Probability: Random Variables

Normal Quantile Plots

The normal distributions provide good models for some distributions of real data. Examples include test scores, heights of people, and errors in measurement. Distributions of some other variables are usually skewed and therefore non normal. Examples include personal income, survival times of cancer patients after treatment, and the lifetime of electronic components.

The decision to describe a distribution by a normal distribution may determine later steps in our analysis of the data such as certain statistical tests. In linear regression, normality of the residuals is a condition for carrying out inference on the slope. How can we judge whether the data (or residuals) are approximately normal?

A histogram or boxplot can reveal obvious non normal features of data such as outliers, skewness, or gaps and clusters. If the histogram or boxplot looks roughly symmetric and unimodel then we need a more sensitive way to judge the adequacy of a normal model. One useful tool for assessing normality is another graph, called the **normal quantile plot**.

Examples Using Simulated Data

Is the normal distribution a good model for the 20 simulated data points shown in the histogram and box plot in Figure 1?

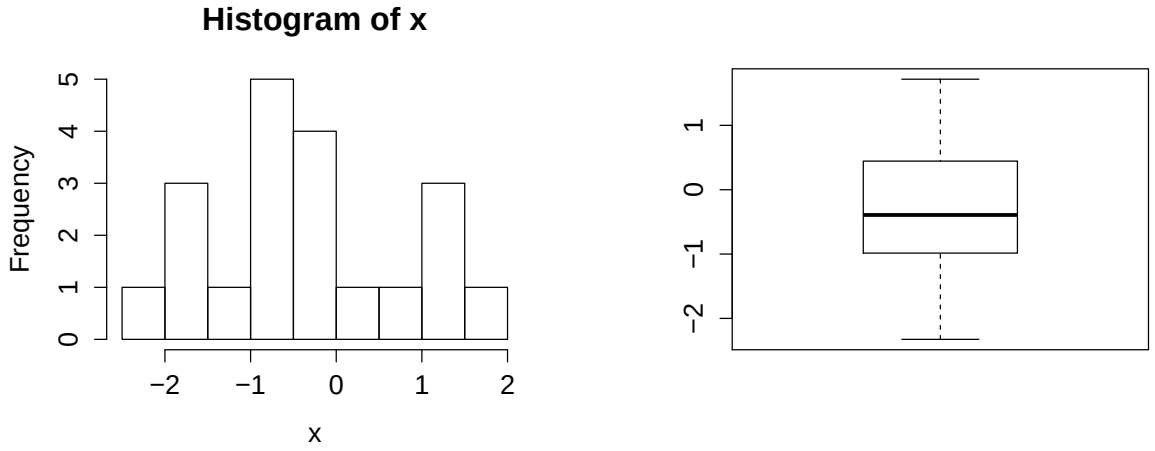

Figure 1: Histogram and Boxplot of Data.

Both plots in Figure 1 indicate the shape of the distribution of the data is symmetric without any outliers. But does this dataset have similar percentiles as a normal distribution?

Recall that any data set that follows the empirical rule (68-95-99.7 rule) has a normal distribution. This idea can be taken one step further by plotting the percentiles of the observed data against the percentiles of the standard normal distribution (that is, the normal distribution with mean 0 and standard deviation 1). Here is the idea:

1. Arrange the observed data values from smallest to largest. Record what percentile the data of each value occupies. For example the smallest observation in a set of 20 observations is the 5^{th} percentile, the second smallest is the 10^{th} percentile, etc.
2. Find the corresponding z-scores for these same percentiles. For example, $z = -1.64$ is the 5^{th} percentile of the standard normal distribution, and $z = -1.28$ is the 10^{th} percentile.
3. Plot each data point against the corresponding percentile from the standard normal distribution. If the distribution is close to standard normal then the plotted points will lie close to a straight line. (Note that R makes a small adjustment to the theoretical quantiles so that the values in the plot are not exactly as described above. But this is a small technical detail that isn't important for your interpretation of the plot.)

The normal quantile plot for the data in Figure 1 is shown in Figure 2.

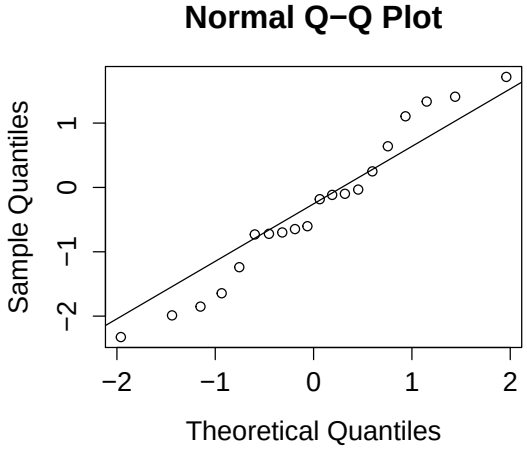

Figure 2: Normal quantile plot of data in Figure 1.

The middle of the distribution is close to a straight line while the tails of the distribution deviate from this line. For example, consider the two smallest data points -2.33 and -1.99. These points are below the line, so these data points are smaller than they would be if these data were normally distributed. This data set shows a small to moderate deviation from the normal distribution. In situations like this, statistical procedures that are robust against

departures from normality, like the t-test, would be appropriate since the deviation from normality is not severe.

Figure 3 shows an example of a normal quantile plot that shows a data set that is approximately normal. Note that the points very closely follow a straight line.

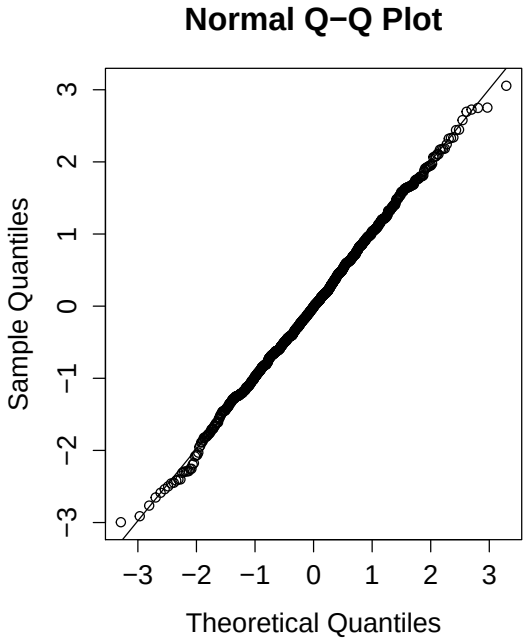

Figure 3: Normal quantile plot of data from a normal distribution.

An example of a data set that shows a high degree of non-normality (due to skewness) is shown in Figure 4. This is a case where a normal distribution would be an inappropriate choice for describing this distribution.

The histogram for the data set in Figure 4 is shown in Figure 5.

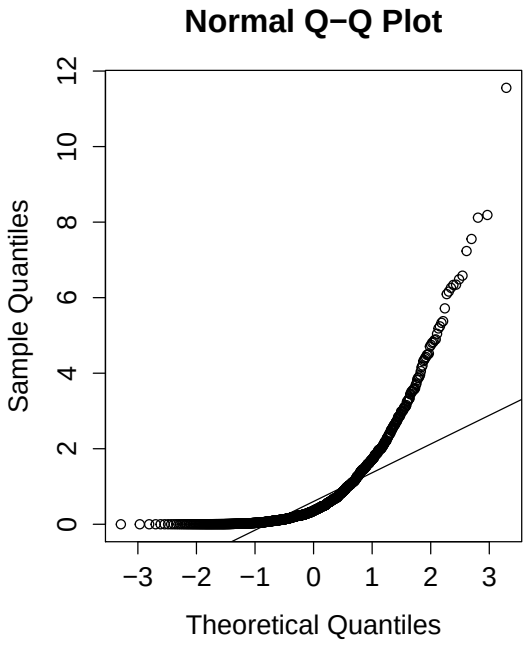

Figure 4: Normal quantile plot of data from a non normal distribution.

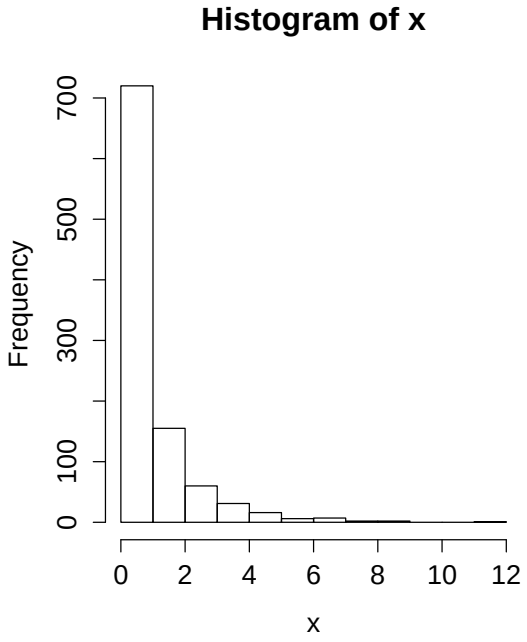

Figure 5: Histogram of data from a non normal distribution. (Correponding normal quantile plot is Figure 4.)

Another example of a normal quantile plot that shows a high degree of non normality (due to outliers) is shown in Figure 6. This is another case where a normal distribution would be an inappropriate choice for describing this distribution.

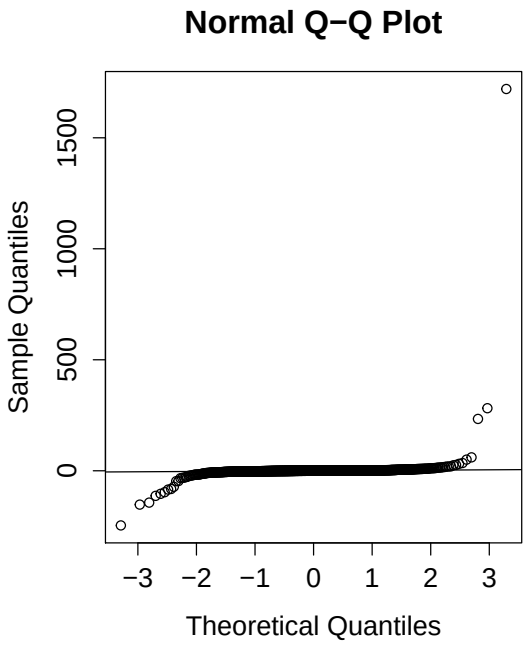

Figure 6: Histogram of data from a non normal distribution.

The boxplot of the data in Figure 6 is shown in Figure 7.

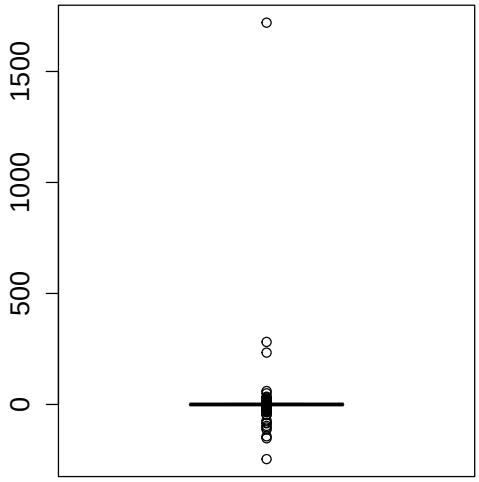

Figure 7: Boxplot of data from a non normal distribution. (Corresponding normal quantile plot is Figure 6.)

It is sometimes possible to find a transformation of skewed data such that the transformed data is closer to normally distributed. A common transformation that reduces skew is a logarithmic transformation. Figure 8 shows a histogram and the corresponding normal quantile plot for a right-skewed dataset. Figure 9 shows a histogram and the corresponding normal quantile plot for the same data, after the natural logarithm has been taken of each data value. (Note that logarithmic transformations with other bases, such as a base 10 logarithm, would have the same effect on the shape.) When using statistical procedures that require the data to be (approximately) normally distributed, it is sometimes helpful to first take a logarithmic transformation of skewed data.

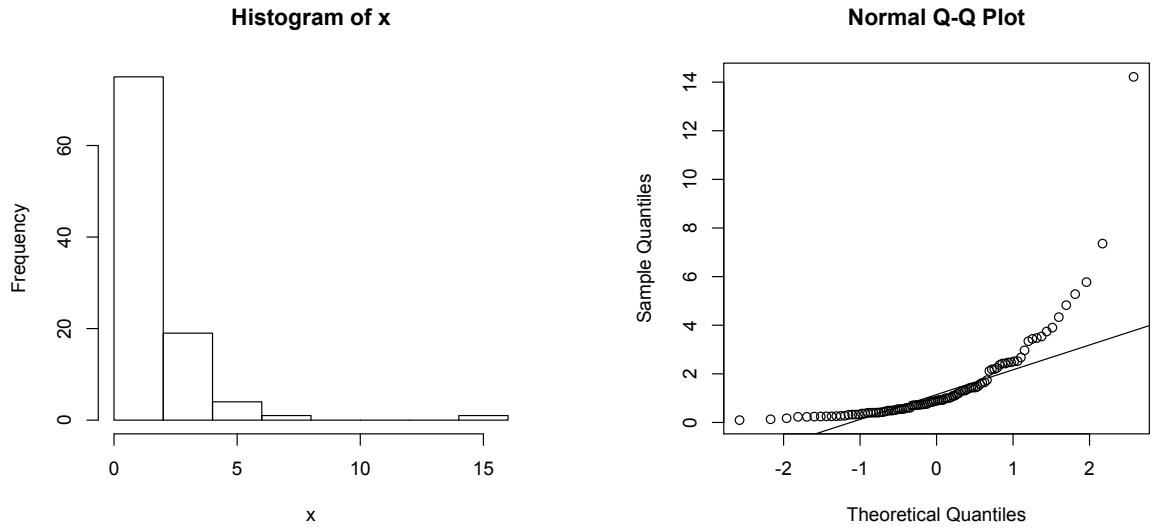

Figure 8: Histogram and normal quantile plot of a dataset with a right-skewed distribution.

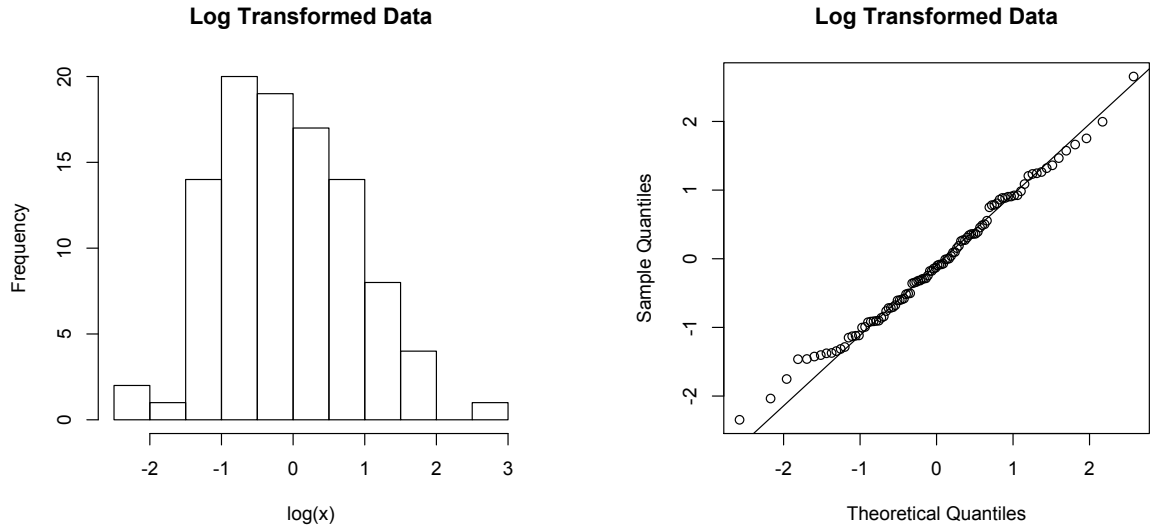

Figure 9: Histogram and normal quantile plot of the dataset in Figure 8 after natural logarithm transformation.