

4.4: Normality

How to decide which test to use, parametric or non-parametric, t-test or Wilcoxon? We need to know if the distribution follows or at least approaches normality. This could be checked visually (Figure 4.4.1):

How does QQ plot work? First, data points are ordered and each one is assigned to a quantile. Second, a set of theoretical quantiles—positions that data points should have occupied in a *normal distribution*—is calculated. Finally, theoretical and empirical quantiles are paired off and plotted.

We have overlaid the plot with a line coming through quantiles. When the dots follow the line closely, the empirical distribution is normal. Here a lot of dots at the tails are far. Again, we conclude, that the original distribution is not normal.

R also offers numerical instruments that check for normality. The first among them is Shapiro-Wilk test (please run this code yourself):

Here the output is rather terse. P-values are small, but what was the null hypothesis? Even the built-in help does not state it. To understand, we may run a simple experiment:

The command `norm()` generates random numbers that follow normal distribution, as many of them as stated in the argument. Here we have obtained a p-value approaching unity. Clearly, the null hypothesis was “the empirical distribution is normal”.

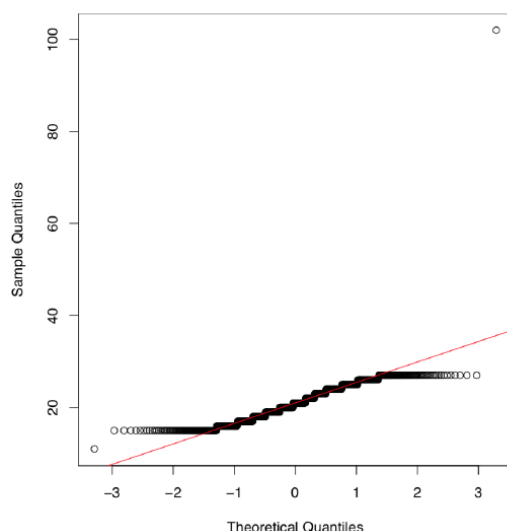


Figure 4.4.1 Graphical check for the normality.

Armed with this little experiment, we may conclude that distributions of both `salary` and `salary2` are not normal.

Kolmogorov-Smirnov test works with two distributions. The null hypothesis is that both samples came from the same population. If we want to test one distribution against normal, second argument should be `pnorm`:

(The result is comparable with the result of Shapiro-Wilk test. We scaled data because by default, the second argument uses scaled normal distribution.)

Function `ks.test()` accepts any type of the second argument and therefore could be used to check how reliable is to approximate current distribution with *any* theoretical distribution, not necessarily normal. However, Kolmogorov-Smirnov test often returns the wrong answer for samples which size is < 50 , so it is less powerful than Shapiro-Wilks test.

`2.2e-16` is so-called *exponential notation*, the way to show really small numbers like this one (2.2×10^{-16}). If this notation is not comfortable to you, there is a way to get rid of it:

(Option `scipen` equals to the maximal allowable number of zeros.)

Most of times these three ways to determine normality are in agreement, but this is not a surprise if they return different results. Normality check is not a death sentence, it is just an opinion based on probability.

Again, if sample size is small, statistical tests and even quantile-quantile plots frequently fail to detect non-normality. In these cases, simpler tools like stem plot or histogram, would provide a better help.

This page titled 4.4: Normality is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by Alexey Shipunov via source content that was edited to the style and standards of the LibreTexts platform.