QQPlots and Normality Testing

MA2003B Application of Multivariate Methods in Data Science

Team Members:

Raul Gomez
Adriana Reyna

August 13, 2025

1 Normality Testing

Although QQ plots are a good way to visually assess the normality of a dataset, they can be subjective and imprecise. Therefore, it becomes necessary to develop more objective and precise statistical tests to determine whether a dataset follows a normal distribution. However, before introducing these tests, we will remind the reader the definition of the skewness and kurtosis of a sample.

Definition 1.1. Let $x_1, x_2, \ldots, x_n$ be a sample of size n. The *sample mean* $\bar{x}$ and *sample variance* s^2 are given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The *sample skewness* g_1 given by

$$g_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3.$$

If $g_1 > 0$ the sample is said to be *right-skewed*, if $g_1 < 0$ it is *left-skewed*, and values near zero indicate approximate symmetry. The *sample excess kurtosis* g_2 is given by

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

A value of $g_2 = 0$ corresponds to the kurtosis of the normal distribution (mesokurtic). Positive values ($g_2 > 0$) indicate heavy-tailed, sharply peaked distributions (leptokurtic), while negative values ($g_2 < 0$) indicate light-tailed, flat-topped distributions (platykurtic).

1.1 The Shapiro-Wilk Test

Definition 1.2 (Shapiro-Wilk Test)**.** Let $x_1, x_2, \ldots, x_n$ be a sample of size $n \geq 3$, and let

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

denote the order statistics of the sample. The *Shapiro–Wilk test* is a statistical test for the null hypothesis

$$H_0 : \text{The data are drawn from a normally distributed population.}$$

against the alternative hypothesis

$$H_1 : \text{The data are not drawn from a normally distributed population.}$$

Let $\bar{x}$ be the sample mean, and let $m = (m_1, m_2, \ldots, m_n)^\top$ denote the expected values of the order statistics of a sample of size n from the standard normal distribution. Let V be the covariance matrix of these order statistics. Define the weight vector

$$a = \frac{m^\top V^{-1}}{\sqrt{m^\top V^{-1} V^{-1} m}} = (a_1, a_2, \ldots, a_n)^\top.$$

The *Shapiro–Wilk test statistic* is given by

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

Values of W close to 1 indicate that the sample is close to normal, whereas small values of W indicate departures from normality. The p-value of the test is computed from the sampling distribution of W under H_0. If this p-value is less than the chosen significance level α, H_0 is rejected, providing evidence that the data are not normally distributed.

Remark 1.3.

- The Shapiro–Wilk test is particularly powerful for small sample sizes ($n \leq 50$), though it remains applicable for larger samples (up to about $n \approx 2000$ in many software implementations).
- The test is sensitive to both skewness and kurtosis departures from normality.
- For very large samples, the test may detect trivial deviations from normality as significant.

1.1.1 Test subsubsection

Just a test subsubsection to check the numbering of the section.