Preparation of data for analysis

Theory R functions Examples Exercise

Before the main analysis, and after the data have been imported into R, it is useful to **explore data** first, check for missing values or outliers, check for range and type of species and environmental data, apply transformation or standardization if necessary, check for possible correlations between environmental variables etc.

Missing values

This is not as trivial as it may sound. Missing data are elements in the matrix with no value, in R usually replaced by NA (not available). Note that there is an important difference between `0` and NA. It makes sense to replace missing value by zero if the entity is really missing (e.g. species was not recorded and gets zero cover or abundance), but it make not sense to replace it by zero if the entity was not recorded (e.g., if I didn't measure pH in some samples because the pH-meter got broken, I should not replace these values by 0, since it does not mean that the pH of that sample is so low). Samples with missing values will be removed from the analysis (often silently without reporting any warning message), and if there are many missing values scattered across different variables, the analysis will be based on rather few samples. One way to reduce this effect is to remove those variables with the highest proportion of missing values from the analysis. Another option is to replace the missing values by estimates if these could be reasonably accurate (mostly by interpolation, e.g. from similar plots, neighbours, values measured at the same time somewhere close, or values predicted by a model).

Outliers

Outliers are those values within given variable which are conspicuously different from other values. Outlier value could get quite influential in the analysis, so it is worth to treat it in advance. How “different” the value should be to become the outlier is often based on a subjective threshold. As a rule, we should not remove some samples as outliers after the analysis just because removing it will improve our result; there must be a more sound reason for it. One option is that the outlier is an error in measurement or sampling; therefore, first, spend a reasonable effort to make sure that such value is not a mistype (either in the field or when retyping the data into a spreadsheet). Another option is that the sample itself really describes conditions which are rather different from the rest of the data set; if there are very few such specific samples, it may be reasonable to remove them, since there may not be enough replications to describe this difference. For example, in the river valley dataset, there are some vegetation plots sampled in limestone, although most of the plots are from an acid bedrock; since there are 97 samples in the data set, and only three are on limestone, it may be reasonable to delete them from the dataset if we are interested how, e.g. soil pH, influences richness or species composition.

There is a number of ways how to detect outliers. A simple exploratory data analysis (EDA) could reveal it graphically, e.g. using a box plot or a histogram. In a box plot, the outlier is defined as a value 1.5 times of interquartile range above upper quartile (Q3) or below lower quartile (Q1); the interquartile range is the range between upper and lower quartile: IQR = Q3-Q1 (Fig. 1).

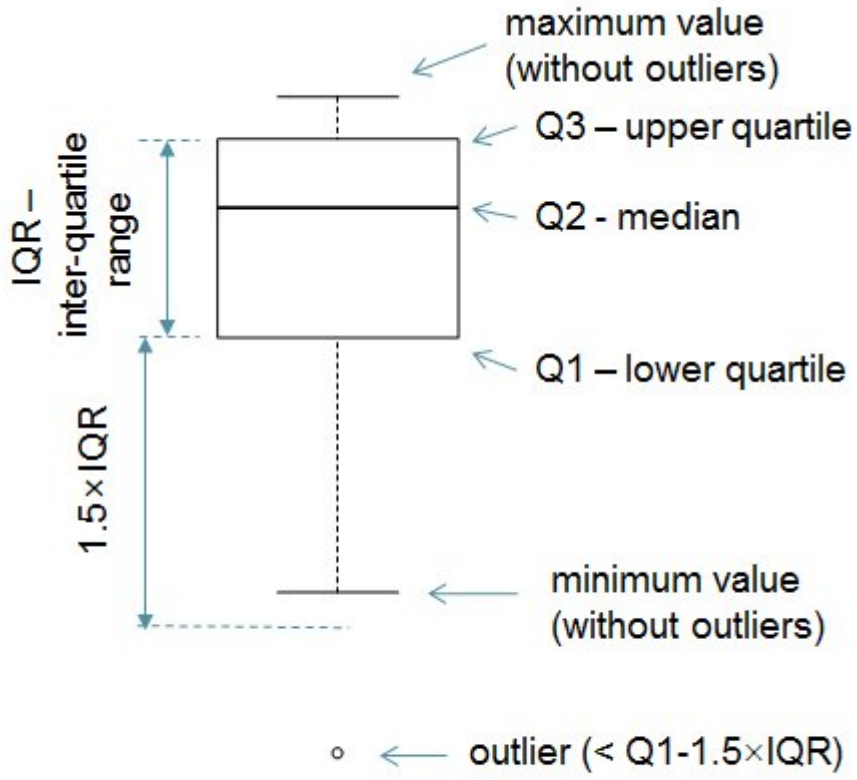

Figure 1:
Definition of outliers in the box plot. An outlier is shown by circle below the non-outlier range of values.

Using pH values available from Vltava river valley dataset as an example, Fig. 2 illustrates the use of box plot and histogram to identify outliers[1)]. Boxplot indicates that there are four outliers with pH value too high. Histogram confirms that three pH values above 6.0 are really separated by a gap from the other pH values. Closer examination reveals that three samples (namely 32, 33 and 34, highlighted by red colour) are from the same transect which was made in a limestone bedrock, which is why they have rather high values of soil pH. When I was sampling the data, I was aware that there is limestone, and I hoped to have high pH samples in my dataset; that time I did not think that it will be the only three plots between all 97 plots which will be on a limestone. These values are therefore not a mistake, but they are outliers since they describe a different phenomenon (forest on limestone bedrock) which does not have enough replicates in the dataset. I may either delete them or go back to the field and try to collect more limestone samples. The fourth value indicated as outlier by boxplot (highlighted by green colour) is a sample done in different area, perhaps also on some small limestone patch; however, since I am not sure with that, I would not remove it as an outlier (according to histogram this value fits to the overall distribution, although this would change if the histogram breaks are set up more fine).

Figure 2: Boxplot (above) and histogram (below) of soil pH values from Vltava river valley dataset.

Data transformation

Data transformation changes relative differences among individual values and consequently also their distribution. We may want to transform data either because (some) statistical analyses and tests require the residuals that tare approximately normally distributed and have homogeneous variance (homoscedasticity), i.e. no relationship between variance and mean, or because linear relationships may be easier to interpret than non-linear. When transforming data, we need to make sure that transformation actually didn't make the distribution of values even worse and didn't actually generate outliers. When commenting results, we should use not transformed values of variables. And, if displayed in the graphs, we should use tick mark labels with untransformed values, or clearly specify that the values shown are transformed.

Types of transformation:

- linear: by adding constant or multiplying by constant (does not change results of statistical tests, e.g. converting temperature measured in °C to °F: T(°F) = T(°C) × 1.8 + 32)
- non-linear:log-transformation, square-root transformation etc. (results of statistical tests are different from tests of not-transformed variables)

A good indicator of whether data need to be transformed is projecting the values using the histograms and checking whether the distribution is symmetrical, right-skewed or left-skewed (Fig. 3). Ecological data are often right-skewed because they are limited by zero at the beginning.

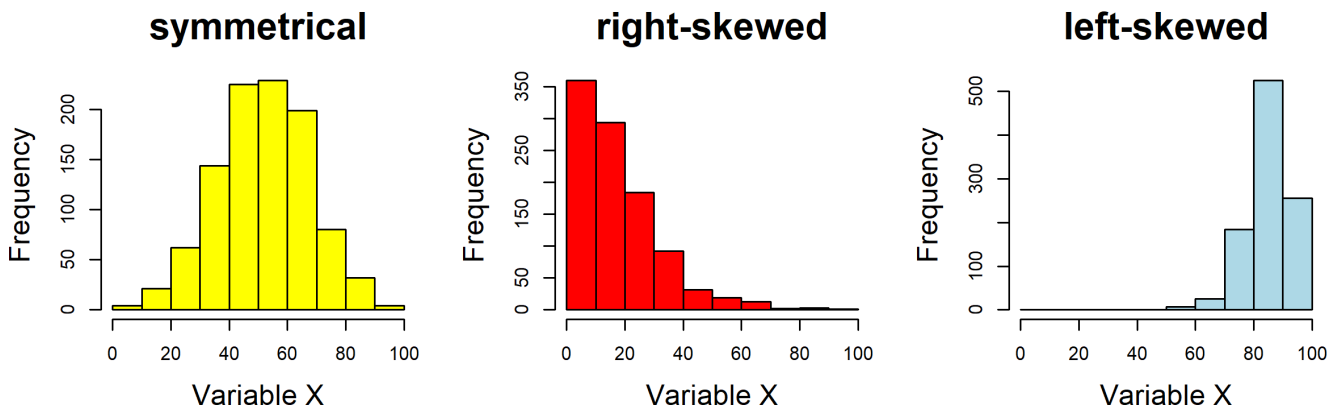

Figure 3

Logarithmic transformation

Log transformation is suitable for strongly right-skewed data with log-normal distribution (with the relationship between mean and variance):

$$y'=\log(y) \text{ or } y'=\log(ay+c)$$

where constant *a* is usually 1, but if *y* is from interval <0;1>, than *a* > 1 (to maintain positive y' values); constant *c* can be added if y contains zeroes, since log (0) is not defined (-Inf), and can be 1 or some arbitrary selected small value (e.g. 0.001). Note that constant *c* can influence results of the analysis (e.g. ANOVA), and it is better to select the value which makes the transformed distribution the most symmetrical. Example on Fig. 4 shows the relationship between the area of the country and it's population; both variables are strongly right skewed, and without transformation, the whole relationship is driven by few large or populous countries; after log transformation, a strong correlation appears.

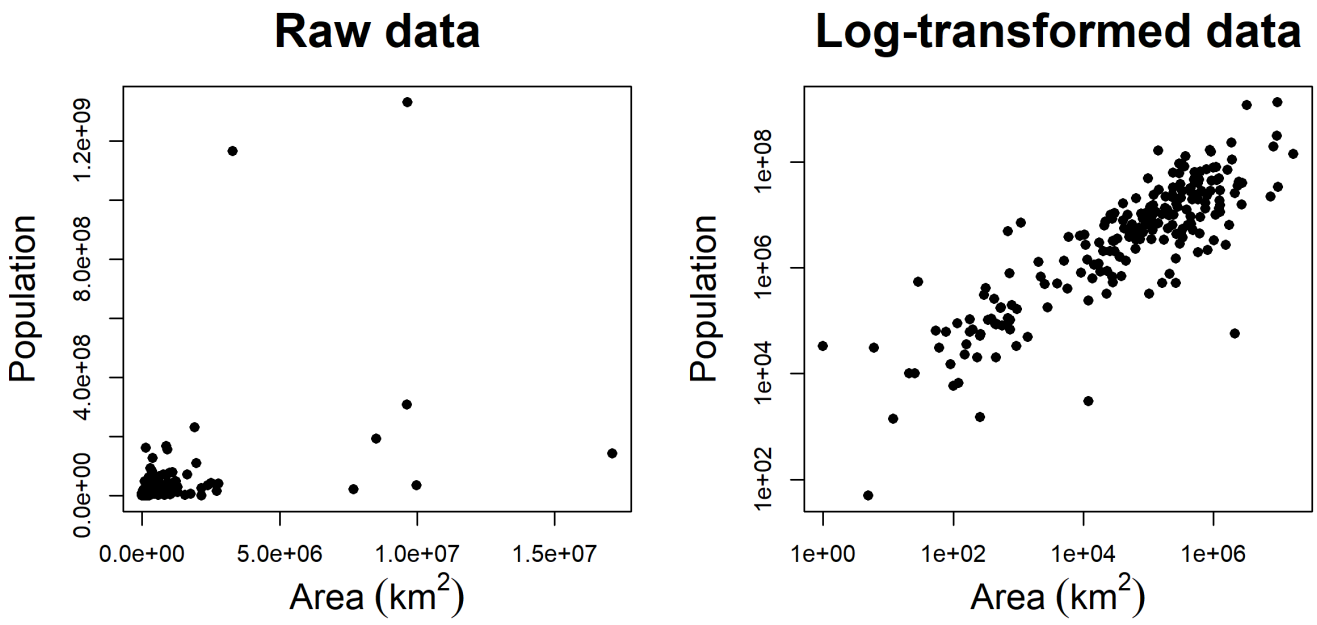

Figure 4

Square-root transformation

Suitable for slightly right-skewed data: $y' = \sqrt{y}$ or $y' = \sqrt{y + c}$

where constant *c* can added if the values contain zeros, and can be e.g. 0.5, or 3/8 (0.325); the higher-root transformation is more powerful for right-skewed data (fourth-or higher root transformation is essentially approaching presence-absence transformation). While log transformation is suitable for strongly right-skewed data, sqrt transformation is suitable for slightly right-skewed data (Fig. 5).

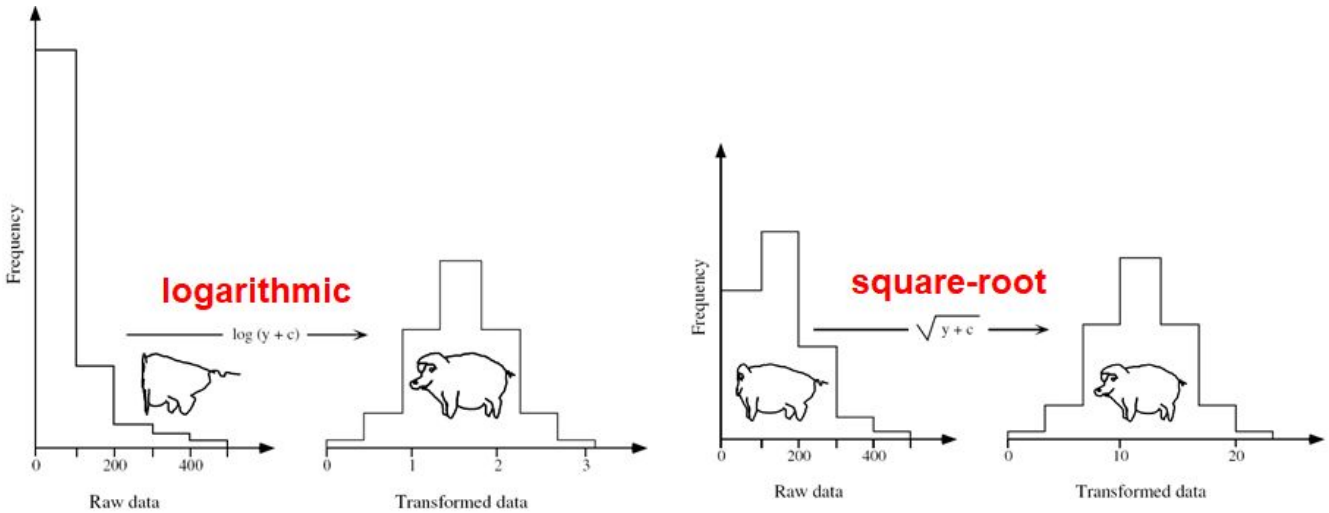

Figure 5: Difference between log and sqrt transformation. For the meaning of the pig shape, see below.

Power transformation

Suitable for left-skewed data:

$y' = y^p$ [to raise *y* on the power of *p*]
which, with *p* values lower than one, becomes root transformation (p = 0.5 – square-root, p = 0.25 – fourth-root etc.)

Arcsin transformation (angular transformation)

Suitable for percentage values (and ratios in general):

$y' = \arcsin(y)$ or $y' = \arcsin\left(\sqrt{y}\right)$
where *y* values must be in the range [−1, 1] and transformed values are in radians within the range [−□/2, □/2].

Reciprocal transformation

Suitable for ratios (e.g. height/weight body ratio, number of children in population per number of females):

$y' = \frac{1}{y}$

Data standardization

Data standardization

[1)]
Script to draw this figure can be found here.