Experimental Design and Data Analysis for Biologists

Gerry P. Quinn
Monash University

Michael J. Keough
University of Melbourne

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521811286

First published in print format 2002

ISBN-13 978-0-511-07812-5 eBook (NetLibrary)
ISBN-10 0-511-07812-9 eBook (NetLibrary)

ISBN-13 978-0-521-81128-6 hardback
ISBN-10 0-521-81128-7 hardback

ISBN-13 978-0-521-00976-8 paperback
ISBN-10 0-521-00976-6 paperback

Linearity
Parametric correlation and linear regression analyses are based on straight-line relationships between variables. The simplest way of checking whether your data are likely to meet this assumption is to examine a scatterplot of the two variables, or a SPLOM for more than two variables. Figure 5.17(a) illustrates how a scatterplot was able to show a nonlinear relationship between number of species of invertebrates and area of mussel clumps on a rocky shore. Smoothing functions through the data can also reveal nonlinear relationships. We will discuss diagnostics for detecting nonlinearity further in Chapter 5.

Independence
This assumption basically implies that all the observations should be independent of each other, both within and between groups. The most common situation where this assumption is not met is when data are recorded in a time sequence. For experimental designs, there are modifications of standard analyses of variance when the same experimental unit is observed under different treatments or times (Chapters 10 and 11). We will discuss independence in more detail for each type of analysis in later chapters.

4.3 Transforming data

We indicated in the previous section that transformation of data to a different scale of measurement can be a solution to distributional assumptions, as well as related problems with variance homogeneity and linearity. In this section, we will elaborate on the nature and application of data transformations.

The justification for transforming data to different scales before data analysis is based, at least in part, on the appreciation that the scales of measurement we use are often arbitrary. For example, many measurements we take are based on a decimal system. This is probably related to the number of digits we have on our hands; characters from the Simpsons would probably measure everything in units of base eight! Sokal & Rohlf (1995) point out that linear (arithmetic) scale of measurement we commonly use can be viewed in the same way. For example, we might measure the length of an object in centimeters but we could just as easily measure the length in log units, such as log centimeters. In fact, we could do so directly just by altering the scale on our measuring device, like using a slide ruler instead of a normal linear ruler.

Surprisingly, transformations are quite common for measurements we encounter in everyday life. Sometimes, these transformations simply change the zero value, i.e. adding a constant. Slightly more complex transformations may change the zero value but also rescale the measurements by a constant value, e.g. the change in temperature units from Fahrenheit to Celsius. Such transformations are linear, in that the relationship between the original variable and the transformed variable is a perfect straight line. Statistical tests of null hypotheses will be identical, in most cases, for the untransformed and the transformed data.

More commonly in data analysis, particularly in biology, are transformations that change the data in a nonlinear fashion. The most common transformation is the log transformation, where the transformed data are simply the logs (to any base) of the original data. The log transformation, while nonlinear, is monotonic, i.e. the order of data values after transformation is the same as before. A log-transformed scale is often the default scale for commonly used measurements. For example, pH is simply the log of the concentration of H^+ ions, and most cameras measure aperture as *f*-stops, with each increase in *f* representing a halving of the amount of light reaching the film, i.e. a $\log_2$ scale.

There are at least five aims of data transformations for statistical analyses, especially for linear models:

- to make the data and the model error terms closer to a normal distribution (i.e. to make the distribution of the data symmetrical),
- to reduce any relationship between the mean and the variance (i.e. to improve homogeneity of variances), often as a result of improving normality,

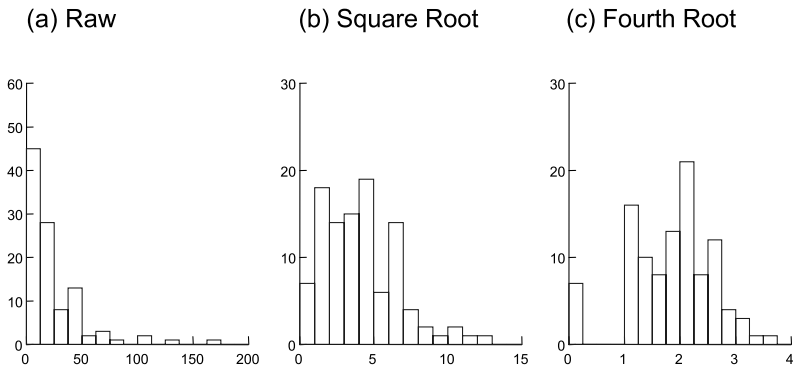

Figure 4.8 Distribution of counts of limpets in quadrats at Point Nepean: (a) untransformed (raw), (b) square root transformed, and (c) fourth root transformed. (M Keough & G. Quinn, unpublished data.)

- to reduce the influence of outliers, especially when they are at one end of a distribution,
- to improve linearity in regression analyses, and
- to make effects that are multiplicative on the raw scale additive on a transformed scale, i.e. to reduce the size of interaction effects (Chapters 6 and 9).

The most common use of transformations in biology is to help the data meet the distributional and variance assumptions required for linear models. Emerson (1991), Sokal & Rohlf (1995) and Tabachnick & Fidell (1996) provide excellent descriptions and justification of transformations. These authors are reassuring to those who are uncomfortable about the idea of transforming their data, feeling that they are "fiddling" the data to increase the chance of getting a significant result. A decision to transform, however, is always made before the analysis is done.

Remember that after any transformation, you must re-check your data to ensure the transformation improved the distribution of the data (or at least didn't make it any worse!). Sometimes, log or square root transformations can skew data just as severely in the opposite direction and produce new outliers!

A transformation is really changing your response variable and therefore your formal null hypothesis. You might hypothesize that growth of plants varies with density, and formalize that as the H_0 that the mean growth of plants at high density equals the mean growth at low density. If you are forced to log-transform your data, the null hypothesis becomes "mean log-growth does not vary with density", or you might say that in the first case, growth is defined as mg of weight gained, whereas after log-transforming, growth is the log-mg weight gained.

4.3.1 Transformations and distributional assumptions

The most common type of transformation useful for biological data (especially counts or measurements) is the power transformation (Emerson 1991, Neter *et al*. 1996), which transforms Y to Y^p, where p is greater than zero. For data with right skew, the square root ($\sqrt{}$) transformation, where $p = 0.5$, is applicable, particularly for data that are counts (Poisson distributed) and the variance is related to the mean. Cube roots ($p = 0.33$), fourth roots ($p = 0.25$), etc., will be increasingly effective for data that are increasingly skewed; fourth root transformations are commonly used for abundance data in ecology when there are lots of zeros and a few large values (Figure 4.8). For very skewed data, a reciprocal transformation can help, although interpretation is a little difficult because then order of values is reversed.

Transforming data to logarithms (the base is irrelevant although base 10 logs are more familiar to readers) will also make positively skewed distributions more symmetrical (Keene 1995; Figure 4.9), especially when the mean is related to the standard deviation. Such a distribution is termed lognormal because it can be made normal by log transforming the values. Use log $(Y + c)$ where c is an appropriate constant if there are zeros in the data set because you can't take the log of zero. Some people use the smallest possible value for their variable as a constant, others use an arbitrarily small number, such as 0.001 or, most

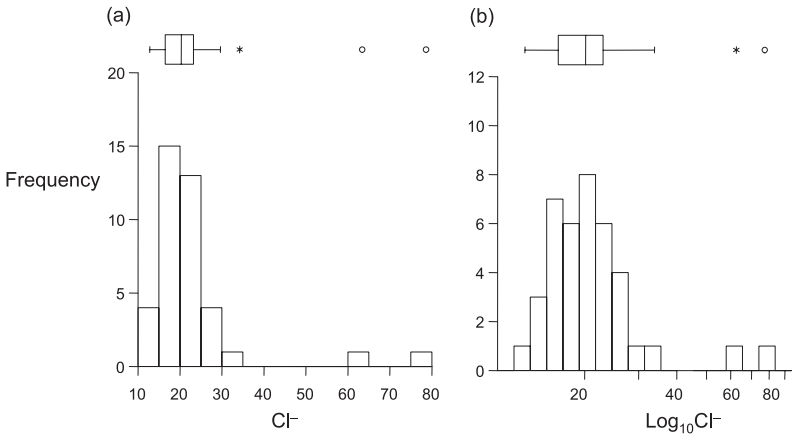

Figure 4.9 Frequency distribution and box plots for concentrations of Cl^- for 39 sites from forested watersheds in the Catskill Mountains in New York State: (a) untransformed and (b) $\log_{10}$-transformed (data from Lovett *et al.* 2000).

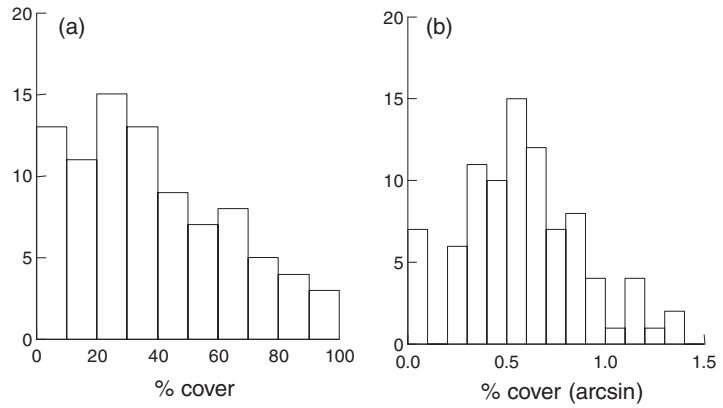

Figure 4.10 Distribution of percentage cover of the alga *Hormosira banksii* in quadrats at Point Nepean: (a) untransformed (raw) and (b) arcsin transformed. (M Keough & G. Quinn, unpublished data.)

commonly, 1. Berry (1987) pointed out that different values of c can produce different results in ANOVA tests and recommended using a value of c that makes the distribution of the residuals as symmetrical as possible (based on skewness and kurtosis of the residuals).

If skewness is actually negative, i.e. the distribution has a long left tail, Tabachnick & Fidell (1996) suggested reflecting the variable before transforming. Reflection simply involves creating a constant by adding one to the largest value in the sample and then subtracting each observation from this constant.

These transformations can be considered part of the Box–Cox family of transformations:

$$\frac{Y^{\lambda}-1}{\lambda} \text{ when } \lambda \neq 0 \tag{4.1}$$

$$\log(Y) \text{ when } \lambda = 0 \tag{4.2}$$

When $\lambda = 1$, we have no change to the distribution, when $\lambda = 0.5$ we have the square root transformation, and when $\lambda = -1$ we have the reciprocal transformation, etc. (Keene 1995, Sokal & Rohlf 1995). The Box–Cox family of transformations can also be used to find the best transformation, in terms of normality and homogeneity of variance, by an iterative process that selects a value of λ that maximizes a log-likelihood function (Sokal & Rohlf 1995).

When data are percentages or proportions, they are bounded at 0% and 100%. Power transformations don't work very well for these data because they change each end of the distribution differently (Emerson 1991). One common approach is to use the angular transformation, specifically the arcsin transformation. With the data expressed as proportions, then transform Y to $\sin^{-1}(\sqrt{Y})$, and the result is shown in Figure 4.10. It is most effective if Y is close to zero or one, and has little effect on mid-range proportions.

Finally, we should mention the rank transformation, which converts the observations to ranks, as described in Chapter 3 for non-parametric tests. The rank transformation is different from the

other transformations discussed here because it is bounded by one and n, where n is the sample size. This is an extreme transformation, as it results in equal differences (one unit, except for ties) between every pair of observations in this ranked set, regardless of their absolute difference. It therefore results in the greatest loss of information of all the monotonic transformations.

For common linear models (regressions and ANOVAs), transformations will often improve normality and homogeneity of variances and reduce the influence of outliers. If unequal variances and outliers are a result of non-normality (e.g. skewed distributions), as is often the case with biological data, then transformation (to log or square root for skewed data) will improve all three at once.

4.3.2 Transformations and linearity

Transformations can also be used to improve linearity of relationships between two variables and thus make linear regression models more appropriate. For example, allometric relationships with body size have a better linear fit after one or both variables are log-transformed. Note that nonlinear relationships might be better investigated with a nonlinear model, especially one that has a strong theoretical justification.

4.3.3 Transformations and additivity

Transformations also affect the way we measure effects in linear models. For example, let's say we were measuring the effect of an experimental treatment compared to a control at two different times. If the means of our control groups are different at each time, how we measure the effect of the treatment is important. Some very artificial data are provided in Table 4.1 to illustrate the point. At Time 1, the treatment changes the mean value of our response variable from 10 to 5 units, a decrease of 5 units. At Time 2 the change is from 50 to 25 units, a change of 25 units. On the raw scale of measurement, the effects of the treatments are very different, but in percentage terms, the effects are actually identical with both showing a 50% reduction. Biologically, which is the most meaningful measure of effect, a change in raw scale or a change in percentage scale? In many cases, the percentage change might be more biologically relevant and we would want our analysis to conclude that the treatment effects are the same at the two times. Transforming the data to a log scale achieves this (Table 4.1).

Table 4.1 Means for treatment and control groups for an experiment conducted at two times. Artificial data and arbitrary units used.

	Untransformed		Log-transformed	
	Time 1	Time 2	Time 1	Time 2
Control	10	50	1.000	1.699
Treatment	5	25	0.699	1.398

Interpretation of interaction terms in more complex linear models (Chapter 9) can also be affected by the scale on which data are measured. Transforming data to reduce interactions may be useful if you are only interested in main effects or you are using a model that assumes no interaction (e.g. some randomized blocks models; Chapter 10). Log-transformed data may better reflect the underlying nature and interpretation of an interaction term.

4.4 Standardizations

Another change we can make to the values of our variable is to standardize them in relation to each other. If we are including two or more variables in an analysis, such as a regression analysis or a more complex multivariate analysis, then converting all the variables to a similar scale is often important before they are included in the analysis. A number of different standardizations are possible. Centering a variable simply changes the variable so it has a mean of zero:

$$y_i = y_i - \bar{y} \tag{4.3}$$

This is sometimes called translation (Legendre & Legendre 1998).

Variables can also be altered so they range from zero (minimum) to one (maximum). Legendre & Legendre (1998) describe two ways of achieving this:

$$y_i = \frac{y_i}{y_{\max}} \text{ and } y_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \tag{4.4}$$