

Prac 3.30**[Computation/variability/variability106]**

The identification of a case as an outlier does not always mean that the case is invalid or abnormal or the result of a mistake. One situation where perfectly normal cases can look like outliers is when there is a mechanism of proportionality at work. Imagine, for instance, that there is a typical rate of production of a substance, and the normal variability is proportional in nature, say from 1/10 of that typical rate to 10 times the rate. This leads to a situation where some normal cases are 100 times as large as others.

To illustrate, look at the `alder.csv` data set, which contains field data from a study of nitrogen fixation in alder plants. The `SNF` variable records the amount of nitrogen fixed in soil by bacteria that reside in root nodules of the plants. Make a box plot and a histogram and describe the distribution. Which of the following descriptions is most appropriate:

- A The distribution is skewed to the left, with outliers at very low values of SNF.
- B The distribution is skewed to the right, with outliers at very low values of SNF.
- C The distribution is roughly symmetrical, although there are a few outliers.

[Prac 3.30-1]

In working with a variable like this, it can help to convert the variable in a way that respects the idea of a proportional change. For instance, consider the three numbers 0.1, 1.0, and 10.0, which are evenly spaced in proportionate terms — each number is 10 times bigger than the preceding number. But as absolute differences, 0.1 and 1.0 are much closer to each other than 1.0 and 10.0.

The *logarithm* function transforms numbers to a scale where even proportions are equally spaced. For instance, taking the logarithm of the numbers 0.1, 1.0, and 10.0 gives the sequence $-1, 0, 1$ — exactly evenly spaced.

The `logSNF` variable gives the logarithm of SNF. Plot out the distribution of `logSNF`. Which of the following descriptions is most apt?

- A The distribution is skewed to the left.
- B The distribution is skewed to the right.
- C The distribution is roughly symmetrical.

[Prac 3.30-2]

You can compute logarithms directly in R, using the functions `log`, `log2`, or `log10`. Which of these functions was used to compute the quantity `logSNF` from `SNF`. (Hint: Try them out!)

log log2 log10 [Prac 3.30-3]

The *base* of the logarithm gives the size of the proportional change that corresponds to a 1-unit increase on the logarithmic scale. For example, `log2` calculates the base-2 logarithm. On the base-2 logarithmic scale, a doubling in size corresponds to a 1-unit increase. In contrast, on the base-10 scale, a ten-fold increase in size gives a 1-unit increase.

Logarithmic transformations are often used to deal with variables that are positive and strongly skewed. In economics, price, income and production variables are often this way. In general, any variable where it is sensible to describe changes in terms of proportion might be better displayed on a logarithmic scale. For example, price inflation rates are usually given as percent (e.g., "The inflation rate was 4% last year.") and so in dealing with prices over time, the logarithmic transformation can be appropriate.