MAT 2379, Introduction to Biostatistics

Section 9.3. Assessing Normality

To check if the data comes from a normal distribution, we should first look at the histogram and see if it has a bell-shaped form, and if resembles the plot of the normal density function.

In this section, we introduce a more accurate method to see if a data set comes from a normal distribution. This method is called the **normal QQ-plot** (QQ comes from quantile-quantile). We explain it below.

Assume that $X_1, \ldots, X_n$ are independent random variables which have a normal distribution with mean μ and variance σ^2.

By the standardization,

$$X_i = \mu + \sigma Z_i, \tag{1}$$

where $Z_1, \ldots, Z_n$ have a standard normal distribution.

Arrange $X_1, \ldots, X_n$ in increasing order and call the arranged values $Y_1 < \ldots < Y_n$. Arrange $Z_1, \ldots, Z_n$ in increasing order and call the arranged values $W_1 < \ldots < W_n$. It follows that

$$Y_i = \mu + \sigma W_i.$$

Hence,

$$E(Y_i) = \mu + \sigma E(W_i) = \mu + \sigma z_i$$

The values $E(W_i) = z_i, i = 1, \ldots, n$ are called the *normal scores*. They are uniquely determined by the value of n. The normal scores are values between -3 and 3 that are calculated as follows: using Table 17.2 and 17.3, find the point z_i such that

$$P(Z \leq z_i) = p_i = \frac{i - 3/8}{n + 1/4}.$$

The values $p_1, \ldots, p_n$ are called the *relative orders* and are uniquely determined by the value of n.

In theory, we obtain that if X_i's are normally distributed, then the average of Y_i is a linear function of z_i.

In practice, we observe $x_1, \ldots, x_n$ and we arrange them in increasing order as $y_1 < \ldots < y_n$. If the plot of the pairs of points (z_i, y_i) is "almost" linear, we say that the assumption of normality is verified. The slope of this line can be used as an approximation for σ.

Example: The following data gives the weight loss (in lb) of 7 women who participated in a weight loss program: (a negative weight loss is a weight gain)

$$x_1 = 7.06 \quad x_2 = -0.61 \quad x_3 = 3.87 \quad x_4 = 3.73 \quad x_5 = 3.61 \quad x_6 = -2.14 \quad x_7 = -5.28$$

We would like to see if it is reasonable to assume that the data comes from a normal distribution.

Arrange the data in increasing order as $y_1 < \ldots < y_7$ as follows:

$$y_1 = -5.28 \quad y_2 = -2.14 \quad y_3 = -0.61 \quad y_4 = 3.61 \quad y_5 = 3.73 \quad 6_6 = 3.87 \quad y_7 = 7.06$$

For $n = 7$, the normal scores are: (roughly)

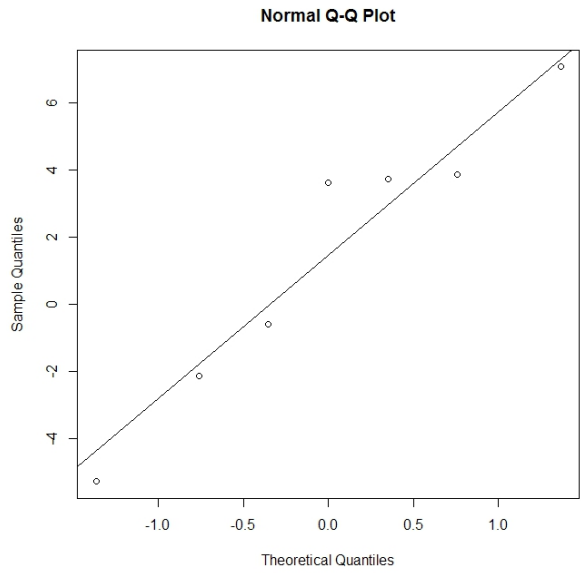
$$z_1 = -1.36 \quad z_2 = -0.76 \quad z_3 = -0.35 \quad z_4 = 0, \quad z_5 = 0.35 \quad z_6 = 0.76 \quad z_7 = 1.36$$

The pairs (y_i, z_i) are given in the following table:

y_i	-5.28	-2.14	-0.61	3.61	3.73	3.87	7.06
z_i	-1.36	-0.76	-0.35	0	0.35	0.76	1.36

Below is the plot of $(z_i, y_i), i = 1, \ldots, 7$, which is called the **normal QQ-plot**, together with the line $y = 1.46 + 4.27z$, where

$$\hat{\mu} = \bar{x} = 1.46 \quad \text{is an estimate of } \mu \quad \text{and} \quad \hat{\sigma} = s = 4.27 \quad \text{is an estimate of } \sigma.$$



The values z_i are called the *theoretical quantiles* and are plotted on the horizontal axis; the values y_i are called the *sample quantiles* and are plotted on the vertical axis. Since the plot is linear, we can say that the data $x_1, \ldots, x_7$ comes from a normal distribution.

Remark: In the book, the plot is done in the opposite order, i.e. the sample quantiles and are plotted on the horizontal axis and the theoretical quantiles and are plotted on the vertical axis. This plot is also called a normal QQ-plot. In this case, since equation (1) can be written also as $Z_i = (-\mu/\sigma) + (1/\sigma)X_i$, one has to plot the line $z = (-\hat{\mu}/\hat{\sigma}) + (1/\hat{\sigma})y$.