What to do when assumptions about your data fail

AJ Smit *University of the Western Cape*

Normal and non-normal data

Throughout the preceding sections I have stressed the importance of testing the assumptions underlying some statistical tests, in particular *t*-tests, ANOVAs, regressions, and correlations. These statistics are called *parametric statistics* and they require that the assumption of normality and homogeneity of variances are met. This is the kind of statistic you would normally be required to calculate, and because they are commonly used, most people are familiar with parametric statistics. However, when the data are not normal (i.e. skewed) or the variances are unequal — as sometimes happens — the resultant parametric test statistics cannot be used. When this happens, we have two options:

1. apply the *non-parametric* equivalent for the statistical test in question, or
2. transform the data.

It is the intention of this Chapter to discuss some options for data transformations. But before we do that, please revise the non-parametric options available as replacements for the main parametric approaches as may be seen in our online textbook and the succinct summary presented in the Methods Cheatsheet.

Data transformations

If the measurement variable is non-Gaussian and/or has widely different variances in different levels of a treatment, you could apply a data transformation. To transform data, one performs a mathematical operation on each observation in a measurement variable, then use these transformed data in the statistical test. Afterwards, back transform the summary statistics such as the mean or median (including the measures of variance, such as SD and CI) to the original units before reporting it in tables and figures or other means of summary.

Below (i.e. the text on log transformation, square-root transformation, and arcsine transformation) I have extracted mostly verbatim the excellent text produced by John H MacDonald from his Handbook of Biological Statistics. Please attribute this text directly to him. I have made minor editorial changes to point towards some R code, but aside from that the text is more-or-less used verbatim. I strongly suggest reading the preceding text under his Data transformations section, as well as consulting the textbook for in-depth reading about biostatistics. Highly recommended!

Log transformation This consists of taking the log of each observation. You can use either base-10 logs (`log10(x)`) or base-e logs, also known as natural logs (`log(x)`). It makes no difference for a statistical test whether you use base-10 logs or natural logs, because they differ by a constant factor; the base-10 log of a number is just 2.303…× the natural log of the number. You should specify which log you're using when you write up the results, as it will affect things like the slope and intercept in a regression. I prefer base-10 logs, because it's possible to look at them and see the magnitude of the original number: $log(1) = 0$, $log(10) = 1$, $log(100) = 2$, etc.

The back transformation is to raise 10 or e to the power of the number; if the mean of your base-10 log-transformed data is 1.43, the back transformed mean is $10^{1.43} = 26.9$ (in R, `10^1.43`). If the mean of your base-e log-transformed data is 3.65, the back transformed mean is $e^{3.65} = 38.5$ (in R, `exp(3.65)`). If you have zeros or negative numbers, you can't take the log; you should add a constant to each number to make them positive and non-zero (i.e. `log10(x) + 1`). If you have count data, and some of the counts are zero, the convention is to add 0.5 to each number.

Many variables in biology have log-normal distributions, meaning that after log-transformation, the values are normally distributed. This is because if you take a bunch of independent factors and multiply them together, the resulting product is log-normal. For example, let's say you've planted a bunch of maple seeds, then 10 years later you see how tall the trees are. The height of an individual tree would be affected by the nitrogen in the soil, the amount of water, amount of sunlight, amount of insect damage, etc. Having more nitrogen might make a tree 10

Square-root transformation This consists of taking the square root of each observation. The back transformation is to square the number. If you have negative numbers, you can't take the square root; you should add a constant to each number to make them all positive.

People often use the square-root transformation when the variable is a count of something, such as bacterial colonies per petri dish, blood cells going through a capillary per minute, mutations per generation, etc.

Arcsine transformation This consists of taking the arcsine of the square root of a number (in R, `arcsin(sqrt(x))`). (The result is given in radians, not degrees, and can range from $-\pi/2$ to $\pi/2$.) The numbers to be arcsine transformed must be in the range 0 to 1. This is commonly used for proportions, which range from 0 to 1, [...] the back-transformation is to square the sine of the number (in R, `sin(x)^2`).

These are by no means the only types of transformations available. Let us classify the above transformations, and a few others, into categories of the types of corrective actions needed:

- Slightly skewed data
  - `sqrt(x)` for positively skewed data
  - `sqrt(max(x+1) - x)` or `x^2` for negatively skewed data
- Moderately skewed data
  - `log10(x)` for positively skewed data,
  - `log10(max(x + 1) - x)` or `x^3` for negatively skewed data
- Severely skewed data
  - `1/x` for positively skewed data
  - `1/(max(x + 1) - x)` or higher powers than cubes for negatively skewed data
- Deviations from linearity and heteroscedasticity
  - `log(x)` when the dependent variable starts to increase more and more rapidly with increasing independent variable values
  - `x^2` when the dependent variable values decrease more and more rapidly with increasing independent variable values
  - Regression models do not necessarily require data transformations to deal with heteroscedasticity. *Generalised Linear Models* (GLM) can be used with a variety of variance and error structures in the residuals via so-called link functions. Please consult the `glm()` function for details.
  - The linearity requirement specifically applies to linear regressions. However, regressions do not *have* to be linear. Some degree of curvature can be accommodated by additive (polynomial) models, which are like linear regressions, but with additional terms (you already have the knowledge you need to fit such models). More complex departures from linearity can be modelled by non-linear models (e.g. exponential, logistic, Michaelis-Menten, Gompertz, von Bertalanffy and their ilk) or *Generalised Additive Models* (GAM) — these more complex relationships will not be covered in this module. The `gam()` function in the **mgcv** package fits GAMs. After fitting these parametric or semi-parametric models to accommodate non-linear regressions, the residual error structure still does to meet the normality requirements, and these can be tested as before with simple linear regressions.

Knowing how to successfully implement transformations can be as much art as science and requires a great deal of experience to get right. Due to the multitude of options I cannot offer comprehensive examples to deal with all eventualities — so I will not provide any examples at all! I suggest reading widely on the internet or textbooks, and practising by yourselves on your own datasets.