Box 1. Definitions of some statistic terms (in an alphabetical order).

- **Akaike information criteria (AIC):** a measure of the relative goodness of fit of a statistical model, developed by Hirotsugu Akaike. It offers a relative measure of the information lost when a given model is used to describe reality (the data). Given a data set, several candidate models may be ranked according their AIC values; AIC provides a means for comparison among models but can tell nothing about how well a model fits the data in an absolute sense.
- **Analysis of variance (ANOVA):** a collection of statistical models in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation. The simplest ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups.
- **Explanatory (predictor) variable:** the independent variables in a statistic model (the X values). It is typically the variable representing the value being manipulated or changed in an experiment, and also refers to already existing characters of the subject involved in the study, such as gender and age for people, or body mass and life history for animals. Note that despite the terminology, values of "independent variables" are rarely statistically independant. Therefore, the term "explanatory variable" is preferable in general.
- **General linear model (GLM):** a flexible generalization of ordinary least squares regression (linear regression). GLM is a way of unfying various other statistical models, including linear regression, logistic regression and Poisson regression.
- **Goodness of fit:** a description of a statistical model describing how well it fits a set of obervations (the data). Measures of goodness of fit typically summarize the descrepancy between observed values and the values expected under the model in question.
- **Histogram:** a graphical representation showing a visual impression of the distribution of data. It is usually an estimate of the probability distribution of a continuous variable. A histogram concists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area (and height) equal to the frequency (or density) of the observations in the interval.
- **Model selection:** the task of selecting a statistical model from a set of potential models, given data. In the simplest cases, a pre-existing set of data is considered.
- **Random forest:** a method using multiple models to obtain better predictive performance by sampling random sub-populations of the observations. The random forest consists of many decision trees, which map obervations about an item to conclusions about the iterm's target value using leaves to represent classifications and branches to represent conjunctions of features that lead to those classifications. Random forest analysis outputs the class that is the most frequent value of the class's output by individual trees.
- **Response variable:** the dependent variable in a statistic model (the Y values). It is the observed result of the independent variable being manipulated or the variance of existing characters.