests that can be used for testing normality include the Martinez-Iglewicz test, the D'Agostino's *K*-squared test, and the Ryan-Joiner normality test, etc.

3. Deviation from the normality assumption

Many statistical tests that require a normality assumption are robust to deviations from the normality. For example, the *t*-test is valid if the sample mean(s) is normally distributed (and sample variance follows a scaled χ^2 distribution), even if the sample data are not normally distributed. This is because, by central limit theorem, the distribution of sample mean is normally distributed for moderately large samples even if the data are not normally distributed. Similarly, in linear regression, it has been shown that the test statistic for testing the regression slope will converge in probability to the standard normal distribution. Therefore, as we have mentioned earlier, if the sample size is large, most of the statistical tests that require a normality assumption are still valid, even if the data distribution deviates from normal. This is consistent with our understanding that a normality test p value is not the only factor to consider to determine whether a statistical test is appropriate, even if from the perspective of the normality assumption.

4. Data that are not normally distributed

If the normality test result shows that the normality assumption is violated, very often a data transformation is recommended. Other times, a non-parametric test can be used. For example, a Mann-Whitney U test (also called the Wilcoxon rank-sum test) can be used to compare two independent samples, and a Kruskal-Wallis Rank Sum Test can be used to compare multiple groups. In addition, some data are known to follow a distribution other than normal, for example, binary outcome data, count data, etc. In such cases, a generalized linear model can be used.

In summary, many statistical tests require a normality assumption. Both graphical and numerical methods are available for evaluating data normality. However, for samples with a small number of observations, none of these methods is satisfactory. On the other hand, for samples with a large number of observations, some of the normality tests might be too sensitive. Therefore, it is preferable to take both graphical and numerical evaluation results into account to reach a conclusion. Nevertheless, many statistical tests that require a normality assumption are robust to deviations from normality due to the central limit theorem, and thus a normality test is not the only factor to consider in order to determine whether a statistical test that requires a normality assumption is appropriate.

Keywords: normality, Shapiro-Wilk W test, sample size

Article citation: Yang S, Berdine G. Normality tests. The Southwest Respiratory and Critical Care Chronicles 2021;9(37):87–90
From: Department of Biostatistics (SY), Pennington Biomedical Research Center, Baton Rouge, LA; Department of Internal Medicine (GB), Texas Tech University Health Sciences Center, Lubbock, Texas
Submitted: 1/4/2021
Accepted: 1/7/2021
Conflicts of interest: none

References

1. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari 1933;4:83–91.
2. De Moivre A. The Doctrine of Chances, 3d ed. 1756. London: Millar.
3. Razali N, Wah YB. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling tests. J Statistical Modeling and Analytics 2011;2 (1):21–33.
4. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52(3–4): 591–611.
5. Smirnov N. Table for estimating the goodness of fit of empirical distributions. Annals Mathematical Statistics 1948;19(2): 279–281.