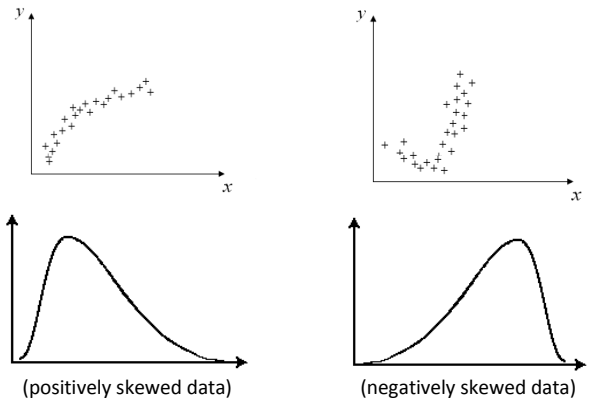
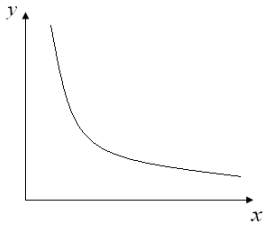
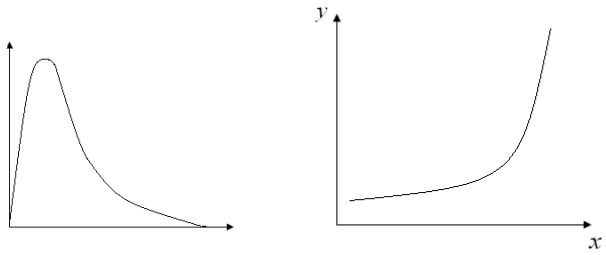
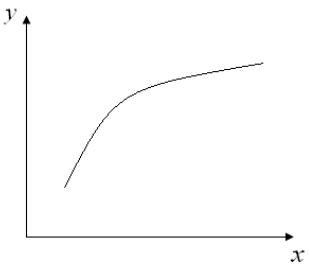
Author: Marta Emmett

Transformations

Date amended: 8/8/2018

Several statistical techniques require data assumptions to hold, the most common assumption failures are on normality, linearity and homocedasticity (constant variance). If the data is not normally distributed then non-parametric methods, generalized linear models (Poisson, exponential, binomial…) or other methods could be used. Another option, however, is to investigate whether data transformations can normalize, linearize the data or improve the homoceasticity. The most commonly used are the logarithm (Econometrics), the square root and the inverse function (Chemistry, Biology).

Data	Transformation
 (positively skewed data) (negatively skewed data)	*Square root* To reduce the curvature: square/cube one variable or take the square root (variance proportional to mean, used for count data) Slight /moderate positive (right) skewness: use square root $\sqrt{y}, \quad y \geq 0$ Slight /moderate negative(left) skewness: use square root $\sqrt{(k-y)}$
	*Inverse/Reciprocal* use for high skewness (variance proportional to mean^4) $\frac{1}{y}$ or $\frac{1}{x}$, $y \neq 0$ or $x \neq 0$ If negative skew $\frac{1}{k-y}$
 Group standard deviation increases with mean	*Logarithmic* use logs for substantial skewness (variance proportional to mean^2) $\ln(y)$ or $\log_{10}(y)$, $y>0$ Using log will make the data less skewed, make the variances more homogeneous and linearise the data! If negative skew $\log(k-y)$
	*Logarithmic* can also be applied to the explanatory variables $\log(x)$

Where k is a constant from which each score is subtracted so that smallest score is 1 (usually largest score +1)

Logarithm

The natural log transformation seems to be used extensively in Financial and Medical data.

Mean: The 'anti-log' of the mean is the geometric mean.

Standard deviation: cannot be back-transformed.

CI: need to be calculated in transform data and then 'anti-logged'. This will be the CI for the geometric mean

Regression: the coefficients and CIs are 'anti-logged'; the interpretation is that y changes by $100\left(e^{\beta_i}-1\right)$% for 1 unit increase of x_i when the other independent variables remain constant.

Correlation: do not 'anti-log'

Reciprocal

Mean: The back-transformation is the harmonic mean

Standard deviation: cannot be back-transformed

CI: cannot be back-transformed

Square Root

Mean: can be back-transformed

Standard deviation: cannot be back-transformed

CI: can be back-transformed

Other transformations

- **Box-Cox transformation** (power transformation):

$$y^* = \begin{cases} \dfrac{(y^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

These include most of the traditional transformations, e.g. $\lambda = \frac{1}{2}$ is the square root, etc. These transformations are not always implemented in packages, but λ can be estimated with the aid of a Box-Cox normal probability plots and by computing correlation coefficients.

- **Arcsine** transformation: this transformation is used for proportions $arcsin\left(\sqrt{p}\right)$

Dealing with zeros

If the data has zeros the logarithmic function cannot be used, an option is to replace zero with a small value (e.g. 1 or half of the smallest observed value or a small values that when added to the lowest observation is 1). This value can either be added to the zeros or to the whole data set.

References

Bland JM, Altman DG (1996). Transforming data. *British Medical Journal*; 312: 770.

Peacock JL, Peacock PJ (2011). *Oxford Handbook of Medical Statistics*. Oxford University Press.

Tabachnick BG, Fidell LS (2007). *Using Multivariate Statistics*. Pearson.

Lewis JP, Traill A (1999). *Statistics Explained*. Addison Wesley.